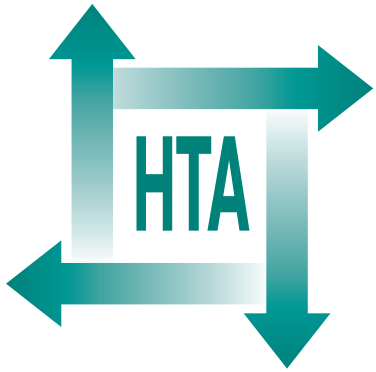
Health Technology Assessment 2005; Vol. 9: No. 13

Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK

BH Willis, P Barton, P Pearmain, S Bryan and C Hyde

March 2005

Health Technology Assessment
NHS R&D HTA Programme



How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK

BH Willis,[1] P Barton,[2] P Pearmain,[3] S Bryan[2] and C Hyde[1*]

[1] ARIF, Department of Public Health and Epidemiology, University of Birmingham, UK
[2] Health Economics Facility, Health Services Management Centre, University of Birmingham, UK
[3] West Midlands Breast and Cervical Screening QA Reference Centre, University of Birmingham, UK

* Corresponding author

Declared competing interests of authors: none of the authors or the units to which they belong has any pecuniary relationship specific or non-specific with manufacturers of automated cervical smear image analysis devices, past or present. C Hyde is a member of the editorial board for *Health Technology Assessment*, although he was not involved in the editorial process for this report.

Published March 2005

This report should be referenced as follows:

Willis BH, Barton P, Pearmain P, Bryan S, Hyde C. Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK. *Health Technol Assess* 2005;**9**(13).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE and *Science Citation Index Expanded (SciSearch®)* and *Current Contents®*/Clinical Medicine.

NHS R&D HTA Programme

The research findings from the NHS R&D Health Technology Assessment (HTA) Programme directly influence key decision-making bodies such as the National Institute for Clinical Excellence (NICE) and the National Screening Committee (NSC) who rely on HTA outputs to help raise standards of care. HTA findings also help to improve the quality of the service in the NHS indirectly in that they form a key component of the 'National Knowledge Service' that is being developed to improve the evidence of clinical practice throughout the NHS.

The HTA Programme was set up in 1993. Its role is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined to include all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care, rather than settings of care.

The HTA programme commissions research only on topics where it has identified key gaps in the evidence needed by the NHS. Suggestions for topics are actively sought from people working in the NHS, the public, consumer groups and professional bodies such as Royal Colleges and NHS Trusts.

Research suggestions are carefully considered by panels of independent experts (including consumers) whose advice results in a ranked list of recommended research priorities. The HTA Programme then commissions the research team best suited to undertake the work, in the manner most appropriate to find the relevant answers. Some projects may take only months, others need several years to answer the research questions adequately. They may involve synthesising existing evidence or designing a trial to produce new evidence where none currently exists.

Additionally, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme is able to commission bespoke reports, principally for NICE, but also for other policy customers, such as a National Clinical Director. TARs bring together evidence on key aspects of the use of specific technologies and usually have to be completed within a limited time period.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this monograph was commissioned and funded by the HTA Programme on behalf of NICE as project number 98/38/01. The authors have been wholly responsible for all data collection, analysis and interpretation and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme, NICE or the Department of Health.

ISSN 1366-5278

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.

T

Abstract

Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK

BH Willis,[1] P Barton,[2] P Pearmain,[3] S Bryan[2] and C Hyde[1*]

[1] ARIF, Department of Public Health and Epidemiology, University of Birmingham, UK
[2] Health Economics Facility, Health Services Management Centre, University of Birmingham, UK
[3] West Midlands Breast and Cervical Screening QA Reference Centre, University of Birmingham, UK
* Corresponding author

Objectives: To assess the effectiveness and cost-effectiveness of adding automated image analysis to cervical screening programmes.
Data sources: Searching of all major electronic databases to the end of 2000 was supplemented by a detailed survey for unpublished UK literature.
Method: Four systematic reviews were conducted according to recognised guidance. The review of 'clinical effectiveness' included studies assessing reproducibility and impact on health outcomes and processes in addition to evaluations of test accuracy. A discrete event simulation model was developed, although the economic evaluation ultimately relied on a cost-minimisation analysis.
Results: The predominant finding from the systematic reviews was the very limited amount of rigorous primary research. None of the included studies refers to the only commercially available automated image analysis device in 2002, the AutoPap Guided Screening (GS) System. The results of the included studies were debatably most compatible with automated image analysis being equivalent in test performance to manual screening. Concerning process, there was evidence that automation does lead to reductions in average slide processing times. In the PRISMATIC trial this was reduced from 10.4 to 3.9 minutes, a statistically significant and practically important difference. The economic evaluation tentatively suggested that the AutoPap GS System may be efficient. The key proviso is that credible data become available to support that the AutoPap GS System has test performance and processing times equivalent to those obtained for PAPNET.
Conclusions: The available evidence is still insufficient to recommend implementation of automated image analysis systems. The priority for action remains further research, particularly the 'clinical effectiveness' of the AutoPap GS System. Assessing the cost-effectiveness of introducing automation alongside other approaches is also a priority.

Contents

List of abbreviations vii

Executive summary ix

1 Background 1
Summary of key points 1
Cervical cancer: realising the potential for its prevention through screening 1
Current cervical screening in the NHS 2
Problems with the current NHS cervical screening programme 3
Limitations of the current slide preparation and analysis 4
Improving slide preparation and analysis using automation and related technology 4
International situation with respect to automated devices 8

2 Aim and objectives of the health technology assessment project 9
Summary of key points 9
Introduction 9
Overall aim and objective 9
Specific components of the health technology assessment project and their objectives 9
Interrelationship of the specific components of the project 10
General method 10

3 Systematic review of the secondary literature on clinical effectiveness of automation and related technologies 13
Summary of key points 13
Introduction and objectives 13
Method 14
Results 15

4 Systematic review of the literature on cost-effectiveness of automation 27
Summary of key points 27
Introduction 27
Methods 27
Systematic review of literature on cost-effectiveness: results 28
Systematic review of literature on cost-effectiveness: conclusions 38
Selective review of approaches to simulation modelling: results 38
Selective review of approaches to simulation modelling: conclusions 39

5 Evaluating automated cervical screening: methodological issues 45
Summary of key points 45
Introduction and objective 46
Part 1: Clinical effectiveness 46
Part 2: Health economics 56

6 Systematic review of the effects and effectiveness of automation 59
Summary of key points 59
Introduction 60
General method 61
Part 1: Test performance 62
Part 2: Reproducibility 87
Part 3: Health outcomes 94
Part 4: Process 99
Overall conclusion from all parts of systematic review of effectiveness 104

7 Likelihood of publication bias in reviews of research on the effects of automation 105
Summary of key points 105
Introduction 105
Locating unpublished literature on automated cervical screening 106
Extending and generalising the search for unpublished literature 114
Reflecting on the nature of publication bias 115

8 Costs of automated cervical screening devices 119
Summary of key points 119
Introduction 119
Method 119
Results 120
Conclusions 127

9 Modelling the health economic impact of automation 131
Summary of key points 131
Introduction 131
Model for cervical cancer progression and screening 132

Alternative approach to the assessment of efficiency 142
Conclusions concerning efficiency of introducing automation 144

10 Discussion and overall conclusions 145
Main findings 145
Findings in the context of past health technology assessments 146
Conclusions 147
Implications for patients and practice 147
Implications for further research on automation in cervical screening 148
Implications for methodological research 148
When to repeat this health technology assessment 149

Acknowledgements 151

References 153

Appendix 1 Secondary literature on clinical effectiveness of automation and related technologies: MEDLINE search 1997–2000 159

Appendix 2 Secondary literature on clinical effectiveness of automation and related technologies: excluded studies 161

Appendix 3 Secondary literature on clinical effectiveness of automation and related technologies: appraising the included studies 163

Appendix 4 Secondary literature on clinical effectiveness of automation and related technologies: detailed evidence tables 165

Appendix 5 Search algorithm used on MEDLINE to identify articles on cost and cost-effectiveness of automated image analysis devices 179

Appendix 6 Search strategies for systematic reviews of evidence on clinical effectiveness 181

Appendix 7 Included and excluded studies in systematic review of test performance 183

Appendix 8 Using verification of discordant cytology as a reference standard 191

Appendix 9 General considerations concerning achieving representativeness of literature included in systematic reviews and health technology assessment 193

Appendix 10 Patient walk-through for the NHS Cervical Screening Programme 195

Health Technology Assessment reports published to date 209

Health Technology Assessment Programme 219

List of abbreviations

AC	Autocyte SCREEN
AGUS	atypical glandular cells of undetermined significance
AHCPR	Agency for Health Care Policy and Research (now AHRQ)
AHRQ	Agency for Healthcare Research and Quality
AHTAC	Australian Health Technology Advisory Committee
AP	AutoPap
ASCUS	atypical squamous cells of undetermined significance
Auto	automated image analysis
AutoPap 300 QC*	AutoPap 300 Quality Control
AutoPap GS System*	AutoPap Guided Screening System
BMS	biomedical scientist
BSCC	British Society for Cervical Cytology
BSCCP	British Society for Colposcopy and Cervical Pathology
CEA	cost-effectiveness analysis
CI	confidence interval
CIN	cervical intraepithelial neoplasia
CIS	carcinoma *in situ*
CRD	Centre for Reviews and Dissemination
DARE	Database of Abstracts of Reviews of Effectiveness
DES	discrete event simulation
FDA	Food and Drug Administration
FIGO	International Federation of Gynaecology and Obstetrics
FNR	false-negative rate
FOV	field of view
FPR	false-positive rate
HPV	human papillomavirus
HSIL	high-grade squamous intraepithelial lesion
HSTAT	Health Services/Technology Assessment Text
ICER	incremental cost-effectiveness ratio
INAHTA	International Network of Agencies for Health Technology Assessment
ISWG	Intersociety Working Group (for Cytology Technologies)
κ	(Cohen's) kappa statistic
κ_w	weighted kappa
LBC	liquid-based cytology
LLETZ	large loop excision of the transformation zone of the cervix
LREC	local research ethics committee
LSIL	low-grade squamous intraepithelial lesion
META	Medical Editors Trials Amnesty
MREC	multicentre research ethics committee
NA	not applicable
NFR	'no further review'
NHSCSP	NHS Cervical Screening Programme

continued

List of abbreviations *continued*

NHS EED	NHS Economics Evaluations Database
NICE	National Institute for Clinical Excellence
NPV	negative predictive value
NRR	National Research Register
ns	not significant
NSI	Neuromedical Systems Inc.
p.a.	per annum
PALGA	Dutch Automated Archive of Pathology Laboratories
Pap smear	Papanicolaou (cervical) smear
PN	PAPNET
PPV	positive predictive value
QA	quality assurance
QARC	Quality Assurance Reference Centre
QSE	quasi-standard error
QUOROM	Quality of Reporting of Meta-analyses
RCT	randomised controlled trial
REC	research ethics committee
SCC	squamous cell carcinoma
Se	sensitivity
SIL	squamous intraepithelial lesion
Sp	specificity
TNR	true-negative rate
TPR	true-positive rate
WNL	within normal limits

* Device and manufacturer names used in this report were correct at the time the report was written. Subsequently, the authors were informed that the device name of AutoPap had changed to FocalPoint slide profiler and the manufacturer from TriPath Imaging Inc to TriPath Care Technologies Inc. (Jackson AK, CellPath plc, UK: personal communication, 28 February 2002). While noting these changes, the authors have not altered manufacturer or device names used in the original version of the report.

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices in which case the abbreviation is defined in the figure legend or at the end of the table.

Executive summary

Background

Cervical cancer is a serious, but fortunately rare disease. Cervical screening programmes have undoubtedly contributed to reductions in incidence and mortality, but the cost, both financial and logistic, has been high. It has been hoped that technological advances, including automated image analysis of cervical smears, would help. However, the technology is expensive, the only currently commercially available automated image analysis devices costing in excess of £0.5 million each in 2001. The implied implementation cost for the NHS in England alone is conservatively estimated at £40 million. Inevitably, there has been concern about whether such costs can be justified.

Automated image analysis involves the translation of a cervical smear into a computerised image, which is then analysed to identify slides with cells likely to be abnormal. Increasingly, the location of abnormal cells on the slide is automatically recorded to facilitate review of the slide. Automated image analysis is incorporated into the existing manual screening procedure. In current devices the attention is on replacing the primary screening step. There has been considerable development of the technology since it was first introduced in the early 1990s. PAPNET and AutoPap have been the two main competing devices; however, commercial pressure has meant that the AutoPap Guided Screening (GS) System is now the only one available.

This assessment was completed in April 2002. Device and manufacturer names used in this report were correct at the time the report was written. Subsequently, the authors were informed that the device name of AutoPap had changed to FocalPoint slide profiler and the manufacturer from TriPath Imaging Inc to TriPath Care Technologies Inc. (Jackson AK, CellPath plc, UK: personal communication, 28 February 2002). While noting these changes, the authors have not altered manufacturer or device names used in the original version of the report.

Objectives

The overall objective of the project was to assess the immediate effects, the wider consequences and costs, and overall cost-effectiveness and cost-utility of introducing automated image analysis to a screening programme with characteristics similar to those currently operating in the UK.

Methods

A health technology assessment was undertaken. This had six interrelated components, each with its own specific objectives. Four systematic reviews of past reviews and health technology assessments, assessments of cost-effectiveness, assessments of clinical effectiveness and cost data, supplemented with a detailed survey for unpublished UK literature, fed into an attempt to model the cost-effectiveness of automated image analysis relative to manual screening alone. A discrete event simulation (DES) model of cervical screening was developed to overcome some anticipated limitations of other modelling approaches. All systematic reviews were carried out in accordance with recognised guidance. The searches for the systematic reviews covered all major electronic databases to the end of 2000. A special feature of the clinical effectiveness review was that studies assessing reproducibility, impact on process and impact on health outcomes were targeted in addition to studies assessing test performance.

Results

The predominant finding from the systematic reviews was the very limited amount of rigorously conducted primary research. For instance, concerning test performance, only two studies (approximately 13,000 slides) assessing impact on sensitivity and specificity of automated image analysis were included; even relaxing these criteria only allowed another five studies (approximately 51,000 slides) to be considered. The results of these studies were difficult to interpret, but

debatably were most compatible with automated image analysis being equivalent in test performance to manual screening. Several studies provided information on reproducibility of assessments, which was often surprisingly poor. Two evaluations of impact on health outcomes were identified, and although they did not contribute directly to the conclusions, they point to a type of evaluation that should be considered more often. Concerning process, there was evidence that automated image analysis does lead to reductions in average slide processing times. In the PRISMATIC trial this was reduced from 10.4 to 3.9 minutes using PAPNET, a statistically significant and practically important difference.

There are two important provisos to these findings. First, none of the included studies above refers to the only currently commercially available automated image analysis device, the AutoPap GS System. The majority of evaluations on test performance and impact on processing times have been performed on PAPNET. Second, detailed searches for UK unpublished literature on the test performance of automated image analysis revealed 13 studies, two of which appeared to be similar in quality to the studies included. This suggests that the findings are possibly highly susceptible to publication bias.

Concerning cost-effectiveness, although the DES model was developed, the authors were not satisfied with its validation. Given the possibility of equivalence of test performance, a cost-minimisation analysis was also used. This tentatively suggested that the AutoPap GS system may be efficient. The key proviso is that credible data become available to support that the AutoPap GS system has test performance and processing times equivalent to those obtained for PAPNET.

Conclusions

As in previous health technology assessments on this subject, the conclusion is that the available evidence on test performance, impact on process and cost-effectiveness is still insufficient to recommend implementation of automated image analysis systems. The priority for action remains further research. An important difference is that previously the insufficiency of evidence was general. Now, a general case for automated image analysis has probably just been made, but is specifically absent for the single device currently commercially available. The findings with respect to other and in many cases older automated image analysis devices need to be confirmed for the AutoPap GS System.

Implications for research on automated image analysis

The areas of greatest priority are:

- 'clinical effectiveness' of the AutoPap GS System relative to existing cervical screening programmes
- further development of the DES model presented in this report, particularly its validation
- further assessment of the cost-effectiveness of the introduction of automation alongside other approaches, including non-technological, to improving cervical screening
- further research on the effectiveness and costs of these other approaches.

Public research funding bodies should consider taking a greater lead in future research to ensure its independence and methodological rigour.

Implications for methodological research

There are many areas that may be pursued, in particular:

- research on the advantages and disadvantages of different research designs assessing the test performance of screening or diagnostic tests, especially two-armed designs
- research on the conduct of systematic reviews of dimensions of the impact of screening and diagnostic tests, other than test performance, especially their reproducibility and impact on process
- further research on publication bias, especially the role and conduct of detailed surveys for unpublished literature.

Chapter 1
Background

Summary of key points

Cervical cancer is a serious, but fortunately rare, disease. Screening for cervical cancer using the Papanicolaou (Pap) smear, to identify treatable preinvasive lesions, is proven to prevent cervical cancer. The programme costs, both financial and logistic, to achieve this have been considerable. Recently the strain on the NHS Cervical Screening Programme (NHSCSP) has become more acute, owing to a combination of media attention, pressure to reduce screening intervals, and difficulty recruiting and retaining screening staff at all levels.

New technologies appear to offer relief. This report considers automation of slide analysis in which slides are translated into computerised images, subject to analysis by computers, and a categorisation into normal or abnormal made. Automated image analysis is not currently intended to replace the manual system completely, but rather to work alongside it. Automated systems used in a primary screening mode are currently receiving the greatest attention. In this situation slides are analysed by an automated device before being passed for manual interpretation. The automated device identifies those slides that are very unlikely to contain abnormal cells, which may then not be examined manually at all ('archived') or just a random sample checked. There has been and continues to be considerable development of automated image analysis devices.

No country has so far universally implemented automated devices. A consistent concern has probably been the perceived high cost of automated devices relative to their benefits. In the UK in 2001 the cost of the only currently commercially available automated screening device was approximately £500,000 per device. Implementation costs for the NHSCSP would be considerable, £40 million being a very conservative estimate for England alone.

The use of technologies other than automated image analysis, such as liquid-based cytology (LBC) and human papillomavirus (HPV) screening, is also being actively considered by many cervical screening programmes, including the NHSCSP. These other new technologies, particularly LBC, may interact with automated image analysis.The role of non-technological approaches to improving cervical screening, such as increasing population coverage or addressing recruitment and retention issues directly, should not be overlooked.

Cervical cancer: realising the potential for its prevention through screening

The detection and treatment of cervical cancer remain an important area of activity for the NHS.[1] However, cervical cancer is a relatively rare cause of death in the UK (crude annual death rate 4.5 per 100,000[2]) and it is mass screening that has heightened the profile of the disease. This screening has a strong rationale. It usually takes 11–12 years for cervical dysplasia to progress to invasive neoplasia,[3,4] so offering the opportunity to detect and treat the disease in the preinvasive stage. Before the introduction of screening for cervical cancer in the mid 1950s, there were over 2500 deaths in England and Wales; now there are around 1200.[2] Although a downward trend was apparent before screening, recent cohort trends show that substantial numbers of cervical cancer deaths are being averted each year by the NHSCSP.[5,6] The rate of decline has increased since the late 1980s, which coincided with the greater coordination and quality assurance of the screening programme nationally, leading in particular to increases in coverage of the target population. The most recent figures indicate that in England about 85% of women aged 25–64 years had been screened once in the previous 5 years, with 84 out of 99 Health Authorities achieving the target coverage of >80%.[7]

However, realising the full potential of cervical screening has presented many difficulties not foreseen in the early optimism surrounding its introduction. Important among these is that the current screening test (the Pap smear) is relatively unsophisticated, which in turn has had far-reaching logistic and financial implications. It is

TABLE I *Interrelationships of different cytology classification systems*

Classification system	Level of abnormality							
Bethesda	Normal	Infection Reactive repair	ASCUS	Squamous intraepithelial lesion (SIL)				Invasive
				Low grade (LSIL) (includes HPV)	High grade (HSIL)			
Richart				Condyloma	Cervical intraepithelial neoplasia (CIN)			
				CIN I	CIN II	CIN III		
Reagen (WHO)		Atypia		Mild dysplasia	Moderate dysplasia	Severe dysplasia	Carcinoma *in situ* (CIS)	
UK system	Negative	Borderline (includes HPV)		Mild dyskaryosis	Moderate dyskaryosis	Severe dyskaryosis		Severe dyskaryosis/ ? invasive
Papanicolaou	I	II		III		IV		V

ASCUS, atypical squamous cells of undetermined significance.

hoped that automated techniques will help to alleviate the resulting burden.

The following paragraphs explore the nature of the current programme, the current problems with it, the nature of automation in cervical screening and the rationale that it may bring about improvements in the current system.

Current cervical screening in the NHS

Any test adapted for a mass-screening programme needs to be simple and reproducible from centre to centre without too great a variation in the results. Since Papanicolaou first realised the potential of studying the cytology of exfoliated vaginal scrapings,[8] it has been used as a method of prescreening for cervical cancer. Scrapes taken from the transformation zone of the cervix are smeared across a slide and stained with the Papanicolaou stain. These smears are initially examined by trained screening laboratory staff (termed cytotechnologists, or cytologists in the USA). Abnormal smears are also then examined by medically trained cytopathologists.

Histologically, cervical neoplasia is classified by a number of systems. One such system, used in the UK, is the cervical intraepithelial neoplasia (CIN) system. The grades, from I to III, represent increasing abnormality, with CIN III being considered a precursor to carcinoma.[9,10]

A screener will predominantly look for cells with abnormal nuclei known as dyskaryotic cells; severe dyskaryosis correlates with CIN III.[11]

Several different cytology classification systems are used worldwide and their interrelationships are illustrated in *Table 1*. This table is based on that reported by McCrory and colleagues,[12] but has been expanded to include the UK cytology system of classification.

In common with many screening programmes worldwide, the NHS relies on analysis of Pap smears.[13] All women aged 20–64 years are eligible for free smears. In response to letters of invitation or reminders sent out by Health Authorities, cervical smears are taken in primary care. These smears are analysed in laboratories. Initially, all smears are screened by primary screeners. A different screener then performs a quick review of normal and inadequate slides, before reporting, and those with an initial diagnosis of abnormal are reviewed by a senior biomedical scientist; all confirmed abnormal slides are then passed to a cytopathologist for review and reporting (*Figure 1*). Non-medically trained 'advanced practitioners' also increasingly review and report abnormal smears in the NHSCSP. Accuracy and quality control of slide analysis are achieved by monitoring the sensitivity of primary screening with respect to the final report issued by the laboratory and the positive predictive value (PPV) of moderate dyskaryosis or worse in relation to histological findings.[14]

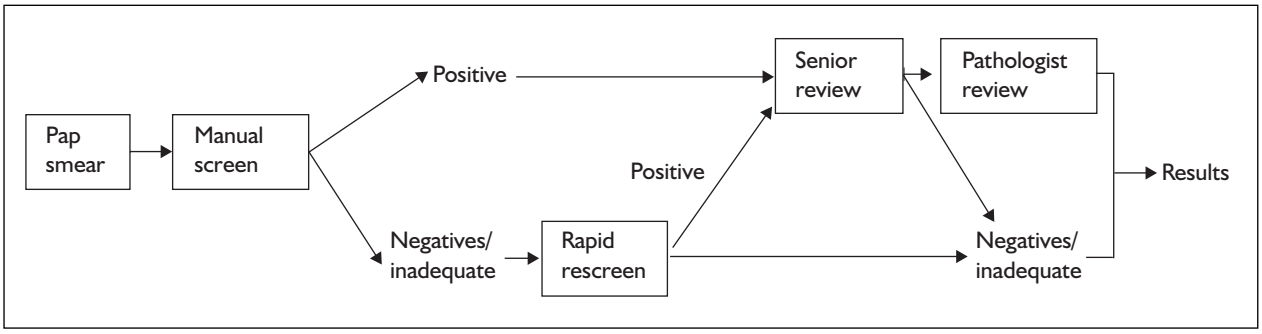


FIGURE 1 *Current system of screening in the NHS (manual screening)*

Management protocols define further action depending on the current and previous result (in the case of inadequate, borderline or mild dyskaryotic findings). Women with smears reported as normal continue with routine screening, whereas those with abnormal smears require either cytological surveillance in the case of inadequate/borderline/mild dyskaryosis or immediate referral for colposcopy for more severe lesions. Cytological surveillance continues for up to three abnormal smears before referral.[14]

The optimal frequency of screening where a normal result is obtained remains uncertain. A mathematical model applied to screening data showed that if the screening interval was 10, 5 or 3 years, the incidence of cervical cancer would fall by 64%, 84% and 91%, respectively, but only an additional 2% decrease would be achieved by annual screening.[15] In consequence, the actual intervals between routine invitations as stated in the NHSCSP guidelines should be no sooner than 3 years and at least every 5 years.[16,17]

Problems with the current NHS cervical screening programme

Despite undoubted success and improved standards of service provision brought about by the quality assurance programme introduced in 1988, many areas remain where programme performance could be improved. For instance, those most at risk of cervical cancer are those least likely to be screened. Around 67% of cervical cancer deaths each year occur in women who have never been screened or have not been screened for more than 5 years, and women with lower socio-economic status are both at the highest risk of abnormalities and the least likely to have the benefit of regular screening.[18] Beyond increasing coverage in those at highest risk, other areas that have been highlighted include improving coverage in general over a 3-year, as opposed to a 5-year period, and increasing the quality of smear-taking to reduce rates of inadequate smears, which were in excess of 390,000 in 2000/01.[7]

Of greatest current concern, however, is the ability of laboratories to cope with the formidable task of manually examining and accurately reporting very large numbers of slides, 4.1 million each year in England and Wales.[7] This demand is likely to rise further with increasing rates of inadequate smears and moves from 5-yearly to 3-yearly routine screening cycles. Additional strain arises from public perception that no cancers should develop if screening is 'done properly', despite knowledge that even with high-quality screening this cannot be achieved. Slides with small numbers of abnormal cells are a significant cause of screening 'misses' in any programme, as slides with small numbers of abnormal cells are more difficult to recognise than those with larger numbers of abnormal cells.[19]

A recent NHS HTA programme-sponsored systematic review of the impact of false negatives in screening programmes reported that although the evidence was mainly anecdotal, there was a consensus opinion that false negatives had a negative impact on public confidence in screening, particularly when the individuals affected took legal action and there was associated publicity.[20] The general problem of false negatives across all screening programmes is not lost on the Medical Defence Union, which in October 1998 had 82 claim files involving false negatives between 1990 and March 1998, with contingency reserves on active files amounting to £1.1 million.[20] Unfortunately, legal proceedings and out-of-court settlements will follow interval cancers irrespective of whether quality standards were met or not.

Screening false negatives that lead to the identification of the responsible screeners can have an impact on the morale and confidence of the other screeners. This was found to be the case when the identification of a substandard screener resulted in the overcalling of inadequate, borderline and mild smears above acceptable levels in the subsequent years; a loss of confidence in the other screeners was considered to be the cause.[20] It is not surprising therefore that recruitment and retention of screening staff are proving increasingly difficult, with many laboratories having insufficient staff, non-medical and medical, to cope with present workload. This leads to a backlog of tests, long waits for women to receive their routine test results, and further demoralisation for staff, who also respond by lowering thresholds for declaring slides abnormal, which further exacerbates the situation. These problems of meeting demand for the processing of slides have refocused attention on the shortcomings of the Pap smear itself.

Limitations of the current slide preparation and analysis

As indicated above, apparent errors, particularly false negatives, cause great concern. Many errors are outside the control of any slide analysis system. Such situations occur where the cervical malignancy develops rapidly or in an occult manner, so that any previous smear does not actually contain cells suggestive of malignancy, or the area of malignancy is not sampled by the smear-taker, or a woman developing cervical cancer has not actually had a smear taken. However, even considering errors directly referable to slide analysis, what is not clearly understood by public, politicians and healthcare professionals is that even when performed to the highest standard, few if any tests are perfect. A systematic review and meta-analysis of 59 studies published to August 1992 which compared the results of Pap smear with histology demonstrates this well.[21] Not only did no included study achieve a false-negative rate of zero (i.e. a sensitivity of 100%), but the highest levels of sensitivity (around 95%) were only achieved with low specificity (around 15%). A similar meta-analysis, published more recently, included 25 additional studies to the above, and demonstrated a combined sensitivity and specificity of 51% and 98%, respectively.[12] Thus, there is an inevitable trade-off between capturing all the important abnormalities actually present in a slide and misidentifying cells that are actually normal. Three factors that may contribute to this are the nature of the slide, the human visual system and fatigue.

- The nature of the slide: even the best Pap smears only present for laboratory examination a small proportion of the material originally harvested, and in a way that may obscure abnormalities actually present because of uneven distribution of material over the slide and the presence of non-epithelial elements such as polymorphs, red blood cells and cellular debris.
- Visual system flaw: an understanding of the human visual system and its associated flaws explains why a manual screening programme may never be totally effective.[22] Screening a slide involves searching for a target object on a background of a large number of distracting objects. Data from psychological literature[23] show that searching for objects with deviating stimuli when the background contained standard objects was more successful than vice versa. Thus, it appears that the human eye is particularly adept at picking out large objects on a background of small objects, or elliptical objects on a background of circular objects. Problems arise when the features of the target object closely resemble those of the distractors, as in the case of pale dyskaryosis or small cell cervical dysplasias.[22] In this instance the distractors (normal cells) are of a similar size to or even larger than the target objects, and the normal cells may be equally pale.
- Fatigue: any repetitive process allows the opportunity for deterioration of performance with increasing length of sessions. Biologically based detection systems are much more susceptible to fatigue of this sort, a fact recognised in the quality assurance of the cervical screening programme by restricting the maximum consecutive length of time that screeners can perform primary screening of slides to 4 hours.

Improving slide preparation and analysis using automation and related technology

The limitations of current techniques have suggested and provided a spur to the development of a number of new approaches directly or indirectly involving automation.

Automated slide analysis devices [PAPNET (Neuromedical Systems Inc.),[24] AutoPap (TriPath)[25] and AutoCyte SCREEN (AutoCyte Inc.)[26]]

Three different devices have been developed, the aim of which is to identify abnormal cells on a slide prepared in the current fashion, or using thin-layer techniques (see below), and so reduce reliance on the human eye as the only means of slide analysis. In essence, each converts the image of a slide into a format that can be interpreted by a computer, and this is then analysed for patterns that denote abnormal cells. The precise means by which this is achieved in each system is complex. Some of the key principles, highlighting differences between systems, are indicated in the following paragraphs.

Although capturing an image of a slide that optimises the contrast between the cells and the other objects presents its own problems,[27] the basic approach to generating something that can be analysed by computers is to create a digitised file. The digitised image consists of a grid of pixels, which when in grey scale have a single numerical value that is represented in a computer's memory. The value of each pixel belongs to a definite range, 0–255 being common in the earlier systems. Each number corresponds to a particular shade of grey; the two extremes, black and white, are represented by 0 and 255, respectively.

The next step, separating (segmenting) abnormal cells from normal cells and non-cell artefacts, has been much more problematic.[27] The earlier attempts used crude tools often revolving around a histogram of grey levels. The histograms were constructed by quantifying the frequency that each grey level appeared in an image. Prewitt and Mendelsohn first realised that the peaks on the histogram corresponded to the nucleus, cytoplasm and background of an image.[28] Using this they set the first minimum point as a global threshold so that in principle the darker pixels of the nuclei remained. Unfortunately, the histograms were not always trimodal. Other attempts smoothed the histogram of any fluctuations before calculating a minimum.[29]

More advanced approaches realised the limitations in thresholding, and used edge detection[30] and contour tracing[31,32] to segment the image, but spurious edges and objects made artefact rejection more difficult. The advent of colour images provided computers with three times more information since each pixel now carried three numbers, one for each of the channels, red, green and blue. This did not necessarily improve matters as noise levels were correspondingly increased.

Some of the more modern systems continue to use low-level programming to extract relevant features, with advancing hardware allowing an increasing number of techniques and calculations to be performed (AutoPap and AutoCyte SCREEN). However, Neuromedical Systems Inc. (NSI) developed a different approach. The PAPNET combined the speed of low-level programming with the decision-making of high-level programming techniques called neural networks. These simulate humans in their ability to learn from experience, and are trained on test sets of images. Neural networks, unlike standard software, are restricted not by the predetermined rules of the program, but by the extent of the variation in the test set.

The distinction between these different automated imaging devices has changed dramatically from when they were first introduced, with a subsequent rationalisation of the technologies available. NSI ceased to trade in 1999, and although the intellectual property of the PAPNET was sold to AutoCyte Inc., the device is no longer available.[33] Furthermore, Neopath Inc., the original producers of the AutoPap device, merged with AutoCyte Inc. to form TriPath. This effectively removed the AutoCyte SCREEN system from the marketplace.[34] Currently, there is thus only one commercially available automated slide analysis device, the AutoPap Guided Screening (GS) System. Even this has undergone (and continues to undergo) considerable development and modification from earlier versions, the AutoPap QC 300 and the AutoPap Primary Screener.

Alternative methods of deploying automated image analysis

In addition to the nature of the automated technology, the way in which it is integrated with existing systems can vary. Although there is a very large number of permutations, two are particularly worth differentiating:

- Use of image analysis as a *primary screening* device: in this modality, all slides are screened by the automated system and an initial diagnosis is made before the suspicious or abnormal slides are passed for review by the manual system. Slides diagnosed as normal by the new technology are usually either archived or submitted to some quality control procedure such as rapid review. In the USA only the AutoPap has gained Food and Drug Administration (FDA) approval for use in this way.

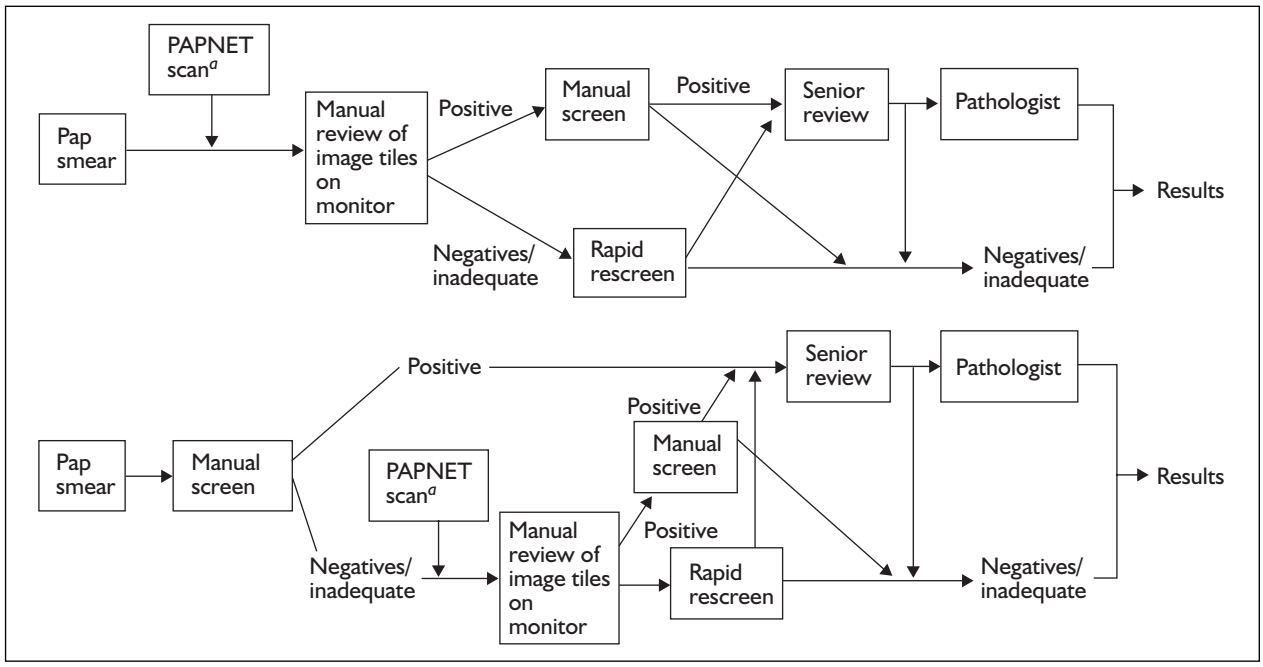


FIGURE 2 *Contrasting PAPNET used as a primary screening system (upper diagram) with PAPNET used in quality assurance mode (lower diagram). [a]A PAPNET scan comprises 164 images or 'tiles' of the most potentially abnormal cells on a cervical smear.*

- Use of image analysis as an aid to *quality assurance*: in this instance all slides are initially reviewed by the existing manual screening system and only those slides reported as negative or inadequate are reviewed by the computerised system. The AutoPap has obtained FDA approval for use in this way, as had the PAPNET, before NSI ceased to operate. *Figure 2* illustrates the two main alternative ways of deploying automated devices, using PAPNET as the example.

Since there are several possible ways of deploying an automated device in a primary or quality control screening system, each possibly having a different impact on the overall performance of the device, manufacturers now indicate the intended use of their system more precisely. Thus, the intended use workflow of the new AutoPap GS System as published by the manufacturer[35] is shown in *Figure 3*.

Automated slide preparation techniques [ThinPrep (CYTYC),[36] AutoCyte PREP, previously known as CytoRich (AutoCyte Inc.)[37]]

In these techniques, after the sample has been obtained, the specimen collection device is thoroughly rinsed into preservative fluid, rather than the specimen being smeared directly onto a slide. This produces a suspension of cells that are then filtered before a slide is made. In this way, more of the cells of cervical origin are captured, extraneous material (blood, pus and mucus) can be excluded and slides composed of uniform monolayers created. A report produced recently on behalf of the National Institute for Clinical Excellence (NICE) showed that although it was difficult to quantify owing to a lack of primary data, it is likely that the rate of inadequate slides would be reduced, specimen interpretation times would be shortened and the technique could lend itself to automated image analysis systems.[38] Pilot studies evaluating LBC in the NHS are underway and are expected to be reported upon in June 2002 (Winder R, NHSCSP: personal communication, 19 December 2001). Anecdotal reports have been positive, but cautious notes have also been voiced on the efficacy of LBC.[39] The results are now available: http://www.cancerscreening.nhs.uk/cervical/lbc.html (accessed 10 January 2005).

HPV screening and other technologies

Other potential advances have emerged in parallel with the development of automation. Chief among these is screening for HPV, where hybridisation and immunoassay techniques are used to locate particular types of HPV in cervical cells collected in fluid-based samples. There has been optimism that such a technique used in conjunction with existing systems could improve the performance of cervical screening programmes.[40] A comprehensive review demonstrated that although no definitive conclusions could be drawn about the clinical effectiveness of HPV testing, there were areas where

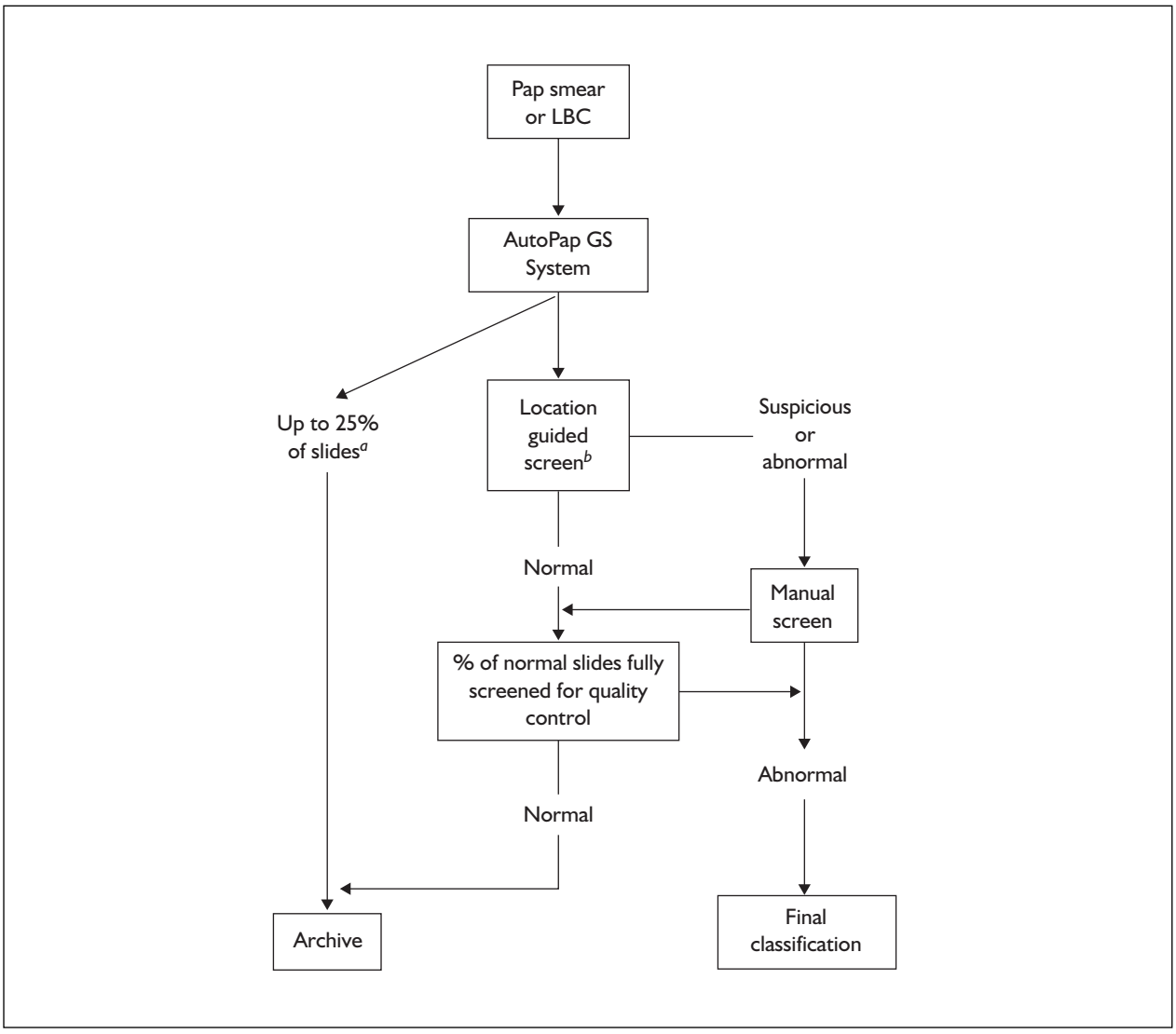


FIGURE 3 *Specified system of deployment for the AutoPap GS System.* [a]*AutoPap ranks slides according to likelihood of abnormality; the slides with the lowest ranks are those designated 'Not for review' and archived.* [b]*The AutoPap system indicates the location on the slide of the cells most likely to be abnormal.*

it could be deployed, most notably in the management of borderline and low-grade smears.[41] This has provided the impetus for UK-based studies evaluating HPV (results now available: http://www.cancerscreening.nhs.uk/cervical/hpv.html, accessed January 2005).[42] Like LBC, there is clear potential for HPV screening to interact with automated image analysis and its parallel development needs to be borne in mind.

Beyond HPV there appears to be a number of other technologies that may have an impact on cervical screening in the future. These include:

- immunostaining of smears[43]
- cervicography (high-quality colposcopic-type photographs of the cervix)[44]
- speculoscopy (examination of the cervix using specialised chemiluminescent light along with acetic acid and low power)[45]
- Polarprobe (Polartechnics Ltd, Australia) (an opto-electronic instrument that detects the existence of cervical precancerous and cancerous lesions by measuring voltage decay and scattering of various wavelengths of light).[46]

Although of possible interest as potential comparators for automated image analysis as means of improving current cervical screening programmes or technologies that may be combined with automation, they are not considered in detail further in this report for practical reasons.

International situation with respect to automated devices

Automation has been considered by several countries. To the authors' knowledge no country has adopted automated screening across an entire national programme, although in The Netherlands 25% of all smears taken are screened by the PAPNET, despite it being no longer commercially available (Boon ME, Leiden Cytology and Pathology Laboratory, The Netherlands: personal communication, 10 January 2002). In the other countries in Europe (e.g. Republic of Ireland, Switzerland, Germany, Italy and Denmark) individual laboratories possess automated systems, but this does not extend across the respective national health systems.

The same may be said of North America, and although there are larger numbers involved [45 laboratories in the USA have a version of the AutoPap (Holt P, TriPath Imaging Inc., USA: personal communication, 17 December 2001)] this is small in comparison to the total number of laboratories. In Australia and New Zealand automated technology has also been evaluated, but neither country endorsed its use. Both countries continue to rely in the main on conventional manual screening.

Irrespective of the country and the degree of uptake, the perceived high cost of automated devices relative to the benefits that may accrue has probably been the main barrier to dissemination of the technology. The device cost in the UK in 2001 for the only currently commercially available automated device, the AutoPap GS System, is in excess of £0.5 million per machine (Jackson AK, CellPath plc, UK: personal communication 14 August 2001). Accepting the manufacturer's claims that 50,000 slides per annum can be processed by each machine, approximately 80 machines would be required to process the current slide throughput for the NHSCSP (approximately 4 million). The substantial implementation costs inevitably raise questions about whether for the same investment, the benefits associated with automation (or indeed other types of new technology) may be more or less than those associated with investment in non-technical initiatives to increase coverage or even, particularly in the context of the NHS, improved pay and conditions for screening staff to improve recruitment and retention.

Chapter 2

Aim and objectives of the health technology assessment project

Summary of key points

The project has a number of components, the methods and results of which are described separately in the following chapters. Each component feeds into the overall conclusions in a defined way.

The overall objective of the project was to assess the immediate effects, the wider consequences and costs, and overall cost-effectiveness and cost–utility of introducing automated image analysis to a screening programme with characteristics similar to those currently operating in the UK.

In this way the project aimed to aid decision-making on whether automated cervical screening devices should be used in the UK and to indicate areas needing further research.

Introduction

This chapter summarises the aim of the project overall, and the specific objectives of the individual components of the health technology assessment, the methods and results of which are presented separately in each of the succeeding chapters. In addition, an overview is given of how each of these components interrelate and contribute to the overall conclusions.

Overall aim and objective

The overall objective was to assess the immediate effects, the wider consequences and costs, and overall cost-effectiveness and cost–utility of introducing automated image analysis to a screening programme with characteristics similar to those currently operating in the UK. In this way the project aimed to aid decision-making on whether automated cervical screening devices should be used in the UK and to indicate areas needing further research.

In this report, automated cervical screening technology means any automated image analysis device, particularly AutoPap, PAPNET and AutoCyte, as described in Chapter 1, used in either a primary screening or quality assurance mode. However, of these it should be immediately noted that the only currently commercially available device is a recent version of the AutoPap, which is specifically designed to operate in a primary screening mode. Automated slide preparation devices (LBC) are not the main subject of this report and are referred to as 'related technology', as is HPV screening. Information on them is included, because the report anticipates that LBC and HPV screening may become components of the NHSCSP in the foreseeable future.

Cervical screening programmes with characteristics similar to those currently operating in the UK are taken to mean population-based screening programmes based on Pap smears taken routinely on women in the approximate age range 20–64 years at 3–5-yearly intervals, the smears being analysed by a manual system involving scrutiny of stained smears by trained individuals using light microscopy. More detail on the precise components of this are provided in Chapter 1. As above, it should be noted that this report anticipates that the nature of current NHSCSP may change in the future, particularly with respect to the incorporation of LBC and HPV screening.

Specific components of the health technology assessment project and their objectives

Systematic review of the secondary literature on clinical effectiveness of automation and related technologies

Based on recent systematic reviews and health technology assessments, taking into account their strengths and weaknesses, the objective was to answer the question, 'What do we already know about the effects and effectiveness of the different new technologies that have been applied to cervical cancer screening?'

Systematic review of the literature on cost-effectiveness of automation

The main objective was to answer the question,

'What do we already know about the cost-effectiveness and cost-utility of automated cervical screening?' More specifically, the project also aimed to identify whether there was variation in cost-effectiveness or cost–utility results for automation, and if so, to use information provided in the published studies to explore the reasons for such variation.

Evaluating automated cervical screening: methodological issues

A key principle adopted in conducting this project was to build on past assessments, and where possible attempt to overcome problems identified with their conduct. This section of the report explores in greater detail the issues identified in the two preceding sections and indicates how the report attempted to address them. If the problems were such that they could not be dealt with in the context of this project, they were carried forward as issues requiring further research.

Systematic review of the effects and effectiveness of automation

Based on a systematic review of the primary literature, the objective was to assess the question, 'What are the effects and overall effectiveness of the introduction of automated cervical screening devices?'

Likelihood of publication bias in reviews of research on the effects of automation

A priori, a concern was identified that unpublished research literature on the effects and effectiveness of automated cervical screening technologies may be particularly prevalent, and that reviews on effectiveness may be susceptible to publication bias as a result. Thus, the objective was to answer the question, 'With respect to reviews of evidence on the effects of automation, what risk to the validity of the overall conclusions is posed by unpublished research (particularly research that has been stopped prematurely)?'

Costs of automated cervical screening devices

The main purpose of this component was systematically to collect source data for the health economic model. To this end, based on primary and secondary research by other investigators and enquiry of manufacturers, the objective was to answer the question, 'What are the costs likely to be associated with the introduction of automated cervical screening devices to programmes similar to those operating, or likely to be operating in the UK?'

Modelling the health economic impact of automation

Based on a *de novo* model, from the viewpoint of the NHS and patients, the objective was to assess the question, 'What is the cost-effectiveness of introducing automated cervical screening devices to programmes similar to those operating, or likely to be operating in the UK?' In this an attempt was made to integrate much of the information on effects and cost identified in the preceding components of the project. However, recognising that uncertainty concerning key data may be a major factor limiting conclusions, ancillary objectives were to develop a model that could incorporate information from future research, and to highlight those parameters that seem to have the greatest influence on cost-effectiveness.

Interrelationship of the specific components of the project

Figure 4 is a schematic representation of how the components of the project and the chapters in which their methods and results are reported impinge upon each other.

In essence, the review of the secondary literature informed the development of the original protocol and helped to shape the approach used in each of the other components of the project. As well as providing necessary input data to the simulation model, each of the main chapters supports conclusions in its own right, including recommendations for further research. Such conclusions are presented at the end of each of the chapters and summarised and integrated in the final chapter, giving the overall conclusions of the project as a whole.

General method

The conduct of the health technology assessment project was based on a predefined protocol contained in Section 3 of the full submission to support the bid for this research, prepared in July 1999. There were no major departures from this protocol. Less emphasis has been placed on some components of the project, particularly directly surveying for information on cost, and greater emphasis on others, such as a specific investigation of the likelihood of publication bias, in the final project compared with that originally anticipated. In addition, the focus of what constituted automation had to be

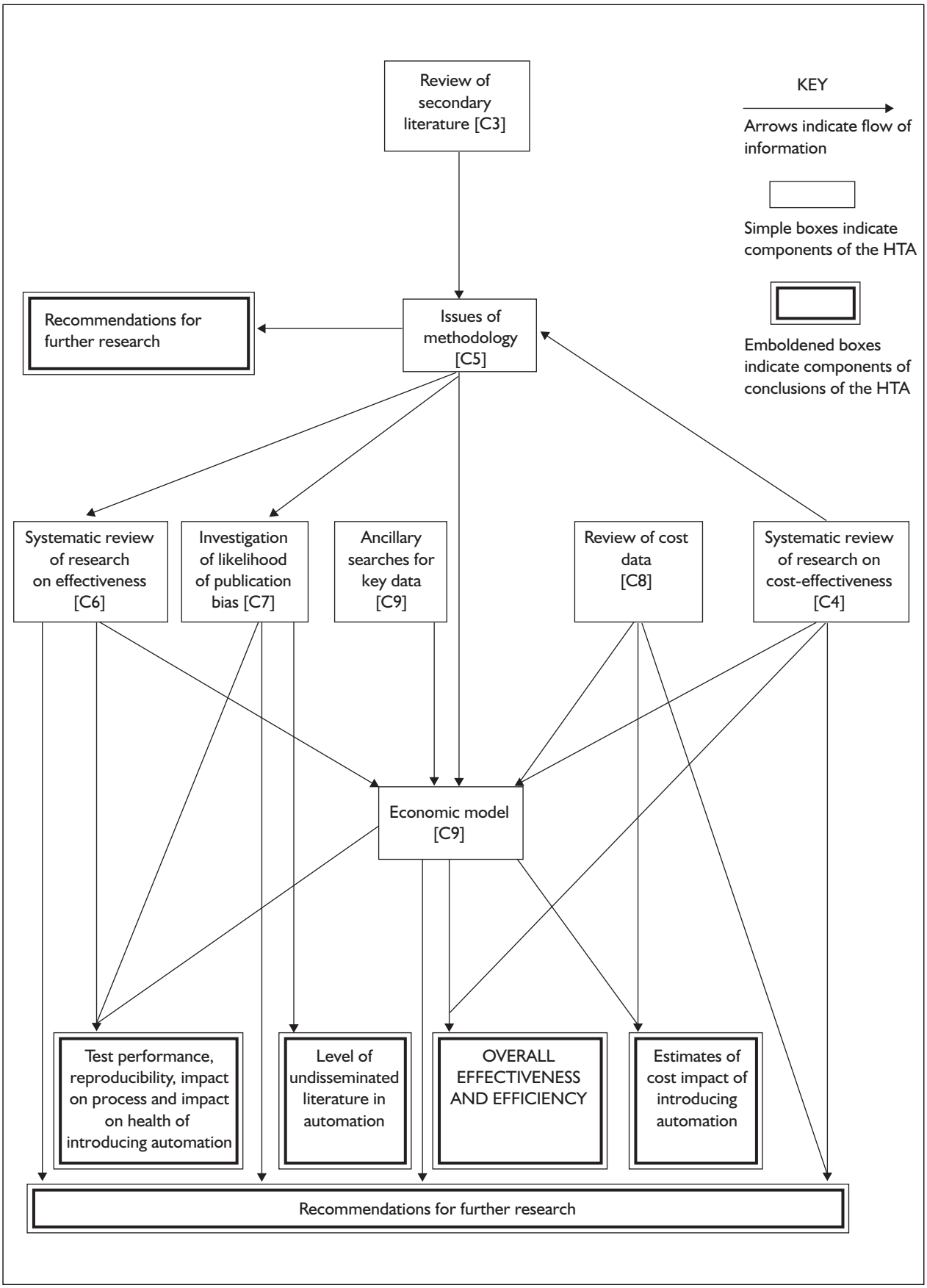


FIGURE 4 *Interrelationship of the different components of the health technology assessment project and how they contribute to its conclusions [chapter numbers in brackets]*

modified to take into account work done between the original submission and the start of the project, although this was anticipated and planned for in the original protocol. The key change in this respect was greater emphasis on automated image analysis in cervical screening and less on automated slide preparation, as the latter had recently been dealt with in guidance from NICE.[38,47] Further, as already indicated, the range of commercially available automated image analysis devices had dramatically decreased. However, in this respect the original intention to consider all devices on which there was research information was maintained, particularly with respect to the systematic review of effectiveness. The need to take care in the generalisation of results as indicating the effectiveness of those actually available was, however, noted.

In general, the methods used in the project were based on standard approaches to systematic reviewing and health economic evaluation defined in recommended texts.[48–51]

Chapter 3

Systematic review of the secondary literature on clinical effectiveness of automation and related technologies

Summary of key points

The main objective was to evaluate the most recent reviews of clinical effectiveness in this field and so define the starting point for knowledge on this topic and help to develop the protocol for the project.

Literature on automated cervical cytology and related technologies (LBC and HPV) was systematically searched for the period from 1 January 1997 to 31 December 2000. Literature prior to this date was felt to be acceptably covered by two health technology assessments published in 1997 and 1998.

The searches identified 44 studies, 16 of which were included, and 13 are appraised in this chapter. Three dealing with cost-effectiveness/modelling, without making a detailed consideration of clinical effectiveness, are considered in Chapter 4.

All reviews of clinical effectiveness had some weaknesses judged against the current methods standards for undertaking systematic reviews. More recent reviews by national health technology assessment groups had fewer weaknesses and were judged to provide more internally valid conclusions on the clinical effectiveness of automation and related technologies.

Based on these systematic reviews, past findings have been that the evidence on the clinical effectiveness of automated image analysis is too weak to justify implementation. However, the technology has been developing and new evaluations have been published since the conclusion of the most recent systematic reviews, confirming that this project remained highly relevant. Indeed, the pace of change is such that further re-examination of the clinical and cost-effectiveness of automation will be required in the future.

Automated image analysis, LBC and HPV testing are complementary, and the potential for interaction should be considered in evaluating effectiveness (and efficiency). However, improved effectiveness (and efficiency) may potentially be achieved without resort to technological advance. It is notable that there appear to have been no attempts to measure or model the relative effectiveness of technological approaches to improving cervical screening programmes with non-technological approaches.

Introduction and objectives

The field of automation in cervical screening has been the subject of a number of publications and reviews in the past decade, and in the past few years in particular this has gathered pace. Therefore, the first task was to appraise the work of others who have reviewed new technologies in cervical screening. Such an approach provided the opportunity to build on the work of previous investigators and to this end the objectives for this component of the health technology assessment project were:

- to confirm that the original protocol was still relevant
- to amend or develop, as appropriate, the framework for the project
- to indicate the expected standard of evidence that would be located from the searches
- to highlight methodological issues that are special to this field
- to summarise the key results from previous work
- to put the potential for improvement of cervical screening through automation of image analysis in the context of other ways of improving cervical screening.

The last objective was particularly important, as the future of other technologies is also currently being considered in the context of the NHSCSP.

It is highly likely that if automation were introduced, it would be part of a cocktail of technologies involving LBC and HPV testing. Although these in particular constitute the other technologies considered in this section, the possibility that other approaches not involving technological advance may prove effective and efficient means to improve existing cervical screening programmes should not be overlooked.

Method

Search

The search was particularly aimed at retrieving secondary research, but was otherwise kept as general as possible (see search algorithms in Appendix 1). The search was restricted to the period from 1 January 1997 to 31 December 2000. Two major reviews extensively covered the field for the period up until 1997,[52,53] and these are used to represent the literature before the search period. The details of the sources of information searched are given below.

Bibliographic databases searched:

- Cochrane Library (2001, issue 4) [includes Database of Abstracts of Reviews of Effectiveness (DARE)]
- MEDLINE (January 1997 to December 2000)
- Health Services/Technology Assessment Text (HSTAT)
- National Research Register (NRR).

Websites and publication lists of key organisations:

- NHS Centre for Reviews and Dissemination (NHS CRD) (particularly NHS CRD reports and Effective Health Care Bulletin) (www.nhscrd.york.ac.uk, accessed February 2001)
- National Coordinating Centre for HTA (www.ncchta.org, accessed February 2001)
- NICE (www.nice.org, accessed February 2001)
- Agency for Healthcare Research and Quality (AHRQ) [formerly Agency for Health Care Policy and Research (AHCPR)] (www.ahrq.gov, accessed February 2001)
- Canadian Co-ordinating Office for Health Technology Assessment (www.ccohta.ca, accessed February 2001)
- Swedish Council on Technology Assessment in Health Care (www.sbu.se, accessed February 2001)
- New Zealand Health Technology Assessment (www.nzhta.chmeds.ac.nz accessed February 2001)
- Medical Service Advisory Committee [formerly Australian Health Technology Advisory Committee (AHTAC)] (www.health.gov/msac, accessed February 2001).

Direct contact with:

- UK National Cervical Screening Programme
- New Zealand Health Technology Assessment
- Canadian Co-ordinating Office for Health Technology Assessment
- Manufacturers (TriPath, Cytyc).

Inclusion criteria

A new technology was considered to mean one of following:

- an automated or semiautomated screening system
- LBC
- HPV testing.

Automation is confined to image processing devices, which use diagnostic software to facilitate the diagnosis of a slide.

A study was included if it was a systematic review of the clinical effectiveness of one of the new technologies applied to cervical screening. A systematic review is a review that uses recognised systematic review methodology.[54]

Systematic reviews of the cost-effectiveness and *de novo* models of cost-effectiveness of one of the new technologies applied to cervical screening were also counted as included, but are generally considered as part of the systematic review of past attempts to assess and model cost-effectiveness reported in Chapter 4.

Inclusion and exclusion decisions were carried out by one reviewer (BW). Brief details of excluded studies and reasons for their exclusion were recorded, and are presented in Appendix 2.

Included study appraisal and analysis

The appraisal of systematic reviews in the majority of cases involves the application of checklists, which test whether individual criteria have been met.[55] The most comprehensive of these is the Quality of Reporting of Meta-analyses (QUOROM) statement,[56] which is aimed at systematic reviews of randomised controlled trials (RCTs). Here, the reviewers decided to take a step back from the checklists and identify the necessary set of objectives that should be met by all systematic reviews, regardless of the study topic. Such an approach allows flexibility in the design of

TABLE 2 *Systematic review of the secondary literature: results of search*

Source	All retrieved	Included	Excluded
MEDLINE	23	9	14
Cochrane Library	20	6	14
Other	1	1	0
Totals	44	16	28

TABLE 3 *Systematic review of the secondary literature: subquestions to which included studies contribute*

Study	Technology	Clinical effectiveness	Cost-effectiveness	Model
Abulafia and Sherer, 1999[57]	Automation	✓		
Abulafia and Sherer, 1999[58]	Automation	✓		
Noorani *et al.*, 1997[52]	Automation	✓		✓
Mango and Radensky, 1998[59]	Automation	✓		
Radensky and Mango, 1998[60]	Automation			✓
Smith *et al.*, 1999[61]	Automation			✓
AHTAC, 1998[53]	Automation, LBC	✓	✓[a]	✓[a]
Broadstock, 2000[62]	Automation, LBC	✓	✓[a]	
Brown and Garber, 1999[63]	Automation, LBC	✓		✓[a]
McCrory *et al.*, 1999[12]	Automation, LBC	✓	✓[a]	✓[a]
Myers *et al.*, 2000[64]	Automation, LBC			✓[a]
Nanda *et al.*, 2000[65]	Automation, LBC	✓		
Austin and Ramzy, 1998[66]	LBC	✓		
Payne *et al.*, 2000[38]	LBC	✓	✓[b]	✓[b]
Cuzick *et al.*, 1999[41]	HPV	✓	✓[b]	✓[b]
Cuzick *et al.*, 2000[67]	HPV	✓		✓[b]
Totals	16	13	5	10

[a] Cost-effectiveness of LBC not considered in detail in main health technology assessment.
[b] Not considered further in main health technology assessment. Reference made to the articles in Chapter 4 from the perspective of the methods of modelling used. Key details provided in appendices.

checklists, which should be seen as a set of questions aimed at testing whether the objectives have been met. These objectives are labelled under the following headings:

- objective
- completeness
- accuracy
- inference
- reproducibility.

Further details on this can be found in Appendix 3, where further discussion is given on the limitations of currently available quality assessment tools for systematic reviews in this sort of situation.

The method of analysis was qualitative, relying on clear tabulation and presentation of characteristics, study quality and findings to generate conclusions and implications.

Results

Yield of searches

The results of the searches are shown in *Table 2*. In total, 44 studies were retrieved from the searches and 16 satisfied the inclusion criteria. However, as *Table 3* shows, three of these studies only contributed information on cost-effectiveness or a *de novo* model, and so are not represented in this chapter,[59,60,64] which considers the 13 studies assessing clinical effectiveness.[12,38,41,52,53,57–59,62,63,65–67]

Results of the appraisal and analysis: systematic review quality

The 13 contributing reports all had some weaknesses relative to gold-standard systematic review technique (*Table 4*). Full details of data abstracted are provided in Appendix 4. Some failed to convince that most relevant articles had been located, either by not documenting the

TABLE 4 *Summary of the included reviews of clinical effectiveness*

Study	Subject	Objective	Search	Inclusion criteria	Quality assessment	Analysis
Abulafia and Sherer, 1999[57]	Auto (AP)	Clear	Poorly documented search strategy; MEDLINE only	Not explicitly stated; English language only	Little attention to nature of reference standard	Thresholds for disease and diagnosis not defined; specificity not considered
Abulafia and Sherer 1999[58]	Auto (PN)	Clear	Poorly documented search strategy; MEDLINE only	Not explicitly stated; English language only	Little attention to nature of reference standard	Thresholds for disease and diagnosis not defined
AHTAC, 1998[53]	Auto (AP, PN, AC), LBC	Clear	Well documented, but searches mainly MEDLINE	Not explicitly stated; probably English language only	Considered in detail	Limited by availability of data
Austin and Ramzy, 1998[66]	LBC	Clear	Minimal documentation of search strategy	Minimal	Not done	Thresholds for disease not defined
Broadstock (NZHTA), 2000[62]	Auto (AP), LBC	Clear	Well documented and comprehensive	Explicitly stated, but restrictive; English language only	Considered in detail	Limited by availability of data
Brown and Garber, 1999[63]	Auto (AP, PN), LBC	Clear	Well documented, but searches mainly MEDLINE	Explicitly stated; English language only	No explicitly stated criteria for assessing the study quality	Thresholds for diagnosis not defined Specificity not considered
Cuzick *et al.*, 1999, 2000[41, 67]	HPV	Clear	Generally well documented and comprehensive	Explicitly stated; English language only	Given some attention but results not made explicit	Full and considered analysis
Mango and Radensky, 1998[59]	Auto (PN)	Clear	Poorly documented search strategy; searches mainly MEDLINE	Not explicitly stated; not clear whether English language restriction	Reference standard considered only	Specificity not considered
McCrory *et al.*, 1999[12]	Auto (AP, PN), LBC	Clear	Well documented and comprehensive	Explicitly stated, but restrictive; English language only	Considered in detail	Full and considered analysis
Nanda *et al.*, 2000[65]	Auto (AP, PN), LBC	Clear	Well documented and comprehensive	Explicitly stated, but restrictive; English language only	Considered in detail	Full and considered analysis

continued

TABLE 4 *Summary of the included reviews of clinical effectiveness (cont'd)*

Study	Subject	Objective	Search	Inclusion criteria	Quality assessment	Analysis
Noorani *et al.*, 1997[52]	Auto (AP, PN)	Clear	Well documented and comprehensive	Not explicitly stated; English language only	Considered in detail	Limited by availability of data
Payne *et al.*, 2000[38]	LBC	Clear	Well documented and comprehensive	Explicitly stated	Considered in detail	Thresholds for diagnosis not defined Otherwise a full and considered analysis

Auto, automated image analysis; AP, AutoPap; AC, AutoCyte SCREEN; PN, PAPNET.

search strategy fully[41,57–59,66] or by searching too few databases to ensure completeness.[57,58,63] Others had vague inclusion criteria.[52,53,57–59]

The reviews having the fewest weaknesses were generally carried out by the different health technology assessment bodies (or equivalent) around the world. Recent reviews[12,38,41,62,65] had fewer weaknesses.

All the included reports restricted included studies to those in the English language, and so the possibility of language bias in the context of systematic reviews in this field has never been explored.[68] Further, although a number of reviews searched sources that may identify unpublished data, none has specifically tried to assess the degree to which publication bias may be operating.

Although not a weakness as such, the more recent reviews have seemingly become a victim of their striving for increased rigour.[12,62,65] That inclusion criteria should be clearly defined and ideally applied as an objective algorithm to minimise selection bias is undeniable. However, if the criteria are too restrictive then few and sometimes no studies are included for appraisal, limiting not only conclusions but also material that can be used to demonstrate why studies perceived to be suboptimal are likely to be misleading. This issue is discussed in more depth in Chapter 5 on methodological issues.

Variation in the assessment of quality of studies included in the included reviews was also noted. For instance, a number of reviewers[12,53,62,65] highlighted problems with the case-mix of the populations used in the primary studies, in terms of either unrealistically high prevalence of disease or unrealistic proportions of the different abnormalities across the full spectrum of disease compared with a national screening population. Furthermore, it has been observed that in a number of primary studies there has been the preferential application of a gold (reference) standard to only a specific subgroup of the study sample. Such observations encompass the concepts of spectrum and verification bias and are again discussed in more depth in Chapter 5. However, in the context of assessing the relative strengths and weaknesses of past reviews on automation, it needs to be noted that some reviews do not seem adequately to address important issues concerning the internal validity of the included studies that they review; indeed, some do not address them at all.

Variation concerning the analysis of included study results was noted among the included past reviews of the clinical effectiveness of automation. For the calculation of new test performance metrics, unambiguous knowledge of the whole study sample is crucial. Without this, all metrics have an associated uncertainty. To measure sensitivity and specificity what is meant by both a positive test and true disease needs to be defined. This is once again discussed in more detail in Chapter 5. However, it is worth noting that distinction between the threshold for the test and the threshold of the reference standard was only clear in two reviews.[12,65] One other review, good in many other respects, by Broadstock,[62] seems to imply that HSIL+ should be the appropriate test positive threshold, but then goes on to say that, "detection of HSIL+ is regarded as the primary objective of the national cervical screening programme". This betrays a lack of appreciation that HSIL+ could refer to both a test diagnosis and a reference diagnosis defining the disease, and that thresholds for both need to be defined.

There were other issues concerning analysis of the results of included studies where there was important variation between one past review and the next. These included:

- summarising or pooling the results for sensitivity, without taking into account important issues of internal and external validity of included studies
- pooling results of specificity independently of the assessment of sensitivity, so failing to take into account that sensitivity and specificity are interdependent
- not considering specificity at all, and so failing to give some indication of the likelihood and consequences of false positives, as well as false negatives
- inappropriate use of meta-analysis, especially the attempt in two reviews to provide summary odds ratios to capture the results of the included studies overall
- failure to consider the possibility that chance variation alone may account for observed differences in say sensitivity; 95% confidence intervals (CI; or the equivalent) for test performance metrics were rarely provided.

Thus, many of the reviews appeared to have made basic errors in their analysis of results of included studies. Although it is readily acknowledged that analysing the results of studies assessing the effectiveness is extremely challenging, it was clear that some reviews performed much better than

TABLE 5 *Summary of results of appraisal: clinical effectiveness of automation*

Study device	Objective	Concluding remarks of review	Quality of review	Independence
Abulafia and Sherer, 1999[57] AutoPap	To estimate the overall FNR of the AutoPap 300 QC system when applied as a (1) primary screening system, (2) quality control (rescreening) system	A core group of authors was responsible for the majority of the publications (13/14). With the independence of these studies being called into question, the authors indicated that "any meta-analysis of this collection of studies should be interpreted with caution" There is a relative paucity of data on the AutoPap 300 QC	Many major weaknesses identified	No issues identified
Abulafia and Sherer, 1999[58] PAPNET	Compared with conventional screening, to determine whether the PAPNET as a primary screener (1) identifies a larger number of abnormalities, (2) has a lower FNR; and (3) whether the PAPNET as a rescreener reduces the FNR	Two studies were rejected for being discordant with the others "We conclude that compared with manual screening, PAPNET identifies 20% more abnormal, has two-fold less false negative, and re-classifies as abnormal one third of manually screened false negative slides"	Many major weaknesses identified	No issues identified
AHTAC, 1998[53] PAPNET, AutoPap, AutoCyte SCREEN	(1) To what extent do the new technologies have the potential to reduce the incidence of, and morbidity and mortality from, cervical cancer? (2) What is the potential of the new technologies to increase the sensitivity and specificity of Pap smear screening?	The following deficiencies in evidence were noted by the authors: a limited number of studies, extensive manufacturer involvement in the studies, absence of RCTs, lack of cytological threshold for positive and negative results, no consistent definition of a positive smear, few studies with biopsy confirmation of results, no definition of gold standard for negative results, sensitivity and specificity generally not reported, tests of statistical significance often not undertaken or reported, lack of consistent comparator, reviewers not always blinded to outcome, study populations displayed spectrum bias Given all of these, the new technologies could not be recommended at the time of the review	Some major weaknesses identified	No issues identified

continued

TABLE 5 *Summary of results of appraisal: clinical effectiveness of automation (cont'd)*

Study device	Objective	Concluding remarks of review	Quality of review	Independence
Broadstock (NZHTA) 2000[62] AutoPap	(1) To assess the clinical effectiveness of semi-automated and automated cervical screening systems (2) To determine the applicability of the evidence in the context of the national screening programme in New Zealand	Sensitivity and specificity could not be reliably determined. No difference was found in the detection of high-grade lesions Increase in sensitivity probably confined to low-grade lesions More research required "The majority of studies appraised were at least partially funded by the industry producing the devices considered" The vast majority of missed lesions in the existing programme would be detected in subsequent screening rounds Increases in sensitivity may come at the cost of decreased specificity Automation cannot be recommended for the New Zealand national cervical screening programme	Minor weaknesses identified	No issues identified
Brown and Garber, 1999[63] AutoPap, PAPNET	To estimate the proportional increase in TPR of (1) manual primary screening and new technology-assisted rescreening, compared with (2) manual primary screening and 10% random rescreening. This was done to provide input data to a model	"Our estimates are subject to uncertainty because the literature on the effectiveness of the 3 technologies (includes ThinPrep) reviewed here is incomplete and sometimes contradictory" "The highest quality studies suggest that the technologies increase the TPR by a modest amount, especially in a laboratory that is already highly accurate"	Major weaknesses identified	No issues identified
Mango and Radensky, 1998[59] PAPNET	To provide effectiveness metrics for the clinical utility of PAPNET	There is a relatively extensive evidence base for the PAPNET, which suggests that its sensitivity for abnormalities exceeds that of unassisted screening Sensitivity used as the main outcome because the patients' safety depends on this metric more than any other	Major weaknesses identified	Authors have potential conflicts of interest

continued

TABLE 5 *Summary of results of appraisal: clinical effectiveness of automation (cont'd)*

Study device	Objective	Concluding remarks of review	Quality of review	Independence
McCrory *et al.*, 1999[12] AutoPap, PAPNET	To estimate the accuracy of the Pap test and the new technologies	"The values reported for sensitivity and specificity in the few studies that use histological or colposcopic reference standards are well within the range of sensitivity reported for the conventional Pap test" "However, including studies that directly compare these new technologies with conventional Pap smear testing (screening or re-screening) using a cytological reference standard results in significant improvements in sensitivity" The authors concluded that the evidence on the new technologies was insufficient for two reasons: (1) There is little evidence on the specificity. (2) Most of the estimates of sensitivity are based on a surrogate reference standard: cytology. Independent consensus agreement by a panel is often confined to the discordant samples only, with biopsy confirmation of high-grade lesions often lacking. Both lead to an overestimation of diagnostic performance	Minor weaknesses identified, least of all reviews identified	No issues identified
Nanda *et al.*, 2000[65] Any, including AutoPap, PAPNET	To evaluate the accuracy of conventional and new methods of Papanicolaou testing when used to detect cervical cancer and its precursors	There were three main deficiencies in the methodologies of the excluded studies: (1) Many of the studies reviewed did not apply the new technology and the conventional Pap test prospectively to the same sample. (2) Where cytology was used as the reference standard, only discordant results were verified, concordant results remaining unverified. (3) There was little evidence on the value of the specificity for any of the new technologies Other deficiencies applied to all studies, and included: the presence of spectrum bias in the study population, the problems associated with no recognised reference standard, and the presence of verification bias in study methodology	Minor weaknesses identified	No issues identified
Noorani *et al.*, 1997[52] AutoPap, PAPNET	(1) To examine the effectiveness of the Pap test (2) To consider different strategies for improving the effectiveness of the Pap test	Operating characteristics of automated systems could not be estimated There is a "lack of a common definition of the gold standard for Pap smear re-screening" The new techniques may increase the effectiveness of the Pap test Resources should not be diverted from recruitment (of subjects), information systems, training and quality control requirements for laboratories to promote the new technologies	Some major weaknesses identified	No issues identified

FNR, false-negative rate; TPR, true-positive rate.

others, and in this respect the review by McCrory and colleagues[12] provides a good model to follow.

Finally, variation in issues that may impinge on the independence or validity, or both, of the included past reviews was noted. First, several reviews appear to have been undertaken by the same groups that undertook the majority of the primary research included in those reviews. In such a situation, unless there is a well-documented, comprehensive search strategy and clear inclusion criteria have been demonstrably applied, it is difficult to convince readers that the possibility of selection bias in included studies has been avoided. Second, two of the reviews included and appraised here were either partially funded or had authors who were funded by the companies producing the devices considered.[41,59]

Results of the appraisal and analysis: conclusions of included reviews on clinical effectiveness of automation

These are summarised in *Table 5*. More extensive analysis is given in the evidence tables detailed in Appendix 4.

This shows that there is variation in the degree of optimism concerning the value of automation, ranging from there being clear evidence for automation (AutoPap or PAPNET) being of value, to clear evidence that it is not. What *Table 5* also shows, however, is that the more optimistic assessments tend to have used review methods that are more open to bias. It is also notable that one of the reviews concluding favourably concerning the effectiveness of PAPNET had authors with clear potential conflicts of interest. All the more recently conducted reviews, which as noted earlier have more systematic approaches to reviewing the available literature, have universally concluded that evidence on the effectiveness of automation is inadequate and that it should not be introduced.

In 1998, the Australian health technology assessment review[53] listed the deficiencies in evidence relating to the evaluation of new technologies in cervical screening as:

- a limited number of studies
- extensive manufacturer involvement in the studies
- absence of RCTs
- lack of cytological threshold for positive and negative results
- no consistent definition of a positive smear
- few studies with biopsy confirmation of results
- no definition of gold standard for negative results
- sensitivity and specificity generally not reported
- tests of statistical significance often not undertaken or reported
- lack of consistent comparator
- reviewers not always blinded to outcome
- study populations displayed spectrum bias.

Given all of the above, the new technologies could not be recommended at the time of the review.

Three years on, many of these deficiencies appear to persist.

A further issue of note is that all reviews have noted the apparent absence of RCT evidence, and have concentrated in lieu of this on estimates of sensitivity and specificity (or their equivalent) to indicate evidence on effectiveness. This was a surprise, given that it is clear from general literature on the assessment of screening and diagnostic tests[69] that test performance is only one potential contributor to effectiveness. Other important components include the repeatability of the test and its impact on the process of care. The apparent absence of any attempt to review evidence on these and other aspects of effectiveness is discussed at greater length in Chapter 4, alongside a consideration of the types of study design that may be best suited to collecting valid information on them.

Results of the appraisal and analysis: conclusions of included reviews on clinical effectiveness of LBC

These are summarised in *Table 6*. More detailed evidence tables may be found in Appendix 4.

Several of the deficiencies in the primary literature on automation (as listed by the Australian review[53]) are directly applicable to the research on LBC, not least the absence of RCTs, the relative paucity of an accurate gold standard and the almost blinkered concentration on estimating the sensitivity of LBC without considering the corresponding specificity. Some of the primary studies were on split samples, a technique where the spatula is used to produce the conventional smear first, and the residue is placed in the vial so a liquid-based specimen may be prepared. Such a method is unique to LBC studies and is not encountered in the studies on automation. There are various opinions as to whether this prejudices[38,53] or benefits[39] the liquid-based technique, and this issue is unlikely to be resolved here. What it does suggest, however, is that other

TABLE 6 *Summary of results of appraisal: clinical effectiveness of LBC*

Study device	Objective	Concluding remarks of review	Quality of review	Independence
AHTAC, 1998[53] ThinPrep, AutoCyte PREP	(1) To what extent do the new technologies have the potential to reduce the incidence of, and morbidity and mortality from, cervical cancer? (2) What is the potential of the new technologies to increase the sensitivity and specificity of Pap smear screening?	Same deficiencies as noted for automated image analysis (see *Table 5*) In addition, all study designs for the assessment of both LBC technologies have been prospective and used the split-sample technique. This may disadvantage LBC, as this is prepared after the conventional smear There appears to be a trade-off between the conventional Pap test and LBC, as each technology detects abnormalities that the other fails to detect There is a significant learning period before becoming competent at monolayer screening	Some major weaknesses identified	No issues identified
Austin and Ramzy, 1998[66] ThinPrep, AutoCyte Prep	To compare the clinical effectiveness of LBC with conventional screening, by measuring the increased number of abnormalities detected	The data show an overall increase in the detection of LSIL for both LBC technologies, compared with conventional techniques Comparisons made to detect the impact of collecting device, however, demonstrated that LBC detects fewer abnormalities than conventional when the Ayre's wooden spatula is used. The residue is used for the LBC system This has led the authors to suggest that direct-to-vial studies would "ultimately be much more relevant than currently available split-sample data in judging the true potential of liquid-based methods to enhance detection".	Some major weaknesses identified	No issues identified
Broadstock (NZHTA), 2000[62] AutoCyte PREP, ThinPrep	(1) To assess the clinical effectiveness of LBC cervical screening systems (2) To determine the applicability of the evidence in the context of the national screening programme in New Zealand	Similar concerns to those for automated image analysis (see *Table 5*) Inadequate verification of the new tests against a suitable reference standard was a consistent failing; only discordant results were verified and concordant positives were often not subject to any verification The introduction of LBC cannot be recommended for the New Zealand national cervical screening programme	Minor weaknesses identified	No issues identified

continued

TABLE 6 *Summary of results of appraisal: clinical effectiveness of LBC (cont'd)*

Study device	Objective	Concluding remarks of review	Quality of review	Independence
Brown and Garber, 1999[63] ThinPrep 2000 system	To estimate the proportional increase in TPR of (1) manual primary screening with ThinPrep and 10% random rescreening, compared with (2) manual primary screening and 10% random rescreening	As for automated image analysis (see *Table 5*)	Major weaknesses identified	No issues identified
McCrory *et al.*, 1999[12] ThinPrep	To estimate the accuracy of the Pap test and the new technology	Identical concerns to those for automated image analysis (see *Table 5*)	Minor weaknesses identified; least of all reviews identified	No issues identified
Nanda *et al.*, 2000[65] ThinPrep	To evaluate the accuracy of conventional and new methods of Papanicolaou testing when used to detect cervical cancer and its precursors	Identical concerns to those for automated image analysis (see *Table 5*)	Minor weaknesses identified	No issues identified
Payne *et al.*, 2000[38] CYTOSCREEN, LABONORD Easy Prep, AutoCyte PREP, ThinPrep	What is the effectiveness of LBC for cervical screening compared with conventional screening?	LBC would lead to: a decrease in the percentage of inadequate specimens, an improvement in sensitivity, a probable decrease in interpretation times, and a potential for easier use with other technologies, such as HPV and automation However, there were the following deficiencies in evidence: no RCTs, specificity is largely unknown and may be worse than the conventional Pap test, and few studies had a gold-standard comparator	Minor weaknesses identified	No issues identified

TABLE 7 *Summary of results of appraisal: clinical effectiveness of HPV*

Study	Objective	Concluding remarks of review	Quality of review	Independence
Cuzick *et al*., 1999, 2000[41,67]	To evaluate the available data concerning the role of HPV testing in primary screening, either alone or as an adjunct to cytology To improve the management of women with low-grade cytological abnormalities To improve the accuracy of follow-up after treatment of preinvasive or early invasive lesions To review the methods available for HPV testing and determine their appropriateness for widespread implementation To determine what future research is required to obtain more reliable answers about its use in screening	All three technologies had similar performance characteristics. Sensitivity and NPV were superior to other technologies used to detect HPV Only one technology is currently available as a commercial off-the-shelf kit (the Digene HCII assay) In primary screening "HPV testing is more sensitive than cytology for detecting CIN II/III"; however, "the specificity is substantially lower" HPV testing in borderline and low-grade cytology cases "greatly improves the specificity and positive predictive value"	Some weaknesses identified	Authors have potential conflicts of interest

NPV, negative predictive value.

methods of appraisal of the technology (such as direct-to-vial studies) should not be discarded in preference for the split-sample methodology.

As with automated image analysis, the reviews encompassed a wide range of views on the effectiveness of LBC. Most of the reviews concluded that there was insufficient evidence either to draw strong conclusions or to recommend its implementation across a national healthcare system. The most recent review, by Payne and colleagues,[38] while noting persisting uncertainties, felt that the evidence on balance indicated benefit. The guidance subsequently issued by NICE suggested a piloted introduction.[47]

Results of the appraisal and analysis: conclusions of included reviews on clinical effectiveness of HPV

These are summarised in *Table 7*. More detailed evidence tables may be found in Appendix 4.

The two papers cited relate to one HTA review published in 1999,[41] as the later paper published in 2000[67] reports on a summary of the findings of this review.

At the time of the review the evidence summarising the HPV screening systems was sparse. In particular, the use of HPV screening as a primary screening tool could not be supported, owing to the lack of evidence on its effectiveness. A sensitivity analysis of the effectiveness parameters used in the economic model in the review demonstrated a huge swing in the expected outcomes from the totally favourable to the totally unfavourable.

This was not the case in the selective use of HPV screening on borderline and mild cytology cases, where it was concluded that such a strategy could improve the existing system even when modelled in the worst case. Such a role of augmenting cervical cytology as a screening tool, rather than replacing it, is currently being tested in pilot studies at a number of sites in the UK.

However, it should be noted that two of the authors of the report had either a direct or an indirect relationship with the one company that produces a commercially available screening system, leading to a potential conflict of interest.

Technologies and approaches to improving effectiveness of cervical screening, other than automation, LBC and HPV

Several authors noted that improving the current cervical programme does not necessarily mean introducing a new technology. In particular, it was recognised that increasing the coverage in many cases may significantly reduce the incidence of cervical cancer.[38,52,53,62] Broadstock[62] suggested that rather than committing resources to the new technologies they would be more efficiently deployed by:

- increasing uptake of routine screening
- 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
- ensuring adequate follow-up and treatment where required.

It has also been noted that the standard Ayre's wooden spatula could well be the worst type of collecting instrument for cervical sampling and that extended tip spatulas would offer greater efficacy.[38]

McCrory and colleagues[12] go as far as to suggest that when all the 'costs' are considered, a cervical screening programme may not be worth it after all: "given the rarity of cervical cancer relative to HPV infection and SIL, the inconvenience, potential discomfort, and psychological distress associated with screening and treatment of cancer pre-cursors (many of which will never progress to become cancer) might well outweigh the negative impact of cancer itself on women's quality of life at a population level."

Chapter 4

Systematic review of the literature on cost-effectiveness of automation

Summary of key points

The main objective was to evaluate what was already known about the cost-effectiveness and cost–utility of automated image analysis devices.

The search used to identify systematic reviews in the previous chapter was amplified, targeting articles on cost-effectiveness. Of the 233 citations examined, 13 studies were included, only eight of which contributed substantial information on cost-effectiveness and are discussed in detail.

Based on the included studies, there is currently a large degree of uncertainty regarding the cost-effectiveness of automated image analysis devices. The cost-effectiveness result seems to be driven, in part, by the estimates made of the changes in test sensitivity and specificity that result from the introduction of automation.

The most rigorous and robust analysis currently available is undoubtedly that reported by McCrory and colleagues[12] and any future analytical work should use their model as a starting point.

On the basis of this review it is clear that an independent assessment is needed of the cost-effectiveness of the recently developed AutoPap technology in primary screening mode in a UK setting. This would provide a way of establishing the robustness of the policy recommendations resulting from the study by Smith and colleagues.[61]

None of the economic analyses identified in this review considered the policy question of combining LBC with automation. This question is particularly pertinent to the UK, where a realistic policy option in the short term is that LBC will become part of the established NHSCSP. Future research needs to address this broader issue.

Data limitations currently prevent full account being taken in economic analyses of the impact of automation on patient quality of life, especially the improvements that may result from reductions in the number of false positives.

Future economic analyses concerned with approaches to improve the efficiency of cervical screening programmes should be broad in their focus to include issues of screening uptake and not just factors relating to the nature of the programme that is delivered to women participating in screening.

Introduction

This chapter has two components. The first section provides a review of previously published economic analyses of automated image analysis devices, and the second reports a supplementary review of studies not included in the first section that have adopted a decision-analytic model to address a policy issue in the area of cervical screening. The main purpose of the cost-effectiveness review is to identify the variation in results reported, and using information provided in the published studies to explore the reasons for such variation. The supplementary review of models provides an insight into existing modelling work carried out in this area. Information from both aspects of this review has informed the primary cost-effectiveness and modelling analysis described later in the report (Chapter 9). The methods section of this chapter describes the methods used for the two components, but the results and conclusions are reported separately.

Methods

The starting point for both aspects of the review reported in this chapter (i.e. for the cost-effectiveness and modelling components) was the output of the search for secondary research reported in Chapter 3. Beyond this, an additional search was undertaken to amplify published information on cost-effectiveness of automated image analysis devices, defined as a study estimating both the costs and consequences associated with the use of automated cervical screening. The same search was also used in the review of cost data reported in Chapter 8.

The sources interrogated for this additional search were:

- Cochrane Library 2001, Issue 2 [includes NHS Economics Evaluations Database (NHS EED) and DARE]
- MEDLINE, 1996 to March 2001
- EMBASE, 1998 to March 2001
- CINAHL, 1998 to May 2001
- CANCERLIT, 1998 to March 2001
- HealthSTAR, 1998 to December 2000
- EconLit, 1998 to May 2001
- NRR
- websites and publication lists of key organisations:
  - NHS CRD
  - National Coordinating Centre for HTA
  - NICE
  - Agency for Health Care Policy and Research
  - Canadian Co-ordinating Office for HTA
  - Swedish Council on Technology Assessment in Health Care
  - New Zealand HTA
  - Australian HTA
- direct contact with:
  - NHSCSP
  - New Zealand HTA, Canadian Co-ordinating Office for HTA
  - Manufacturers TriPath, Cytyc.

The search was deliberately focused on recent literature to maximise the applicability of any results to the current day. The search strategies used combined series of terms capturing the condition of interest (e.g. cervix neoplasms/), the intervention of interest (e.g. image processing computer assisted/) and the type of literature desired (e.g. economics medical/). Full details of the search strategy for MEDLINE are provided in Appendix 5.

Searches of the bibliographic databases were supplemented by examining the reference lists of reviews and included studies identified, looking particularly for citations which indicated that an article had dealt with both automated screening and had considered costs or efficiency, cost-effectiveness or cost–utility. The outputs of the searches for reviews and effectiveness were similarly scrutinised.

Inclusion criteria for this section on cost-effectiveness were simply that the study should provide information on both the costs and consequences associated with an automated cervical screening device. The only automatic exclusion criterion was if the study was a review or an editorial, where the main aim was to report others' assessments of cost-effectiveness without any significant additional analysis or modelling. The most systematic and up-to-date systematic review was, however, retained for the purposes of comparison. All of the included studies were formally appraised, allowing important strengths and weaknesses of the studies to be identified. The criteria used for the appraisal are broadly based on the Drummond checklist for health economic assessments.[70] The data abstracted from included studies falls under two headings:

- *comparisons made and methods used* (including such factors as the study comparators and screening population, the form of analysis used and the approach to modelling used)
- *reported results* (including the base-case results for incremental costs, effects and the cost-effectiveness ratio, and the sensitivity analysis results).

Inclusion criteria for the modelling review were that the study reported a decision-analytic model that addressed a policy question relating to cervical screening and that it was not already included in the cost-effectiveness review. The method of analysis was qualitative, relying on clear tabulation and presentation of characteristics, study quality and findings to generate conclusions and implications.

Systematic review of literature on cost-effectiveness: results

Yield of search

In total, 223 citations were examined. This included all of the articles identified for the review of secondary literature on clinical effectiveness reported in Chapter 3. Of the 233, 140 were immediately excluded, mostly for reasons of duplication (133 studies). Full text of 82 of the 83 provisionally included studies was examined; one study in the *Journal of Clinical Ligand Assay* could not be obtained.[71] Sixty-six of the 82 were excluded, 59 because they did not deal with cost or cost-effectiveness and/or did not deal with automation, and seven because they did not provide any new primary data or undertake new analysis or modelling. Thus, 16 studies were included in either the cost-effectiveness review or the review of cost data (see Chapter 8) or both. Thirteen papers contributed some information on cost-effectiveness.[12,18,52,53,60,61,63,64,72–76] However as shown in *Table 8*, only eight of these papers contributed substantial information on cost-

TABLE 8 *Included cost-effectiveness studies, and those analysed in detail*

Study	Met inclusion criteria?	Analysed in detail?	Comment
AHTAC, 1998[53]	Yes	Yes	
Brown and Garber, 1999[63]	Yes	Yes	
McCrory *et al.*, 1999[12]	Yes	Yes	
Noorani *et al.*, 1997[52]	Yes	Yes	
O'Leary *et al.*, 1998[72]	Yes	Yes	
Radensky and Mango, 1998[60]	Yes	Yes	
Schechter, 1996[73]	Yes	Yes	
Smith *et al.*, 1999[61]	Yes	Yes	
Brotzman *et al.*, 1999[74]	Yes	No	Not a full economic analysis; crude analysis in terms of cost per additional abnormal result; valuable for providing direct information on cost and so considered further in Chapter 8
Hutchinson, 1996[18]	Yes	No	Not a full economic analysis; crude analysis in terms of cost per additional abnormal result
Myers *et al.*, 2000[64]	Yes	No	No explicit consideration of automation; considered in review of modelling approaches
Raab *et al.*, 1999[75]	Yes	No	No explicit consideration of automation; main aim to assess cost-effectiveness of manual screening
Troni *et al.*, 2000[76]	Yes	No	Not a full economic analysis; crude analysis in terms of cost per additional abnormal result; valuable for providing direct information on cost and so considered further in Chapter 8

effectiveness or cost–utility, and are discussed in detail.[12,52,53,60,61,63,72,73] The data abstracted from the eight papers are presented in *Tables 9–12*. Concerning the five included studies not discussed in detail, two studies, although mentioning automation, focused mainly on the cost-effectiveness technological advances in general[64] or the cost–effectiveness of manual screening,[75] and three studies presented very rudimentary analyses expressing results in terms of cost per additional abnormal case detected.[18,74,76] Of these latter three, however, two presented valuable directly collected information on costs which is discussed in detail in Chapter 8.[74,76]

Contextual issues

Much of the previous cost-effectiveness work in the area of automation of cervical screening programmes has been undertaken in the USA, and none of the studies relates directly to the UK context. Given this observation, it is necessary that care be taken in extrapolating the results to the UK, especially those data relating to costs.

It is also important to be aware that three of the eight studies considered in detail were either funded directly by a manufacturer of an automated device or undertaken by researchers who declared the receipt of payments from one of the device manufacturers.[60,61,73] The study with the most positive result, by Smith and colleagues,[61] where the automation technology (AutoPap) was found to be a dominant technology (i.e. a technology associated with both lower costs and greater effectiveness), was supported by a research grant from the manufacturer of AutoPap.

Comparisons made in studies

Table 9 provides information on the technologies that have been compared in the studies reviewed. There is clearly considerable variation across studies in the targets for valuation, which makes the direct comparison of study results inappropriate. Although the principal focus has been on the PAPNET technology, in several studies AutoPap was considered either as an alternative to conventional screening[61] or as a competitor to PAPNET.[52] In a position mirroring the results of the effectiveness review (see Chapter 6), all published assessments of the cost-effectiveness of automated devices relate essentially to old and potentially outdated versions of the technology.

TABLE 9 *Study comparators and effectiveness data*

Study	Alternative technologies compared	Screening population	Effectiveness data used in CEA	Sensitivity (%)	Specificity (%)
AHTAC, 1998[53]	(1) Conventional cervical screening (2) Conventional screening plus use of automated device for normal readings	Not stated	Automated devices assumed to lead to an increase of 10% and 6% in high and low-grade readings respectively, and about 300 additional potential cancer cases in a 2-year screening cycle		
Brown and Garber, 1999[63]	(1) Pap smear with 10% random rescreening (2) ThinPrep with 10% random rescreening (3) Pap smear with AutoPap-assisted rescreening (4) Pap smear with PAPNET-assisted rescreening	Women between the ages of 20 and 65 years: screening begins for all women at the age of 20 years	Pap smear and 10% rescreening ThinPrep smear and 10% rescreening Pap smear with AutoPap-assisted rescreening Pap smear with PAPNET-assisted rescreening	81.6 92.6 95.4 97.0	95.8 95.8 95.4 95.4
McCrory *et al.*, 1999[12]	(1) No Pap smear screening (2) Conventional Pap smears (at 1-, 2- and 3-year intervals) (3) A technology that improves test sensitivity (4) A technology that allows 100% rescreening	US women from the age of 15 to 85 years. Six screening strategies considered beginning at the age of 15, 18, 20, 35, 50 and 65 years	Conventional Pap smears A technology that improves test sensitivity A technology that allows 100% rescreening	53.5 84.3 80.4	Assumption: no effect of new technologies
Noorani *et al.*, 1997[52]	(1) Manual rescreening of random 10% of negative cervical smears (2) AutoPap rescreening of all negative smears (3) PAPNET rescreening of all negative smears	Not stated	Manual rescreening AutoPap rescreening PAPNET rescreening	75 80 83	95% Not stated Not stated
O'Leary *et al.*, 1998[72]	(1) PAPNET-assisted rescreening of smears (2) Manual rescreening of all smears	Female members of the US Air Force and their dependants aged 12–88 years	Of the 5478 Pap smears rescreened, the use of PAPNET led to five being reclassified as ASCUS and one as AGUS. No additional SIL was identified		
Radensky and Mango, 1998[60]	(1) Unassisted manual examination of smears (2) Interactive neural network-assisted screening (i.e. PAPNET rescreening)	Women between the ages of 20 and 75 years: screening begins for all women at the age of 20 years	Unassisted manual examination of cervical smears Interactive neural network-assisted screening	85 89–100%	99.5 99.5

continued

TABLE 9 *Study comparators and effectiveness data (cont'd)*

Study	Alternative technologies compared	Screening population	Effectiveness data used in CEA	Sensitivity (%)	Specificity (%)
Schechter, 1996[73]	(1) Screening by cervical smear (2) Rescreening negative smears with PAPNET	Women between the ages of 20 and 64 years: screening begins for all women at the age of 20 years	Screening by cervical smear Rescreening negative smears with PAPNET	75 LSIL 85 HSIL 80 LSIL 88 HSIL	98 95
Smith *et al.*, 1999[61]	(1) Conventional manual Pap screening (2) Primary screening using AutoPap	Women from the age of 18 years: screening for all women begins at the age of 18 years	Conventional manual Pap screening Primary screening using AutoPap	81.6 91.0	94.0 95.0

AGUS, atypical glandular cells of undetermined significance; CEA, cost-effectiveness analysis.

Studies also vary in the assumptions that have been made in how the technologies are to be used. In most studies the expectation has been that the automated screening device would not be used in primary screening, but instead would be used as part of the rescreening strategy for smears initially classified as 'negative'. The study by Smith and colleagues represents the exception where AutoPap is evaluated as a device to be used in primary screening.[61] Other factors making comparison between studies problematic are:

- The population at which the screening is targeted is not always common.
- The assumed interval between screening varies.

The study by McCrory and colleagues highlights the uncertainties relating to these two parameters and, helpfully, treats the age at which screening commences and the screening interval as variables.[12] They, therefore, provide cost-effectiveness results for screening strategies beginning at the age of 15, 18, 20, 35, 50 and 65 years, and for intervals of 1, 2 and 3 years.

Variation in comparators used in cost-effectiveness studies of automated cervical screening results, in part, from the observed variation in the nature of screening programmes being delivered in various countries. For example, an important difference between the USA and the UK in the delivery of cervical screening programmes is what constitutes 'manual Pap screening'. In the USA, after initial examination manually, a 10% manual random rescreen is mandatory. In contrast, in the UK, the initial manual screen is followed by a rapid rescreen of all slides. This almost certainly has implications about the baseline costs and effectiveness against which automation is being compared and makes direct comparison of results from studies conducted in the USA and the UK problematic. Again in contrast to the USA, the UK system has much more complete screening coverage, which inevitably has potentially important implications for the true prevalence of abnormalities likely to exist in the screened population. The focus here has been on comparison with the USA, given that most studies have been undertaken in that setting. However, it is important to be aware that variation in what constitutes manual Pap screening exists across Europe too.[77]

None of the economic analyses identified in this review considered the policy question of combining thin-layer slide preparation (i.e. LBC) with automation. This question is particularly pertinent to the UK, where a realistic policy option in the short term is that LBC will become part of the established NHSCSP.

Estimates of effectiveness used in cost-effectiveness analysis

Table 9 also provides information on the effectiveness data used in the economic analyses. Caution is required in making direct comparisons of test sensitivity and specificity estimates across studies because of variation between studies in how such variables are defined. For example, as highlighted by Schechter, the test sensitivity for the detection of LSIL is likely to be different (i.e. lower) to the sensitivity of the same test for HSIL.[73] The same argument applies for the estimation of test specificity, but this issue is not considered by any of the economic analyses. This explanation of variation in definitions of sensitivity may help to explain some of the variation in sensitivity estimates. However, there is almost universal support across studies for the fact that automation (regardless of the nature of the device) brings about an increase in sensitivity, with little or no impact on specificity. For some studies, for example, Brown and Garber, the increase in sensitivity is dramatic: increasing from 81% for the conventional Pap smear to 97% when PAPNET is used in rescreening.[63] Once again, the study by McCrory and colleagues deserves comment in that the researchers adopted a threshold approach to the issue of gain in sensitivity, asking the question: 'What increase in sensitivity would be required in order that the automated device might be viewed as cost-effective?'[12]

It is worth making a link at this point to the review of effectiveness evidence (see Chapter 6). Although a reasonable number of economic analyses has considered AutoPap, especially in rescreening mode, the review of effectiveness identified no rigorous assessments of test performance of AutoPap.

Although several of the studies report results in terms of a cost-effectiveness analysis with life-years gained as the measure of effect, none takes the leap to cost–utility analysis, where quality-adjusted life-years would be the measure of effect. This reflects another data limitation, this time concerning the impact of improvements in the performance of the screening programme on patient quality of life. Quality of life issues may be of great importance, especially if automation were able to reduce the rate of false positives.

Methods used in studies

A common characteristic of all studies is that cost-effectiveness analysis (as opposed to either

TABLE 10 *Economic analysis methods*

Study	Form of analysis	Perspective	Time horizon	Resource use/costs	Currency, price year	Discounting
AHTAC, 1998[53]	CEA: cost per additional potential cancer case	Health sector costs only	Initial screening costs plus subsequent diagnostic and treatment costs	Use of resources indicated by model Unit costs from published sources	Aus$, not stated	Not considered, given short time horizon
Brown and Garber, 1999[63]	CEA: cost per year of life saved	Societal	Lifetime costs and health effects	Use of resources indicated by model (e.g. number of smears and treatments) Unit costs obtained from a variety of sources, both published and unpublished	US$, 1996	3% per annum for both costs and effects
McCrory *et al.*, 1999[12]	CEA: cost per life-year gained, cost per cervical cancer death prevented, cost per cervical cancer case prevented	Health sector costs only	Lifetime costs and health effects	Use of resources indicated by model (e.g. number of smears and treatments) Unit costs obtained from a variety of sources, both published and unpublished	US$, 1997	3% per annum for both costs and effects
Noorani *et al.*, 1997[52]	CEA: cost per additional abnormal case found	Third party payer	Short-term costs associated with the screening process	Use of resources indicated by model Unit costs from published sources and from manufacturers	Can$, not stated	Not considered, given short time horizon
O'Leary *et al.*, 1998[72]	CEA: cost per additional LSIL identified	Health sector costs only	Short-term costs associated with the screening process	Crude analysis of resource use and costs Unit costs based on local and published sources	US$, not stated	Not considered, given short time horizon
Radensky and Mango, 1998[60]	CEA: cost per year of life saved	"Modified payer perspective using costs borne by payers combined with patient deductibles and co-payments"	Lifetime costs and health effects	Use of resources indicated by model (e.g. number of smears and treatments) Unit costs obtained from a variety of sources, both published and unpublished	US$, 1997	3% per annum for both costs and effects
Schechter, 1996[73]	CEA: cost per day of life saved	"An agency responsible for providing health care to a defined population of women"	Lifetime costs and health effects	Use of resources indicated by model (e.g. number of smears and treatments) Unit costs obtained from a variety of sources, both published and unpublished	US$, 1994	5% per annum for both costs and effects
Smith *et al.*, 1999[61]	CEA: cost per day of life saved	Societal	Lifetime costs and health effects	Use of resources indicated by model (e.g. number of smears and treatments) Unit costs obtained from a variety of sources, both published and unpublished	US$, 1997	3% per annum for both costs and effects

cost–utility or cost–benefit analysis) was undertaken (see *Table 10*). The implication of this is that issues relating to improvements in quality of life have not been captured in the reported studies since the focus for the cost-effectiveness ratio was predominantly on life-years or life-days gained. However, some studies focused on intermediate measures of effectiveness, with cost-effectiveness ratios such as cost per additional LSIL identified or cost per cervical cancer case prevented.

The perspective adopted by studies tended to be either the healthcare sector or society more broadly. Given the long-term nature of cervical cancer progression, a time horizon of the lifetime of screened women is appropriate, but some studies considered short-term costs only, which matched a focus on intermediate measures of effectiveness.[72] A remarkably consistent position with regard to discounting was used, with only one study, for which discounting issues were relevant, not discounting both costs and effects at a rate of 3% per annum. This almost certainly results from the researchers' adoption of the US Panel on Cost-effectiveness statement on the discount rate to be used in base-case analyses.[78] When costs and effects are being considered over lifetimes, results inevitably become sensitive to variation in the discount rate used. It is, therefore, surprising that the sensitivity analyses used in these studies have not explored more fully the robustness of their results across alternative discount rates. The exception to this rule is the study by McCrory and colleagues, which demonstrates the high degree of sensitivity of the results of cost-effectiveness analyses of these technologies to the discount rate adopted.[12]

Table 10 reveals that only one study did not use a formal decision-analytic model as the framework for conducting the cost-effectiveness analysis. There is widespread use of Markov models, but it is commonly the case that researchers other than the authors of the paper in question originally constructed the model. For example, the Markov model originally constructed by Eddy[15,79] has been used in two of the economic analyses. As acknowledged by most of the authors, there are clearly concerns relating to the robustness of the original model assumptions to the situation being considered by the model, that is, current healthcare practice. Most of the modelling exercises have only considered women entering the model at the starting age for screening (e.g. 15 years) and so the policy issue of whether automation should be introduced for women who are currently part of the screening programme has not been explicitly addressed.

Again with the notable exception of the study by McCrory and colleagues,[12] most studies have conducted very limited sensitivity analyses. The vast majority conducted one-way analyses only (i.e. values on uncertain parameters are varied one at a time) on a limited range of parameters. However, test sensitivity has usually been included in sensitivity analysis, with the finding that results vary dramatically as the assumption made in terms of the gain in sensitivity achieved by automation is varied. The implication of the limited use of sensitivity analysis is that the true level of uncertainty surrounding the cost-effectiveness results is greater than that reported.

Cost-effectiveness study results

As indicated above, the use of different measures of effectiveness across studies limits the extent to which direct comparison of study results can be made (see *Table 12*). Most studies report an incremental cost-effectiveness ratio (ICER), which indicates the change in costs and change in effects associated with automation, relative to a comparator of the screening programme without the use of automation. Clearly, care must be taken in comparing ICERs across studies given that the automation technology and the comparator technology are not constant. Where comparable ICERs have been reported (e.g. the studies by Schechter[73] and Brown and Garber[63]) then very different ICER values have sometimes been estimated (e.g. US$48,000[73] versus US$29,000[63] per life-year gained, for the studies in question). Some of this variation appears to result from marked disparities in the estimates of the additional lifetime costs associated with women taking part in the screening programme where automation is used, even where similar technologies are being compared.

Given the lack of consensus on cost-effectiveness resulting from the base-case analyses reported in these papers, the results of sensitivity analyses become of great interest. It is, therefore, particularly disappointing that most studies have undertaken only very limited investigation of the importance of uncertainties in their analyses. From the work that has been done, a consistent finding (unsurprisingly) is the importance of the estimate of the gain in test sensitivity. It appears that the value on this parameter represents an important driver of the result of the cost-effectiveness analysis, and there is marked variation in the estimates quoted. The extensive

TABLE 11 *Modelling approaches, assumptions and sensitivity analyses*

Study	Model type	Selected model assumptions	Sensitivity analyses performed
AHTAC, 1998[53]	Decision tree model	Model based on Australian data concerning the performance of the cervical screening programme Increases in effectiveness from the introduction of automation taken from published sources	One-way and two-way, exploring variation in: costs and savings associated with automation, and effectiveness of automation
Brown and Garber, 1999[63]	Time varying transition state model (i.e. Markov model), developed by Eddy (1987)	No decline in mortality from cervical cancer since late 1980s All cancers develop from preinvasive lesions that may regress spontaneously The majority of cancers (80–95%) develop from a long preinvasive stage	One-way only, exploring variation in three sets of assumptions: characteristics of populations screened, cost and sensitivity of conventional Pap testing, and cost and sensitivity of new technologies
McCrory *et al.*, 1999[12]	Markov model, developed by Sonnenberg and Beck (1993)	Diagnostic evaluation of abnormal smears would detect all true histological abnormalities All patients with abnormal smears would receive appropriate follow-up and treatment New technologies would increase sensitivity without any decrement in specificity Detection of other diseases using the Pap smear (e.g. Chlamydia) not considered	Extensive one-way, two-way and threshold analysis, exploring variation in: diagnostic strategies, costs of diagnosis and treatment, incidence of cervical cancer, test sensitivity and specificity, hysterectomy incidence, discount rate, age at start of screening and screening frequencies
Noorani *et al.*, 1997[52]	Decision tree model	Proportion of true abnormal smears is 10% Sensitivity and specificity estimates provided above	One-way only, exploring variation in: true rate of abnormal smears, sensitivity and specificity, and rescreening and review costs
O'Leary *et al.*, 1998[72]	No formal modelling used	Not applicable	Adjustments made for the cost of PAPNET, but no formal sensitivity analysis reported
Radensky and Mango, 1998[60]	Time varying transition state model (i.e. Markov model), developed by Eddy (1990)	No decline in mortality from cervical cancer since late 1980s All cancers develop from preinvasive lesions that may regress spontaneously The majority of cancers (80–95%) develop from a long preinvasive stage	One-way only, exploring variation in: sensitivity and specificity, and costs of managing cervical cancer
Schechter, 1996[73]	Markov model	Model tracked cervical neoplasm-related events in a hypothetical cohort of women aged 20–64 years, with an age distribution and mortality pattern matching that of the US female population	One-way only, exploring variation in: costs of care and equipment, screening interval, sensitivity and specificity, follow-up regimens and natural history of cervical neoplasia
Smith *et al.*, 1999[61]	Markov model	Model run by entering a hypothetical cohort of 18-year-old women whose mortality pattern represents that of the general US population for women of that age Model continues to iterate until all women die	One-way only, exploring variation in: sensitivity and specificity, variable costs for both manual and automated screening, and disease prevalence

TABLE 12 *Results reported in cost-effectiveness studies*

Study	Base-case results: costs	Base-case results: effectiveness	ICERs	Sensitivity analysis results
AHTAC, 1998[53]	Additional screening costs: Between Aus$10 and $30 per screen Potential savings per screen: Between Aus$7.5 and $15 per screen	Automated devices assumed to lead to an increase of 10% and 6% in high- and low-grade readings, respectively, and about 300 additional potential cancer cases in a 2-year screening cycle	ICER for use of automation compared to conventional programme: Aus$240,000 per additional potential cancer case per screening cycle	Major determinant of variation in cost-effectiveness is the value of the increased effectiveness attributed to the use of automation
Brown and Garber, 1999[63]	Lifetime cost per woman screened (3-year screening interval) (US$) Pap smear with 10% random rescreen: 614 ThinPrep with 10% random rescreen: 695 Pap smear with AutoPap: 657 Pap smear with PAPNET: 700	Additional days per woman screened (3-year screening interval): Pap smear with 10% random rescreen: 24.93 ThinPrep with 10% random rescreen: 25.73 Pap smear with AutoPap: 25.89 Pap smear with PAPNET: 26.00	ICER for Pap smear with AutoPap-assisted rescreen against Pap smear (with 10% random rescreen): US$16,360 per life-year saved ICER for Pap smear with PAPNET-assisted rescreen against Pap smear (with 10% random rescreen): US$29,356 per life-year saved	Results insensitive to all but very large changes (i.e. ±50%) from baseline in estimated TPR and cost. But, "such large changes in TPR are within values reported in literature"
McCrory *et al.*, 1999[12]	Average lifetime cost per woman (3-year interval, age 15 at start of screening) (US$) No Pap: 893 Conventional Pap smears: 1108 A technology that improves test sensitivity: 1240 A technology that allows 100% rescreening: 1276	Additional life expectancy (days) over No Pap screening option: Conventional Pap smears: 19.2 A technology that improves test sensitivity: 21.4 A technology that allows 100% rescreening: 21.2	ICER for improved primary screening compared to conventional Pap screening: US$21,915 per life-year gained ICER for improved rescreening compared to conventional Pap screening: US$30,528 per life-year gained	Small decreases in specificity can significantly affect the cost-effectiveness estimates of any technology that improves sensitivity. Other factors causing large variation include age at start of screening, cancer incidence and discount rate
Noorani *et al.*, 1997[52]	Programme costs for screening 4 million women (Can$) Manual rescreening: 36 million AutoPap rescreening: 59 million PAPNET rescreening: 85 million	Additional abnormal cases (compared to no rescreen): Manual rescreening: 7500 AutoPap rescreening: 60,000 PAPNET rescreening: 62,250	ICER (cost per additional abnormal case) for strategy below compared to no rescreen: Manual rescreening: Can$251 AutoPap rescreening: Can$418 PAPNET rescreening: Can$810	Results sensitive to the screening cost for the automated devices, and the review cost by the pathologist
O'Leary *et al.*, 1998[72]	Not stated in disaggregate manner	Not stated in disaggregate manner	ICER for PAPNET rescreening against manual re-screening: US$17,475 per additional case of LSIL identified	Result sensitive to cost of PAPNET; using "costs quoted in advertisements" resulted in ICER of US$101,343 per additional case of LSIL identified

continued

TABLE 12 *Results reported in cost-effectiveness studies (cont'd)*

Study	Base-case results: costs	Base-case results: effectiveness	ICERs	Sensitivity analysis results
Radensky and Mango, 1998[60]	Lifetime cost per woman screened (3-year screening interval) (US$): Unassisted manual examination of cervical smears: 197 Interactive neural network-assisted screening (using sensitivity estimate based on biopsy and CIN+): 287	Additional days per woman screened (3-year screening interval): Unassisted manual examination of cervical smears: 25.492 Interactive neural network-assisted screening (using sensitivity estimate based on biopsy and CIN+): 26.333	ICER for neural network-assisted screening against unassisted manual screening: US$39,087 per life-year saved	Results highly sensitive to variation in values of sensitivity, but not sensitive to variation in treatment costs
Schechter, 1996[73]	Not stated in disaggregate manner	Not stated in disaggregate manner	ICER for PAPNET rescreening against conventional manual screening (2-year screening interval): US$48,474 per life-year gained	Results highly sensitive to the screening interval (ICER for 3-year interval: US$25,185 per life-year saved), to variation in values of sensitivity and specificity, and to change in assumptions concerning natural history
Smith *et al.*, 1999[61]	Lifetime cost per woman screened (3-year screening interval) (US$): Conventional manual Pap screening: 9388 Primary screening using AutoPap: 9385	Days of life saved by automated screening compared to conventional manual screening (3-year screening interval): 13.1	ICER for automated screening with AutoPap against conventional manual screening): –US$976 per life year saved Note: negative ICER indicates dominance for AutoPap (i.e. AutoPap associated with lower cost and greater effectiveness)	Results highly sensitive to variation in values of sensitivity and specificity

sensitivity analysis undertaken by McCrory and colleagues supports this general finding.[12] In addition, they indicate that other factors causing large variation include policy variables (e.g. age at start of screening), epidemiological factors (e.g. incidence of cervical cancer) and analysis parameters (e.g. the rate at which future costs and effects are discounted).

From a policy perspective it is important to consider specifically evidence relating to the use of AutoPap in primary screening mode. The reason for this is that effectively it represents the only variant of the automated technology under active consideration for implementation. There are currently no other commercially available devices and this AutoPap device is not designed for rescreening slides. From the present cost-effectiveness review, there is only a single study which provides a specific estimate of the cost-effectiveness of AutoPap in primary screening mode.[61] The study by Smith and colleagues appears to represent a highly optimistic assessment, with a base-case finding of dominance for AutoPap (i.e. AutoPap is associated with lower costs and greater effectiveness). Considerable caution is required in the interpretation of this result, for several reasons:

- It is based on a US setting and so extrapolation to the UK is problematic (as discussed above).
- The result is driven by the effectiveness estimate that indicates a dramatic improvement in sensitivity (and an increase in specificity).
- The funding for this study came from the manufacturer of the device in question.

Systematic review of literature on cost-effectiveness: conclusions

There currently exists a large degree of uncertainty regarding the cost-effectiveness of automated devices. The cost-effectiveness result seems to be driven, in part, by the estimates made of the changes in test sensitivity and specificity that result from the introduction of automation. The most rigorous and robust analysis currently available is undoubtedly that reported by McCrory and colleagues, and any future analytical work should use their model as a starting point.[12]

On the basis of this review it is clear that an independent assessment is needed of the cost-effectiveness of the recently developed AutoPap technology in primary screening mode in a UK setting. This would provide a way of establishing the robustness of the policy recommendations resulting from the study by Smith and colleagues.[61]

None of the economic analyses identified in this review considered the policy question of combining thin-layer slide preparation (i.e. LBC) with automation. This question is particularly pertinent to the UK, where a realistic policy option in the short term is that LBC will become part of the established NHSCSP. Future research needs to address this broader issue.

Data limitations currently prevent full account being taken in economic analyses of the impact of automation on patient quality of life, especially the improvements that may result from reductions in the number of false positives.

Future economic analyses concerned with approaches to improve the efficiency of cervical screening programmes should be broad in their focus to include issues of screening uptake and not just factors relating to the nature of the programme that is delivered to women participating in screening.

These conclusions are consistent with the most systematic and up-to-date review of economic evaluations identified before this work, by Broadstock.[62] She included six evaluations, all of which are considered in this chapter, and placed particular emphasis on uncertainty arising from costs, estimates of sensitivity and estimates of specificity.

Selective review of approaches to simulation modelling: results

The purpose of this review was to identify additional decision-analytic models (i.e. excluding those referred to already in this chapter) that relate broadly to the area of cervical screening and that may be of some relevance to the review of automated devices. The inclusion criteria for the review just reported were thus relaxed to accommodate not only automated screening, but also the impact of other associated technologies that may have a bearing on how automation is deployed. Thus, models that were specifically developed to look at LBC, HPV or a generic new technology have been appraised here.

Interventions

Three studies were included.[38,41,64] These modelled the impact of introducing a generic new

technology,[64] LBC,[38] and HPV testing.[41] Full details of the data abstracted from each study are provided in *Tables 13, 14* and *15*, respectively.

The first study, by Myers and colleagues,[64] was a US study, and the other two, by Payne and colleagues[38] and Cuzick and colleagues,[41] were both UK studies. Each was compared to the existing conventional screening system as a primary screening system, but in addition HPV testing was considered in combination with cytology and in surveillance. The study by Myers and colleagues[64] was an extension of the work carried out for the AHCPR by McCrory and co-workers,[12] but as the Myers study did not directly model automation it was excluded from the analysis of cost-effectiveness.

Type of model

All three papers used a variation of a Markov model, with the study by Cuzick using a semi-Markov approach.[41] Explanations of the key issues relating to Markov models are given in Chapter 5.

Selected assumptions

There were marked similarities and some differences between the assumptions of the three models. In terms of natural history, HPV is used as the only aetiological factor of cervical cancer in the model produced by Myers,[64] 95% of the time in that developed by Cuzick[41] and not at all by Payne.[38] A constant incidence of CIN was assumed in the latter. The progression and regression rates of CIN were either dependent on the severity of the CIN (i.e. state dependent)[38] or both state and age dependent, as in the other two models.[41,64]

Clinical effectiveness data

The standard of input data used on test effectiveness varied across the three studies. Payne and colleagues,[38] whose model was based on previous work by Sherlaw-Johnson and co-workers,[80] used the corresponding data on performance characteristics. The LBC was modelled to increase Pap test sensitivity by 15% for CIN2/3, but only by 2% for CIN3. In the US group's model,[64] the performance of the Pap test was considered constant across all disease categories, and the new test would add to the Pap test sensitivity. The knowledge on the natural history, prevalence and performance characteristics of HPV testing was sufficiently uncertain for Cuzick and colleagues to adopt a 'best case' to 'worst case' approach to their model (for details see the evidence tables).[41] The specificity of the Pap test was not varied in any of the models.

Input costs

All were from a health system perspective. The two UK studies[38,41] based their costs around UK hospital trusts and data from the Department of Health, while the US study[64] based its cost information on MEDSTAT and Medicare claims data. The marginal costs for the new technology and LBC were $0–15 and £3.50, respectively. The base year for costs was US$1997 and UK£1999. In Cuzick's review of HPV screening,[41] the test cost for the Pap test is £16 and for the HPV test is £17, giving a marginal cost of £1 for changing from one primary screening mode to another, but the base year of costing was not given. Discount rates varied between studies, and between costs and health benefits.

Outcomes

Myers and colleagues found that increasing sensitivity of the Pap test from 51% to 99% resulted in an ICER of US$7206, for a 3-year cycle.[64] This is without any extra unit cost assigned to the new technology. LBC incurred an additional cost per additional life-year saved of £2500 for a 3-year cycle. It also reduced the inadequate smear rates and reduced the lifetime number of smears by 6%, although this appeared to be at the expense of a corresponding increase of 6% in the lifetime number of colposcopies. The cost-effectiveness of HPV testing as a primary screening tool or in combination with cytology very much depended on whether a worst case or best case scenario was chosen. However, in surveillance the HPV testing of mild to borderline cases in follow-up would improve the mortality reduction in both scenarios. This has provided the impetus for the current pilot studies in the UK.

Sensitivity analyses

The models' outcomes were found to be sensitive to the disease progression rates, the sensitivity and specificity of the test, the discount rate and the cost of the new technology.

Selective review of approaches to simulation modelling: conclusions

All of the studies appraised here used either Markov or a variation on Markov models. Cohort simulation with fixed time cycles was used, whereby the entire cohort moved through the model at fixed points in time. Such models may be inappropriate for the case of cervical screening,

TABLE 13 *Abstracted data for the model by Myers and colleagues: generic automated technology*

Study model	Intervention strategy	Assumptions and clinical effectiveness data	Input costing data
Myers *et al.*, 2000[64] 320-state Markov model	A new technology vs conventional smear The new technology was not defined, but has been assumed to be either LBC or automation for the purpose of this appraisal Note: LSIL+ received immediate referral for colposcopy and was treated The main outcome measure for the comparison of introducing a new technology into the screening programme is the incremental cost per life-year saved. The new technologies are treated generically, assuming that their effect would be to reduce the FNR by 40–90%, but with a reduction in specificity by up to 25%	*Assumptions* Baseline values and the assumptions remain the same as McCrory *et al.*[12] New technologies are treated generically All women with abnormalities received appropriate follow-up Colposcopy and biopsy sensitivity is 100% Sensitivity of cytology and pelvic examination for stages 2–4 or cervical cancer is 100% Abnormal smears were defined as ASCUS+ HPV was considered the main aetiological factor (although serotypes were not distinguished), so age-dependent incidences and regression rates were included as part of the natural history in the model Rate of progression to LSIL was modelled as constant Rates of progression/regression of LSIL were age dependent Rates of progression/regression of HSIL were constant Adenocarcinomas were not dealt with separately *Clinical effectiveness data* Test / Metric / Base case Pap test / Sensitivity across all types of CIN / 51% Pap test / Specificity across all types of CIN / 97% New technology / Increase in sensitivity / 40–90% New technology / Decrease in specificity / 0% The first two entries are based on a number of meta-analyses, two of which relate to the AHCPR report (McCrory *et al.*[12]), which this work has followed	Health system-based costs, societal costs not included Base year for all costs was US 1997 and discounting 3% p.a. Costs estimated by using primary claims and secondary data sources. Costs for 20–64-year-olds estimated from claims data from the MEDSTAT group. Costs for ages 65 and older estimated on a combination of sources, including Medicare and various ancillary service fee schedules The perspective was to consider the costs of whole episodes of care and not just estimates based on procedure specific costs (thus complications and co-morbidity are included in the former but not the latter) Conventional screening costs were considered to be lower than other estimates, but diagnosis and treatment costs for cervical cancer substantially higher Marginal cost of new technology range = US$0–15
Results			**Sensitivity analysis**
Insufficient data to perform a full quantitative analysis Increasing sensitivity of the test from 51% to 99% incurred an ICER of US$7206 for a 3-year cycle and US$194,083 for a 1-year screening cycle, without any extra cost to the test Costs increase as specificity decreases Multiway analysis demonstrated that for a given cost-effectiveness ratio threshold, the sensitivity and specificity would have to increase in combination towards 100% as the screening interval was decreased. This is without considering any marginal cost of the new technology. At any marginal cost above US$3, cost savings were not possible			A one-, two-, three-way analysis was performed for different combinations of sensitivity and specificity at 1-, 2- and 3-year screening cycles Different thresholds of incremental cost per life-year saved were also defined at $25,000, $50,000 and $75,000

continued

TABLE 13 *Abstracted data for the model by Myers and colleagues: generic automated technology (cont'd)*

Conclusions	Comments
The broad results were that the new technologies with higher sensitivities, even at low marginal costs per slide, had large ICERs regardless of the screening interval. The authors suggest that this reflects the natural history of the disease, as most low-grade abnormalities would regress without intervention Furthermore, a small decrement in the specificity of new technology over the conventional Pap test greatly increases the ICER. This leads the authors to suggest that the pursuit of improved sensitivity alone is insufficient to improve the cost-effectiveness; the new technology must have an associated improvement of specificity	The shortcomings of this analysis are the lack of inclusion of societal costs and quality of life measures The model also assumes the same incremental change in test performance characteristics across the full spectrum of disease The study suffers from the reporting constraints (in particular the limited space) encountered by publishing the results in a popular journal. There is obviously more information that could not be reported owing to the space constraint The use of multiway sensitivity analyses demonstrates the interaction between the different varying parameters The next stage of improvement on such a model would be to find a way to model societal costs and quality of life issues

p.a., per annum

TABLE 14 *Abstracted data for the model by Payne and colleagues: LBC*

Study model	Intervention strategy	Assumptions and clinical effectiveness data	Input costing data
Payne *et al.*, 2000[38] Markov model	LBC screening vs conventional screening Two separate intervention policies considered: Policy A: immediate colposcopy for abnormal smears of borderline or above Policy B: immediate colposcopy for moderate/severe and rescreen at 6 months for borderline/mild. A colposcopy to follow a second borderline diagnosis or above	*Assumptions* Cohort of 100,000 women followed from the age of 18 to 95 years Screened between the ages of 18 and 64 years State transition model based on Sherlaw-Johnson *et al.*[80] (reference) States calculated on a 6-monthly basis Disease progresses through each of CIN1 to CIN3 before becoming invasive. Rates of progression/regression are independent of age Regression may only occur from CIN1 Incidence of CIN1 is constant for all ages until 64 years, then, zero Risk of mortality from invasive cancer is constant Coverage is estimated to be 85% Women either attend the screening programme regularly or not at all Inadequate slides are assumed to require an immediate rescreen. Subsequent slides are assumed to be adequate Borderline and mild smears are treated together Colposcopies are assumed to be 100% sensitive and specific *Clinical effectiveness data* Taken directly from the Sherlaw-Johnson model for the conventional Pap test characteristics LBC modelled to give an absolute increase in sensitivity: for CIN1 and CIN2 lesions of 15%; for CIN3 and invasive of 2% The specificity of LBC is taken to be the same as conventional screening = 98% baseline	Health system-based costs Total direct costs for tests, diagnosis and treatment are used in model and have been taken from a typical NHS trust, the Department of Health and manufacturers. Conventional screening costs are taken from the NHSCSP 1994, projected forward to a base year of 1999 Baseline marginal costs for LBC = £3.50 These are likely to be underestimated as they do not include transportation, storage and training costs Discount rates for costs 6%, and for health benefits 1.5%

continued

TABLE 14 *Abstracted data for the model by Payne and colleagues: LBC (cont'd)*

Results									**Sensitivity analysis**
Outcome: ICER, average lifetime number of colposcopies per woman and incidence of cancer (outcomes are given for policy B only)									The model was found to be sensitive to the following key variables: disease progression rates, sensitivity of the tests, improvement in the inadequacy rate, marginal cost of LBC and discount rate taken
	Incremental cost per additional life-year saved, ICER (£/life-year)			Average lifetime no. of colposcopies per woman			Annual incidence of invasive cancer (%)		
	5 years	3 years	2 years	5 years	3 years	2 years	5 years	3 years	2 years
Conventional	1,197	31,519	342,358	0.086	0.104		0.018	0.014	0.013
LBC	1,095	2,522	4,446	0.091	0.110		0.016	0.013	0.012
Change (%)				+6%	+6%		–11%	–7%	–8%

Conclusions	**Comments**
The introduction of LBC would have an incremental cost of around £2500 and £1000 per life-year gained for a 3-year and 5-year screening cycle respectively. However, these results are sensitive to the disease progression and discount rates used Other improvements in the programme (increasing coverage, more effective sample collection) may need to be considered Further research, particularly on a low-prevalence population, would add further information to the model Owing to the uncertainties surrounding certain values and assumptions in the model, further research is recommended in the following areas: marginal cost per sample of the new technologies over conventional screening; and improvements in the inadequate rate and the relative specificity of LBC techniques	The following problems with the model were identified: A constant incidence of CIN is assumed, with no age variation The progression of CIN to invasive is constant for all ages No steady state was achieved before input to the model The costs for treating invasive carcinoma may be underestimated The baseline smear costs may be high The marginal costs for LBC do not include training, storage and transportation costs Incremental changes in sensitivity were not accompanied by a corresponding decremental change in specificity LBC would effectively decrease the lifetime number of smears taken by 6%, owing to a projected decrease in the number of inadequate smears, but as a result would increase the lifetime number of colposcopies by 6% All conclusions must be treated with caution owing to the large uncertainty surrounding the input data (particularly test performance) and only marginal benefits being demonstrable

TABLE 15 *Abstracted data for the model by Cuzick and colleagues: HPV*

Study model	**Intervention strategy**	**Assumptions, clinical effectiveness data and input cost data**
Cuzick *et al.*, 1999[41] Semi-Markov (MISCAN)	Four strategies considered Cytology alone Cytology + HPV HPV as a stand-alone primary screening test Adding HPV testing to surveillance	Population: cohort born in 1955 followed Population screened between the ages of 20 and 64 years at 3- and 5-year intervals CIN may occur and progress without HPV infection, but once the CIN has developed after HPV infection, the HPV is not permitted to clear Percentage of invasive cancers that are HPV positive is fixed at 95% Lifetime risk of developing low-grade CIN and high-grade CIN is set at 15% and 5%, respectively Coverage for screening 85%, 10-year coverage 95% Surveillance leads to treatment if there are three consecutive borderline, two consecutive mild dysplasia, one borderline and one mild dysplasia, or one moderate/severe dysplasia smears Surveillance ends and the woman returns to the screening programme with two consecutive negatives Owing to incomplete evidence on natural history, prevalence and test performance, a best case and worst case scenario, represented by model A and model B, respectively, were used. The essential differences are shown below

continued

TABLE 15 *Abstracted data for the model by Cuzick and colleagues: HPV (cont'd)*

Clinical effectiveness	**Cost data**
Duration of detectable preclinical phase over Pap smears (years): A, 10; B, 1 Period before clearance of HPV (years): A, 1; B, 10 HPV sensitivity per disease category (%): Normal: A, 90; B, 50 Low-grade CIN: A, 90; B, 50 High-grade CIN: A, 90; B, 60 Invasive: A, 90; B, 70 HPV positivity in cytologically negative women per age group (%): 20–25 years: A, 15; B, 20 30 years: A, 5; B, 8 40 years: A, 3; B, 6	Health system-based costs. Cost of HPV test £17, cost of Pap test £16. Taken from Watford General Hospital (smear capacity 60,000 p.a.) annual report 1988 Management of CIN taken from the NHS price tariff for Healthcare Resources Group. Curative treatment costs taken from van Ballegooijen *et al.*[40] No discounting and no base year for analysis given

Results

Outcome: cost per life-year gained over no screening and percentage reduction in mortality

	Cost (£) per life-year gained				% reduction in mortality			
	Model A		Model B		Model A		Model B	
	3 years	5 years	3 years	5 years	3 years	5 years	3 years	5 years
Cytology	390	155	390	155	76	64	76	64
HPV+ cytology	900	420	1050	570	88	83	85	77
HPV	400	100	715	425	79	70	68	55
Surveillance (existing)	390		390		76		76	
Surveillance (HPV)	325		360		86		84	

Note: costs per life-years gained are given relative to a policy of no screening and not relative to alternative screening strategies. Therefore, they are not ICERs

Conclusions	**Comments**
Increasing the frequency of conventional Pap smear screening from 5 years to 3 years improves the mortality rates at the expense of cost-effectiveness Using model A, the combined test of cytology and HPV testing on a 5-yearly interval decreased mortality greater than on the existing 3-yearly Pap test The HPV test only, carried out at 5-yearly intervals, decreased the mortality by 70% compared with 76% for the Pap test on a 3-yearly interval, but the corresponding costs for the HPV test were 75% lower than the Pap test. In model B, however, the use of HPV testing either to supplement or to replace cytology resulted in worse cost-effectiveness rates than cytology screening In surveillance the HPV testing of mild to borderline cases in follow-up would improve the mortality reduction in both model versions	The transition probabilities from one disease to the next, and the dwelling times in a particular state, are age dependent. These have been used to fit data from British Colombia. Details of these parameters are not given in the report It is not clear whether the values for HPV positivity in cytologically negative women refer to HPV positivity in actually disease-free women (as defined by the model), or allow for false-negative cytology. In the former instance, some of the values taken for model B (the worst case) still appear low compared to those reported in the prevalence chapter of the review The model demonstrates the gaps in knowledge on age-specific HPV prevalence and incidence in different subpopulations; these gaps need to be filled before accurate estimates (i.e. with less uncertainty) of the impact of HPV screening can be made

but provide a useful starting point for consideration of the decision problem relating to automation.

The models have not allowed a steady state to establish before outcomes are measured. At any point in time there is a distribution of age groups entering, continuing or leaving the programme. To follow a cohort where everyone is at the same age is unrealistic.

The problems with modelling a cervical screening programme start with adequately dealing with the disease's natural history without oversimplifying key factors that have significant impacts on outcomes. As the natural history of cervical cancer is not totally understood, sensitivity analyses on natural history parameters such as the incidence of HPV and CIN are important.

Chapter 5

Evaluating automated cervical screening: methodological issues

Summary of key points

Clinical effectiveness

Clinical effectiveness of a screening test requires more than assessment of sensitivity and specificity. Past reviews and health technology assessments have tended to concentrate on these outcome measures. This technology assessment tried to respond by looking specifically at evidence on reproducibility, impact on process (especially average time to process a slide and rejection rates) and impact on health outcomes.

The report anticipates that the current NHSCSP may be in a state of change. The norm against which introduction of automation should be judged may not be just the current manual system, but a system which incorporates LBC.

Cervical screening is a highly complex intervention. It is tempting to assume that all manual programmes are alike. However, small differences in the quality assurance procedures (e.g. 10% random rescreening as opposed to rapid rescreening of all negatives and inadequates) may have major implications for the interpretation and generalisability of research undertaken in one country to another. This technology assessment responded to this by documenting as closely as possible the precise nature of the screening programme against which any system involving automation was being compared.

Complexity of intervention is also an issue for the intervention, automation and how it is incorporated into the existing system (primary screening and rescreening). This, in turn, may affect whether the impact has to be assessed on a whole-system basis or whether it is adequate to consider the output of the specific stage of the screening process in which automation is involved.

The fact that the automated devices differ and are in a state of development is also important. The response of this technology assessment was again to pay particular attention to the nature of the intervention.

Within-subjects designs (such as those commonly used to assess sensitivity and specificity) are poorly adapted to giving an indication of the ability to prevent invasive disease and death in cervical cancer. Thus, it does not seem sensible that assessments of the effectiveness of automated cervical screening should focus solely on such designs. Between-subjects designs are better suited to measure health outcomes. Random allocation is ideal, but may be practically difficult; historically controlled studies, although subject to greater bias, may offer a useful alternative source of evidence.

The most valid research design that is practically achievable in making an assessment of the impact of introducing automation in cervical screening depends on the nature of the outcome. An attempt was made to define the range of study designs that would provide some evidence of reasonable validity for each of the groups of outcomes previously identified:

- sensitivity/specificity: within-subject study design ideal
- reproducibility: within-subject study design ideal; needs to consider alternative sources of variation, especially machine–machine and observer–observer; no ideal statistical measure, but kappa used in this project
- process: difficult to be dogmatic on which study design is likely to provide the most valid information; both within-subject and between-subjects study designs were considered
- health outcomes: between-subjects design, and RCTs in particular, ideal; however, historically controlled (pre–post) studies also sought and included.

Studies evaluating sensitivity and specificity need to be carefully assessed with respect to how open they are to spectrum and verification bias. The nature of the reference standard also needs consideration, as does whether a two-armed study design is more appropriate where there is uncertainty about what constitutes the gold standard for assessment of disease state.

Problems also occur in the analysis of results of evaluations of test performance. Particular care was taken in this assessment concerning clear definition of both the test threshold and the threshold used to designate presence of disease.

The possible effect of publication bias has not been examined in previous reviews and health technology assessments. A component of this report was devoted to exploring how great a threat unpublished literature might be to conclusions based on published and more easily accessible reports of evaluations of automated cervical screening alone.

Past reviews have tended to exclude non-English language articles; this review deliberately searched for evaluations that may have been published in languages other than English.

Health economics

The models that have previously been used to estimate the cost-effectiveness of introducing automation and related technology have generally been Markov models. The main methodological issue from a health economic perspective was whether there are more appropriate approaches.

Discrete event simulation (DES) offers advantages that may be particularly applicable to evaluating the cost-effectiveness of introducing automated image analysis, and was the modelling approach attempted in this project.

Introduction and objective

Several issues, some of which were anticipated when preparing the protocol and some arising from the detailed appraisal of the literature reported in Chapters 3 and 4, require further explanation. These are expanded upon below. The authors have emphasised how particular issues have informed the conduct of this project. This chapter has been divided into two parts: issues concerning clinical effectiveness, and issues concerning cost-effectiveness.

Part 1: Clinical effectiveness

Defining the clinical effectiveness of a new screening test

In their treatment of screening tests, Sackett and colleagues[81] give the following guidelines as to whether a screening test does more harm than good:

- Does early diagnosis really lead to improved survival, or quality of life, or both?
- Are the early diagnosed patients willing partners in the treatment strategy?
- Is the time and energy it will take us to confirm the diagnosis and provide (lifelong) care well spent?
- Do the frequency and severity of the target disorder warrant this degree of effort and expenditure?

The introduction of automation will not alter the nature of the existing test, it will only alter its interpretation. Yet, despite the reported success of the conventional Pap smear, which with time has provided affirmative answers to all of these questions, there may be a need to re-address these questions if automated systems are introduced.

In such a case, it is not certain whether the answer to any of these questions would be yes. If the new technology were to produce greater false positives, and thereby increased invasive investigation, or miss important lesions or only identify lesions at a stage where their course could not be altered, then the undoubted extra cost involved in converting to automated screening would dictate no future for such a technology without its further development.

The need to evaluate a new technology to ensure that it not only satisfies these guidelines, but also constitutes an improvement on the existing test is imperative. A screening test represents a special type of diagnostic test, which unlike a standard diagnostic test is not about reducing the uncertainty of diagnosis in the sick and symptomatic, but is concerned with identifying accurately potential lesions in the presymptomatic individual.[82] However, the standards required of a screening test are the same as those demanded of a diagnostic test.

It is apparent from Chapter 3 that reviewers in the main have concentrated on measuring the clinical effectiveness of automated cervical screening in terms of sensitivity and specificity or their equivalent. However, while these are of importance, other features of a screening test also need to be considered.

They need, for instance, to give an indication of how consistently the test performs. Reproducibility is one of seven standards described by Reid and colleagues[83] that should be met when evaluating a diagnostic test, but it applies equally to a screening test. For this reason, the present authors

decided to assess evidence on reproducibility as part of evaluating the clinical effectiveness of automated cervical screening.

In this report, the interpretation of clinical effectiveness has been extended to include impact on process; that is, how many slides need to be processed in order to screen 100,000 women. This is partly in the light of work from the forerunning reviewers, indicating that differences in sensitivity and specificity between the new technology and the existing conventional system are clearly marginal, suggesting equivalence rather than superiority. If, however, equivalence in accuracy is associated with significantly shorter processing times, then the technology could be perceived to be more effective. This is particularly true where a key current concern in the UK, and probably other countries too, is that the number of slides needing to be processed apparently exceeds the system's capacity to train and retain sufficient technical staff to read the slides.

There may also be specific drawbacks with the new technology, such as increased rejection rates and lower tolerance levels to imperfections in slide preparation, which would not be associated with the existing system. These elements have also been included in the appraisal of clinical effectiveness.

Finally, undeniably the highest level of impact that could be sought on the clinical effectiveness of automated cervical screening devices would be on health outcomes; that is, incidence or mortality from cervical cancer. Ideally, this would require an RCT comparing populations exposed to conventional screening, with those with a screening system incorporating automated devices. From past reviews it seemed unlikely that such an RCT had taken place or was in progress. However, in the absence of RCTs it appeared that past reviews had not considered other research designs that might assess such outcomes. Thus, to maximise the chance of capturing any reasonably valid information on outcomes of such clear importance, the authors specifically looked for and included evidence on impact on health outcomes from quasi-experimental research designs, particularly historically controlled studies (i.e. pre–post studies or interrupted time series). Historically controlled trials were targeted because it was likely that experimental study designs with a parallel control group might be very difficult to organise. In contrast, it was felt plausible that following first FDA approval for a number of automated devices in the mid-1990s, health systems locally, regionally or nationally may have systematically introduced a change from conventional to automated screening and evaluated its impact.

In summary, in evaluating the clinical effectiveness of automated cervical screening systems, this study has specifically evaluated

- test performance in terms of sensitivity and specificity
- test reproducibility
- impact on process measures, such as processing time
- impact on clinical outcomes, particularly as measured by historically controlled trials.

Nature of intervention and comparators

The current screening programme in the NHS, and indeed most other organised screening programmes, consists of the conventional Pap smears being examined by trained screening laboratory staff. Yet, other technologies may not be far off the horizon (e.g. LBC and HPV testing) and it is possible that these will become part of mainstream practice. Based on consultations with the NHSCSP, the possible combinations of deployment of these three technologies are illustrated in *Figure 5*.

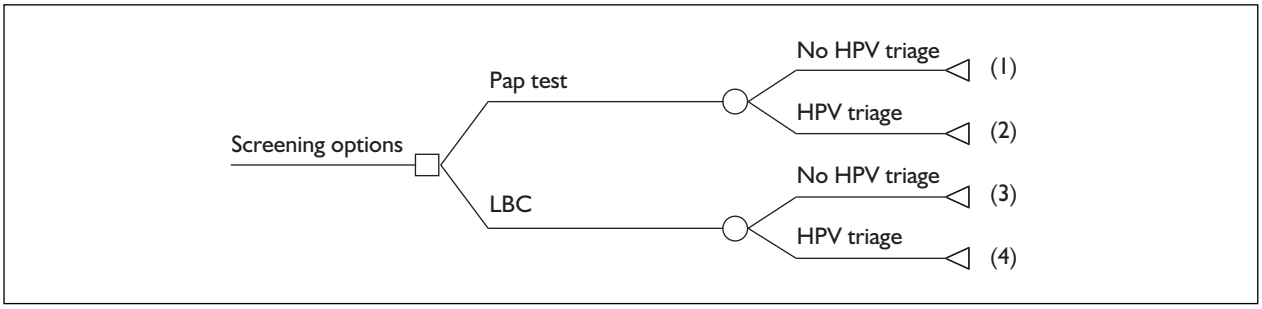


FIGURE 5 *Possible future screening strategies: the base cases, without automation*

Thus, either the existing Pap test or LBC would be the primary screening technology and HPV triage would act as an adjunct to either of these. In the UK HPV screening is currently being piloted for the triage of borderline and low-grade cases, and is represented by options (2) and (4). Note that primary HPV screening is not considered here as its effectiveness is still very unclear, and it represents a far more distant possibility for implementation (see Chapter 3).

These are the base cases into which the effect of adding automation will be assessed. It should be noted that incorporated within this strategy, although not shown, is that all negative and inadequate cases in the UK are subject to a rapid review.

In *Figure 6* the ways in which automation may impinge on the four main alternatives are added. This generates 12 potential combinations of the main competing technologies. Of these, the comparison of Pap or LBC + no HPV triage + automation (primary/rescreening) (options 2, 3, 8 and 9) versus Pap or LBC + no HPV triage + no automation (options 1 and 7) represents the one that would currently be of greatest interest to policy makers in the UK.

It should be noted with respect to this diagram that the individual screening options do not always summarise the individual stages of the process that are sometimes found in comparative studies. For instance, an option with 'no automation' consists of the primary screening of all slides followed by the rapid reviewing of the negative and inadequate slides, followed by their final reporting. This results in a number of possible ways in which primary automated screening may be compared with no automation in a study. These include comparing the results of the whole systems, one with automation and one without, or confining the comparison to the individual stage of the process in which the device is deployed with the corresponding stage in the existing system.

This latter type of comparison is especially common when using automation as a rescreening device. Ideally, in this instance the yield of rescreening negatives/inadequates with an automated system would be compared with the yield from the rapid review of such slides as used in the UK. Rapid review is the method of rescreening in New Zealand, and as a result Broadstock, in her review,[62] justifiably excluded rescreening trials where the control arm did not involve rapid review.

The vast majority of studies in the field arise from the USA, where the screening system involves the rescreening of a 10% random sample of negatives as a quality control measure. The UK uses rapid

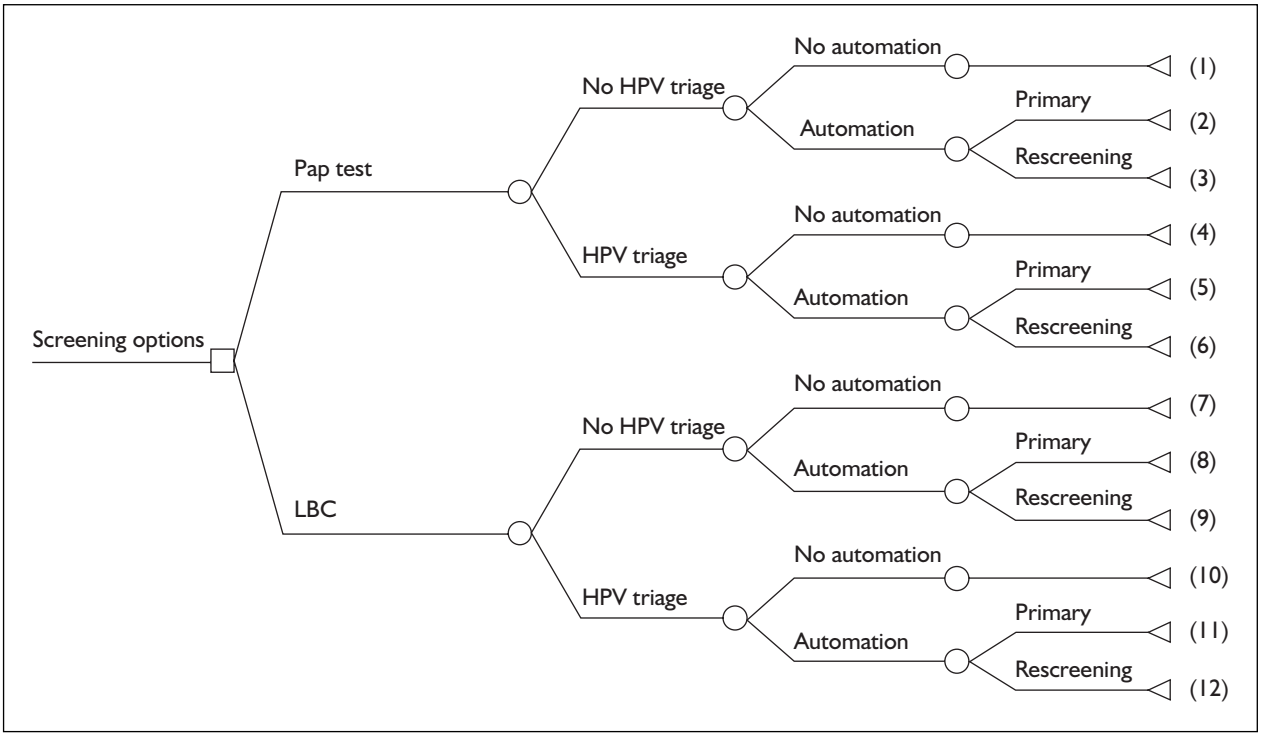


FIGURE 6 *Possible future screening strategies: base cases incorporating automation as either a primary screening or a rescreening tool*

rescreening with the belief that it is more effective than the system in the USA. Therefore, it may be argued that a rescreening device that does not prove itself to be more effective than the US rescreening system will have nothing to offer the UK system of screening. For that reason, these studies have been included.

In summary, both the intervention and potential comparators are highly complex. This has implications for interpretation and generalisability of results, and the technology appraisal attempted to respond to this by abstracting as much detail as possible about the nature of the interventions and the comparators in the research identified in the review of effectiveness. Anticipating that the NHSCSP may be in a state of change, an attempt was made to anticipate the need to identify and review research that used a screening system incorporating LBC as the comparator.

General issues concerning study design

In evaluating the clinical effectiveness of automated cervical screening technology (see Chapter 3), reviewers have tended to target studies of **within-subject design**, where both arms (no automation arm and automated arm) are applied to the same set of slides. This is in accordance with guidelines produced by the Intersociety Working Group for Cytology Technologies.[84,85] However, this assumes that sensitivity and specificity are the only dimensions of effectiveness that need to be examined.

Two of the main objectives of cervical screening programmes are to prevent invasive disease and to prevent deaths from cervical cancer. When using a within-subjects design there is an a priori assumption that the detection of preinvasive disease prevents the occurrence of invasive disease. Thus, detecting invasive disease in a subject represents a failure to meet one of the objectives, as it should have been detected in the preinvasive phase, and its non-detection still represents a failure, since invasive disease has occurred but has not been detected yet. In using an outcome such as the number of cervical cancers detected, a within-subject design does not allow for either technology to demonstrate its superiority over the other.

If the number of preinvasive lesions detected is accepted as a reasonable surrogate marker for the prevention of invasive disease, then the two arms can be compared in a within-subject design. It is well recognised in the literature that preinvasive disease does lead to invasive disease and such a design is valid on those grounds. However, the proportion of, say, HSIL+ that progresses to invasive cancer is still not accurately known. Thus, the key point about studies of within-subject design is that they do not allow for the direct measurement of the prevention of disease or death.

The effect that a device has on the level of invasive cervical cancer in the local population can be ascertained directly in a between-subjects design. In this instance the subjects are tested by one device only, and not both. The problems posed by this design are ensuring that both study populations (one for each arm) have the same risks of developing invasive disease before intervention (screening). The best way of achieving this is by random allocation.

Thus, for measuring the impact of automation in preventing cervical cancer, the best study design is an RCT with an outcome measure of invasive disease detected. Alternatively, the number of deaths from cervical cancer could be used as an outcome measure, if the effect that automation has on the mortality from cervical cancer is the objective.

Random allocation represents the most effective way of ensuring that the two populations start with the same risks of invasive disease. Unfortunately, such studies will need to be larger than their within-subject counterparts, as the prevalence of invasive disease is somewhat lower than preinvasive disease, and to demonstrate a significant difference in the number of invasive lesions detected between the arms requires large samples.

The alternative is to use non-random allocation, but ensuring that populations have similar risk is more difficult. One approach less valid than the RCT is to use historical controls. Here, the arms are not contemporaneous, but separated by time. A screening centre elects to change its technology, so the new test arm represents all the slides screened by the automated system from that point onwards. The control arm represents a similar period where the conventional system was still in place before the change. The assumption here is that the risk characteristics of the population under investigation have not changed significantly over the period of the investigation. This may not be an unreasonable assumption where whole populations in a defined area are being considered, and the likelihood of differential selection in the pre- and post-period seems low.

More difficult is the possibility that something other than the introduction of automation may have occurred between the pre- and post-period, and also that the likelihood of detection bias may be greater, as it is hard to envisage how outcomes can be assessed blind of knowledge of whether the outcome is being assessed in the intervention or control period. As in the RCT, a suitable outcome measure could be the number of invasive lesions detected or mortality from cervical cancer.

The key points about between-subjects designs, therefore, are that they allow for the direct measurement of prevention of disease or death, and that the RCT represents the most valid way of doing this, but historically controlled studies are more likely to be carried out, and can still provide valuable information provided one is aware of the biases that may be operating and interprets them cautiously in this light.

This section emphasises that with respect to assessing the clinical effectiveness of automated cervical screening technologies the most appropriate study designs to target depend on the dimensions of effectiveness that one is trying to assess. The next section discusses which study designs may be the most appropriate for each of the groups of outcomes previously identified.

Appropriate study designs for the different elements of clinical effectiveness

Test performance

Producing an accurate estimate for the sensitivity and specificity of the automated systems has been the area of clinical effectiveness most evaluated by other reviewers. As discussed in the previous section, when the test is concerned with using preinvasive lesions as a surrogate marker for prevention of invasive disease, then sensitivity and specificity are suitable outcomes and a within-subject design represents the best form of study to evaluate these. There are, however, important specific issues about how such studies should be conducted, which are discussed in detail in the next section. Aside from these, it has been repeatedly noted by Reid and colleagues,[83] among others, that outcome measures should be given with confidence intervals. However, this has rarely been achieved in past reviews and this report will attempt to respond to this.

Reproducibility

An area that has not received much attention in previous reviews has been the evaluation of the new technology's ability to generate reproducible results (e.g. generate the same decision for the same slide presented on two occasions). This may not have been considered because it may appear obvious that a system that is not dependent on human interpretation is necessarily likely to be more reproducible than the current manual system, which is completely dependent on human interpretation (and subject to its inevitable frailties). However, this assumption needs to be tested. Low reproducibility not only is a failing in it its own right, but will limit test performance.

The fact that reproducibility requires the input to be held constant (e.g. the same set of slides) immediately suggests that a within-subject design is the most appropriate. However, there are different potential sources of variation that may affect an automated cervical screening system in respect of which reproducibility needs to be assessed.

Machine–machine

Automated devices are not perfect and some variation in performance should be expected and quantified. There is intramachine performance, where the ability of the machine to produce the same set of results on the same slide scanned on two separate occasions is assessed. Similarly, there is the intermachine performance, where the ability of two different devices (but of the same manufacturer's specification) to give the same result on the same slide is evaluated.

Observer–observer

In addition to machine–machine variation, which is relevant to all automated systems, observer–observer variation may need to be considered. The main object of all automated devices is to generate a targeted subset of the slides for further manual examination. However, some of the automated systems display image tiles intended to summarise the most abnormal features of the slide scanned. This interface requires a screener to make an interpretation and a decision based on the images displayed, therefore introducing intraobserver and interobserver variation. The PAPNET and the AutoCyte SCREEN would be subject to this type of variation. Thus, as with the existing manual system, the potential for both interobserver and intraobserver error needs to be assessed. This is the variation in results that occurs when different screeners, or the same screeners at different times, view the same slides or computer-generated images.

Machine+observer–machine+observer

The observation that different automated devices interface with the human interpretive component

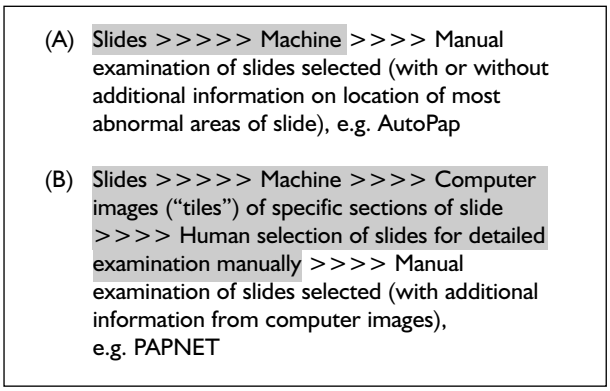


FIGURE 7 *Illustration of the varying number of steps in the automated process whose reproducibility may be assessed with different automated image analysis systems. Shaded areas represent that portion of the process that may be considered to be 'automated'.*

to different degrees in different ways (*Figure 7*) suggests that what should be being measured to give a fair comparison across different automated devices is the combined variation attributable to both machine and observer components. Further, this variation, in turn, should be compared with the variation that would be seen without the involvement of automated devices.

As part of the review, studies that evaluated the reproducibility of the automated cervical screening systems under the conditions mentioned above were considered for appraisal. Unfortunately, there is no statistical test that is ideally suited for measuring agreement between two observations. The one used and probably the one most used in clinical studies is a variant of the kappa score, which corrects for agreement beyond chance. Where possible in this review, weighted kappa scores were calculated from the data presented in the article for comparability of the arms. However, whenever a weighted or standard kappa score is quoted in the review the following limitations should be borne in mind:

- The score is dependent on a number of parameters, including disease prevalence, making interstudy comparisons more difficult.
- It does not give any recognition to the level of discordance unless it is weighted.
- What constitutes good and poor agreement is not defined absolutely. Fleiss suggests values greater than 0.75 as representing excellent agreement beyond chance and values below 0.40 as representing poor agreement beyond chance. Values between these limits are taken to represent fair to good agreement beyond chance.[86]
- It is not an appropriate statistic for the evaluation of new test performance (see previous section).

Process

Another area that has received scant attention in past reviews is the effect that a new technology would have on the day-to-day operations of a laboratory. Thus, this project attempted to review systematically evaluations indicating the likely impact of automation on workflow and other measures of process compared with conventional screening. Specifically, information was sought on:

- measures of workflow, such as average time to process a slide or the number of slides needing to be processed to screen a population of a given size
- other workflow consequences, such as the number of slides that the new system is unable to process
- changes to the system.

It is difficult to be dogmatic about what sort of study design may be feasible and provide valid indications of the impact of the introduction of automation on process. A priori, it was anticipated that the design most likely to have been undertaken and to provide valid information was a historically controlled (pre–post) design. In the event, however, the authors were inclusive with respect to acceptable design, provided there was a comparator group/period, as it was anticipated that such measures would tend not to be the primary subject of the published literature and hence would be uncommon occurrences. Despite this inclusiveness, full account of threats to validity was taken in the interpretation of any studies providing information on impact on process.

Health outcomes

As already stated, it is clear that with respect to impact on health outcomes, between-subjects research designs are likely to be the most valid. RCTs would be the ideal. However, because of the practical difficulty of organising an RCT comparing a screening system incorporating automation and one relying on a traditional manual system, and the fact that such an RCT had not been encountered in previous reviews, the range of study designs sought and included for these outcomes was extended to historically controlled trials (pre–post designs). As for process measures, there was a clear recognition that there are important intrinsic threats to validity in such designs, but this was taken into account in the interpretation of their results.

Particular issues in assessing test performance

As described in Chapter 3, several reviewers[12,38,52,53,62] have noted important deficiencies in the quality of evidence underpinning estimates of the sensitivity and specificity of automation and related new technologies in cervical screening. The following potential threats to validity and problems were also carefully considered in this health technology assessment.

Spectrum bias

In evaluating a diagnostic test, the importance of the spectral composition of the study population being representative of that encountered in practice has been highlighted by a number of authors.[87,88] When this is not the case, irrespective of whether a diagnostic or screening test is being evaluated, spectrum bias may arise.

The term spectrum bias encompasses the phenomenon that the sensitivity and/or specificity of a test varies with the different populations tested.[89] The different populations are distinguished by a range of features that differentiate the diseased from the non-diseased.[90] These features could be pathological, such as a study population where the diseased group consists mainly of HSIL+, which are easier to identify than the lower grade lesions more likely to be found in a normal screening population. They could arise from co-morbidity, and in cervical cytology age could be considered a co-morbid factor, since with increasing age there is an increasing likelihood of a woman developing a withdrawn atrophic cervix, which can hide potential lesions from sampling. Thus, the distribution of age groups tested should reflect that encountered in practice. These and other features could increase or decrease the likelihood of false negatives and false positives of the cytology test, and should be considered as a potential cause of variation in estimates of sensitivity and specificity from one study to another. It should be noted that equal prevalence of 'disease' between the study sample and the actual screened population does not necessarily imply an absence of spectrum bias.

Attempts have been made to evaluate the effect of applying different study populations with different disease spectra to the performances of a test.[91] Ultimately, the effect depends on the case-mix of each of the populations. However, in the context of cervical screening it may be inferred that with all other things being equal, in a population where there is a relatively high proportion of high-grade cervical cytology, this will have the effect of biasing the estimated sensitivity upwards.

Verification bias

Verification bias arises when only a subset of the study population's test findings is verified by the gold standard. It is a flaw found in a number of studies on diagnostic tests[83] and is of relevance to cervical cytology. A common example of this is the biopsy confirmation of only those subjects with positive cytology. Ransohoff and Feinstein,[90] who described the bias as work-up bias, indicate that the effect of confirming by biopsy only those with positive cytology would be to overstate the sensitivity of the test.

This type of bias is also intrinsically linked to diagnostic review bias, where the reviewer applying the gold standard may be influenced into overdiagnosing disease when aware of the test result, hence the need for independent verification. Unfortunately, the study protocol is often known to the reviewer, and so will be the referral protocol of test positives only, even if the test result is not known. In theory, as Ransohoff and Feinstein state, "work-up bias can lead to under-diagnosis (missed abnormalities) but not to over-diagnosis".[90]

To avoid verification bias the reference standard should be applied to the whole study population, regardless of the test result, but when the reference standard is invasive, verification bias may be unavoidable (see next section). To mitigate this the reference standard may be applied to a random selection of negatives, to try to correct for it, or to consider alternative, less invasive, but often less valid reference standards for the negatives (e.g. cytology or prolonged follow-up).[92]

Alternatively, if verification bias cannot be eliminated, then another approach is to calculate its effect. Methods have been suggested that derive formulae that calculate sensitivity and specificity in the presence of verification bias. Using Bayes' theorem these metrics are calculated if the PPV and prevalence are known.[93] When all test positives are verified by the gold standard, the PPV is easily calculated; however, the prevalence of disease in the population is not usually known accurately. Following Nanda and colleagues' suggestion[65] of randomly sampling the test negatives for verification, the NPV may be estimated, and together with the PPV used to estimate the prevalence.

Reference standard

In principle, a gold standard should always determine the true level of disease in the study population. However, in cervical cytology this has proved to be extremely difficult, fraught with ethical and practical problems. The more recent reviews, in trying to address the need for clear reference standards in the studies they have considered, have often resulted in very restrictive inclusion criteria leading to no or few articles being appraised.[12,62] Whether such an approach is appropriate is debatable and the issues surrounding the choice of reference standard are discussed below.

In cervical cytology a range of abnormal outcomes is possible, from borderline and mild changes (low grade) to suspected invasive disease. The significance of mild changes is not entirely clear, but it has been reported that only 21% of low-grade cytological lesions will progress to a high-grade lesion, with over half regressing to normal over 24 months.[94] Over a 10-year period less than 1% of low-grade lesions have the potential of developing into a neoplasm, which equates to a 3-year screening programme having up to three attempts to identify the lesion in the preinvasive phase. In the worst case this would remain undetected until it manifests itself clinically, at which stage a biopsy would be performed establishing the diagnosis. This could provide the basis for the definitive gold standard. The long-term follow-up (of a decade or more) of a screening cohort would allow sufficient time to establish the level of disease in the study sample.

Understandably, in most cases this has not been adopted, for a number of reasons. To allow an abnormality to progress to a higher grade or potentially cancerous lesion without intervention has ethical implications, although such a study was carried out in New Zealand in what has become known as the "unfortunate human experiment".[95] Another reason is that, when using such a reference standard where there is protracted follow-up, there is the opportunity for *de novo* lesions to arise. It is impossible to know whether the lesion represents a new one or has been present all the time. Finally, all research is subject to resource constraints, and such follow-up could prove to be expensive. This has provided the impetus for the development of alternative reference standards. The two most often used are histology and cytology, but neither is without its problems.

Histology as a reference standard

For ethical reasons a biopsy is usually only performed on those cases with positive cytology, and therefore verification bias cannot be ruled out. In this instance the sensitivity tends to be overestimated at the expense of specificity.[90]

Further, the process of taking a biopsy is subject both to sampling error, particularly in the instance of borderline and low-grade cytology, as this may not be observable at colposcopy, and to variable interpretation leading to disagreement in diagnosis. The Intersociety Working Group for Cytology Technologies (ISWG), therefore, has suggested that any histology should be subject to a consensus diagnosis;[84,85] this would reduce the potential for both verification and diagnostic review bias without totally removing either. They advocate the use of biopsy on a "statistically significant subset of patients with positive cytology" when evaluating the PPV of a primary screening device, PPV being used to approximate specificity.

As negative smears often remain unverified by this reference standard, the level of true negatives and false negatives in the study sample cannot be known, and so neither can the sensitivity and specificity. Instead, the relative true-positive and relative false-positive rates may be calculated.[96]

The time delay between taking the smear and the biopsy is also important, as there is scope for the development of *de novo* lesions in the interim. Concurrent sampling leaves no scope for this error; however, in practice the delay may be up to 2 years, calling into question whether the histology taken later is a true reflection of the disease status of the patient when the smear was taken.

As it may be difficult to justify taking a biopsy in the case of repeated borderline abnormalities when full colposcopic view demonstrates no lesion, for the purpose of this review histology was considered an adequate reference standard if all those with a demonstrable lesion at colposcopy received a biopsy.

Cytology as a reference standard

An alternative reference standard is to use cytology, usually in the form of consensus agreement of two or more skilled screeners, cytologists or cytopathologists at a multiheaded microscope.[84,85] This assumes that a group of assessors working together is less likely to make an error than one assessor working alone.

This method has the advantage that in theory negatives may also be verified, although only rarely implemented. Often the verification process

is applied only to those discordant samples between the two study arms, which Miller argues is not without its problems.[97] One of the fundamental principles of evaluating any test is that both the test and the reference standard should be applied to the subjects, and that they are independent and blind of each other's results.[81,83] In discordant analysis the risk of the test result influencing the reference standard assessment is high.[97] Further, if concordant smears results are not verified then any measures of sensitivity and specificity are likely to be overestimated, as both slide results may be wrong.

The ISWG in its proposed guidelines for rescreening technologies,[85] suggests that the "detection of missed abnormals requires two or more skilled technologists or pathologists to independently review all putatively negative slides". Any positives identified are subject to a consensus review with at least one additional expert present.

The ISWG considers adjudicated cytology to be the appropriate gold standard to use in a study design when determining the sensitivity of the new technology. An apparent inconsistency, however, is in the treatment of concordant positive results, and may account for the number of studies that appear to ignore such cases. In primary screening, it is recommended that discordant results from the two study arms be subject to the panel review (already described) without any reference to concordant results. Yet, to calculate the sensitivity of the new technology requires full knowledge of the level of true positives and false negatives, which cannot be enumerated unless the concordant results are also subject to comparison with the reference standard. Reference is made to taking a biopsy of a subset of those with positive cytology to establish the PPV, but this relies on the subset being representative of the total number of positive cases.

In its second paper (on rescreening devices), the ISWG recommends that in effect the entire study population should be reviewed by at least two experts in the first instance, with positive slides receiving a panel review at a multiheaded microscope with an additional expert present.

For the purpose of this review, a cytological reference standard was considered adequate if the diagnosis had resulted from the consensus opinion of a panel of two or more skilled screeners, cytologists or cytopathologists who had independently assessed at least all of the discordant samples (and preferably all abnormal slides) blind of the results from the two study arms.

Whether histology or cytology is used as a reference standard, a corollary of the above is that no reference standard can be considered perfect. This, in turn, challenges whether a traditional approach of judging the advantage of one test over another by comparing each to an apparently common gold standard (often in separate studies) is feasible or acceptable. It raises the possibility that two-armed designs may be more valid. In this case, an existing and a new test are compared to a reference standard in a single study, and estimates made of the advantage that one test offers over another in the context of that study. Although there may be uncertainty about the validity of the absolute values of, say, sensitivity and specificity calculated in a two-armed design, one can be reasonably certain that identification of an advantage of one test over another is reasonably robust. For this reason and the acknowledged uncertainties about an ideal reference standard in cervical screening, two-armed study designs were preferentially sought in this review.

Defining both test threshold and diagnostic threshold

When attempting to estimate the sensitivity or specificity of any test, two definitions have to be made:

- the definition of a positive test
- the definition of disease or abnormality.

The former affects the number of positives that the test allocates, and the latter determines how many the reference standard allocates as abnormal. Both definitions may be varied, in effect changing the threshold in each. Thus, if the disease status is defined as being high-grade lesions and above (HSIL+), then the threshold for a positive test may be set at inadequate+, borderline+, LSIL+ or HSIL+. Perhaps one of the more confusing aspects of using a Bethesda system is that the same term for the threshold of a test may be used for the reference standard. Thus, the term HSIL+ could refer to those slides in which the cytology shows a high-grade lesion or worse, but it could also refer to the disease status determined by the reference standard, which may be cytology or histology. The US group[12] addressed this potential for confusion by referring to the Bethesda system of terminology when describing the cytological test, and using the histology threshold system when describing the disease status or reference standard (CIN).

Oddly, this distinguishing feature between the threshold for the test and the threshold of the reference standard has seldom been acknowledged, and only two reviews[12,59] considered it important enough to give clarification.

In this project, for the estimates of both sensitivity and specificity a range of both test thresholds and diagnostics thresholds will be used. For each estimate, however, both the threshold for the test and the reference standard will be made explicit.

Particular issues concerning reviewing the clinical effectiveness literature

Publication bias

Basing the findings of a systematic review entirely on the published literature is to risk ignoring valuable information that has not reached publication. This is important as it is possible that the results of such research may be systematically different from research that is published. Such a phenomenon is known as publication bias.[98,99] This is thought to be a demonstration of the researchers' belief that a study with equivocal results will not be published, and is therefore not worth submitting for publication.[100] It could also be that the editors of journals live up to these expectations by rejecting such studies.

In fields of interest where there is a dominant influence of manufacturers, there is significant potential for such bias. The independence of the trialists may be particularly compromised if devices have to be borrowed owing to their prohibitive purchase costs. Furthermore, the researchers may have to submit their data for review by the corporation before permission for publication is granted. Such controls may lead to the suppression of data that shed an unfavourable light on the device under study.[100] This pressure is compounded by the potential gain in revenue for a company, by demonstrating that their device significantly improves the current situation. In the field of automation, this could relate to whole countries adopting the technology across their entire health system. Independence of published studies may be further called into question when the vast majority of papers on a subject seem to be conducted by a very limited number of investigators.

Evaluating the level of publication bias in any field can be a major undertaking for a systematic review and is usually impractical with the resources available. From the review of the secondary literature in Chapter 3 it was an area highlighted as having received very little attention so far, no previous reviews having examined for its possible effect. This report has attempted to do so.

Although there is no universally agreed definition of publication bias, a well-established one by Dickersin is, "The tendency on the parts of investigators, reviewers and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings".[101] There are two aspects to this definition, the study findings and the publication. The study findings are usually interpreted in the form of statistically significant or insignificant, positive or negative, important or unimportant. These interpretations, according to Dickersin's definition, determine whether the study is likely to be published. Unfortunately, what is meant by publication is not given a great deal of explanation. This is particularly problematic when it can be increasingly argued that publication is not a dichotomous event but a continuum, with a range of media formats, lying between a highly accessible article published in a widely read journal and an incomplete researcher's manuscript located in the top drawer of his filing cabinet.[100] Therefore, a full analysis of publication bias requires us either to define what is meant by a publication or to attempt to define publication bias differently. Although other definitions and approaches have been suggested, it was decided to give the term 'published study' a rigorous definition for the purpose of this review (see Chapter 7).

To determine the existence of publication bias and measure its impact on the conclusions drawn from the published literature requires gathering indirect and direct evidence.[102] Indirect evidence comes in the form of observing that there is a higher proportion of significant results in published studies. Such an approach is not considered reliable, however, as the true proportion of significant results is unknown.[100] Also, it has been argued that, if hypotheses to be tested are not selected at random, then one can presume a higher number of studies with significant results. Other indirect evidence is provided by theoretical techniques that assume an association between publication bias and the sample size of the studies. In studies with small samples there is greater potential for random error and therefore the results are more widely spread. Funnel plots, in particular, demonstrate this phenomenon, and publication bias is considered a possibility if the normally symmetrical funnel-shaped curve is asymmetrical.[100] Unfortunately, asymmetry may not necessarily be due to publication bias; other

causes include chance, different intensity of intervention, choice of metric, poor design of small studies and fraud.

Direct evidence arises from searching for unpublished or partially published evidence. Most systematic reviews will today include some attempts to ascertain unpublished literature, by contacting researchers identifiably active in the field, but few can be more exhaustive than this. Among the possible sources of information searched in the past for evidence of publication bias are any potential investigators, charity bodies or cohort follow-up of registered studies such as those submitted for ethics committee approval.[98] Hetherington and colleagues surveyed more than 42,000 clinicians over 18 countries, which resulted in their being notified of 18 unpublished RCTs in the field of obstetrics and gynaecology (although an unpublished study was not defined).[103] Their conclusions were that "publication bias will not be addressed successfully by attempts to obtain information about unpublished studies retrospectively" and "prospective registration of trials at inception appears to be a feasible approach to reducing publication bias".

Certainly, in the future, the prospective registration of trials as a prophylactic measure should go a long way towards reducing publication bias. However, this does not address the current situation, in that there may be unpublished studies in existence that will affect conclusions of existing systematic reviews and meta-analyses. Others also argue[100] that the inclusion of identified unpublished data (from surveys or otherwise) may not necessarily reduce the bias. Their reasoning is that unless the unpublished studies retrieved are a representative sample of all unpublished studies in the field then the potential for publication bias cannot be satisfactorily solved by location alone.

Despite these concerns, a retrospective approach to identifying unpublished studies was the main approach adopted in this health technology assessment project. While accepting that there may be a danger in obtaining an unrepresentative sample of unpublished literature, it was felt that it was important to demonstrate whether or not large volumes of unpublished literature existed, given that there was such a strong suspicion that this may be the case. If unpublished literature was identified, the authors may not be confident of the effect that this may have on the overall estimates of effectiveness, but would probably be able to confirm or refute the suspicion that publication bias was a major factor needing to be taken into account in the interpretation of the more readily available research findings.

There is no accepted general methodology for undertaking an enquiry sufficiently detailed to identify reliably most unpublished literature, so a strategy that seemed appropriate to research on automated cervical screening was devised from scratch. Recognising that such an approach was in a sense a pilot and that there were time constraints, this assessment reports only the results as applied to interrogation of sources in the UK. It is hoped that this will be sufficient to assess the likelihood of significant amounts of unpublished literature and to provide the basis for future ascertainment of unpublished literature on this topic internationally if the likelihood seemed high.

Language bias

A related phenomenon to publication bias is language bias. In this, it appears that research published in non-English-language journals may have systematically different results to that in English-language journals.[68] Since previous reviews did not systematically search outside the English language, in this assessment the search was specifically extended to include articles published in languages other than English.

Part 2: Health economics

Use of DES in modelling of cervical screening

The need to go beyond what exists in the literature

The key methodological issue concerning cost-effectiveness relates to the most appropriate modelling framework to use in evaluating the efficiency of cervical screening in general and the impact of introducing automated image analysis in particular.

The existing models are essentially derived from three sources.[12,38,41] The AHCPR model by McCrory and colleagues[12] and the models by Payne and colleagues[38] are two independent Markov models. The main difference between them is that AHCPR includes HPV and Payne does not. These models operate in a fixed cycle (yearly for AHCPR and 6-monthly for Payne). Each model assumes that screening occurs at fixed ages and does not allow for surveillance screening. Payne and co-workers allow for a small number of inadequate smears, but assume immediate retesting; "subsequent slides are assumed to be adequate".

Cuzick and colleagues[41] use a semi-Markov approach; this model has the advantage that it allows individual patient histories to be followed and includes surveillance screening. This model does not seem to allow for inadequate smears, although its structure would presumably accommodate them without much difficulty.

None of the models takes account of delays in screening results caused by queuing effects. The only way this can be done is to use a simulation methodology which allows for interaction between individuals. One such methodology is DES.

DES models

A population-based DES model tracks individuals (in this case, women) from an entry point (in this case, reaching the age of 15 years) to an exit point (in this case, all women are followed through to death). Women may change state at various times through the model; it is assumed that such changes take no time. Pseudo-random numbers are used where appropriate to generate the necessary variations within the model. These relate to both time spent in a given state, and the pathway followed between states. The important feature of DES of relevance to this report is that it allows for the effects of limited resources. In this case, if the number of smears sent to the laboratory over a certain period exceeds the capacity to process the smears, then the turn-around time will increase until the rate of receipt of new smears at the laboratory is reduced below the handling capacity. Note that in this model, women of different ages are screened at the same calendar time, and so are competing for resources in a realistic way. A further relevant feature of DES modelling is that it is possible to assess the transient effect of a change in policy as well as the new steady state to be reached. Full details of the cervical screening model are given in Chapter 9.

Other areas that prospective cost-effectiveness analysis should consider

In considering the costs and benefits of any change in the running or operation of a screening programme, it is inappropriate to consider sensitivity in isolation from specificity. Not only has the latter been shown to influence significantly the cost-effectiveness results, but sensitivity and specificity are interdependent. The implication of this is that basing the sensitivity on the best identified sensitivity figure and specificity on the best identified specificity, as has frequently been done, makes no sense given that they will be based on completely different populations.

On the basis of the review of economic analyses it is possible to identify a number of factors that appear to influence strongly the result of the analysis. These include: sensitivity and specificity estimates, thresholds used to calculate sensitivity and specificity, discount rates, unit costs and age at which a woman enters new screening arrangements.

In considering the use of the reviews and analyses detailed in this report, it is important to remember that the review did not uncover a single cost-effectiveness study with a focus on the combination of automation and other relevant technologies (i.e. LBC and/or HPV testing). It is likely that there are significant interactions on both costs and effects between the introduction of automation and the use of these other technologies; the costs and effects of the technologies are unlikely simply to be additive. Future economic analyses should consider this more complex picture, especially given the likely reality of the UK cervical screening programme operating with LBC.

Chapter 6

Systematic review of the effects and effectiveness of automation

Summary of key points

Methods

The objective was to evaluate the effects and overall effectiveness of the introduction of automated cervical screening devices. Information was sought and reported on four groups of outcomes:

- test performance (sensitivity and specificity)
- reproducibility (inter- and intra-rater reliability)
- health outcomes (invasive disease and deaths)
- process measures (rejection rates and slide processing times).

A comprehensive search in a previous systematic review by McCrory and colleagues[12] was updated from 1998 to 2000 and amplified to capture foreign-language articles and historically controlled (pre–post) studies.

Inclusion and exclusion criteria were defined, and were specific to each group of outcomes. Conclusions on impact on test performance targeted studies where the same set of slides were examined by an automation arm and a manual arm, and were submitted to a reference standard.

An important feature of the component of the review on test performance was to perform a main analysis with restrictive inclusion criteria and a secondary analysis where inclusion criteria with least impact on validity were relaxed. The resulting sensitivity analysis was used to examine the robustness of the conclusions from the main analysis and to extend the applicability of the results taking into account potential increases in openness to bias.

Results for test performance

Two studies (approximately 13,000 slides) were included in the main analysis and five additional studies were included in the sensitivity analysis (approximately 51,000 slides).

Variability in what was being assessed, threats to validity and actual variation in results all made interpretation difficult.

The conclusion most compatible with both the results of the main analysis and the sensitivity analysis is that automation offers no advantage over manual systems. However, potentially important improvements in sensitivity, specificity or both are not excluded; indeed, although less likely, nor was deterioration in test performance.

There are no evaluations of the currently commercially available automated device. This has features which mean that its test performance may be considerably different from its predecessor, which itself was only evaluated in one study included in the sensitivity analysis.

There are no evaluations comparing automation with systems incorporating other innovations such as LBC and HPV triage, which may become part of screening systems in the near future.

Implementation of automation cannot be recommended on the currently available evidence on test performance; rigorous and relevant research is an urgent priority.

Results for reproducibility

Six studies were included, all on PAPNET. They demonstrated less than perfect observer and machine variability, and sometimes marked variation, with weighted kappa scores less than 0.4.

The need for assessments on reproducibility of AutoPap was reinforced, and this should be an urgent priority for further research.

Results for health outcomes

Two studies, an RCT on PAPNET and a pre–post study on AutoPap 300 QC, were included.

Each provided information on the impact on numbers of invasive cancers. Given important limitations to their external and internal validity, little can be concluded directly from these studies.

Similar, properly conducted studies on current automated image analysis devices should be a component of future research.

Results for process

Several studies provided reasonably valid information on rejection rates.

Rates for AutoPap appear to between 7 and 8%; what evidence there is suggests that the rejection rates for newer versions of AutoPap are no lower than this, and rejection rates for LBC slides are only slightly lower than those for conventional Pap smears. Rejection rates for PAPNET seem to be less than for AutoPap, best estimates being rejection rates of between 2 and 3%.

Only one study provided clear, reasonably valid information on the impact of introducing automation on mean slide processing times. Using PAPNET to replace the primary screening step in a setting very similar to that in the UK, mean total slide processing time was two and a half to three times faster than the manual system. Although there are grounds to challenge the exact size of the effect, it is clear that the effect on processing times is substantial and statistically significant.

Independent confirmation of whether a similar effect may be expected from the currently commercially available version of AutoPap is probably the single greatest priority for further research on effectiveness of introducing automation. Other system changes should be anticipated from the introduction of AutoPap, including changes in staining protocol, increased vigilance in slide preparation, additional staff to run the automated devices, and an air-conditioned room for each device.

Overall conclusions from systematic review of effectiveness

There continues to be inadequate research evidence on which to base a decision on whether automation should be implemented. Even with the additional research identified in this review, the scope and volume of research are inadequate to address a problem of the complexity of introducing automated image analysis.

Existing research is barely generalisable to the current situation, particularly with respect to the nature of the intervention. No published research exists on the only currently commercially available automated device. No research compares automated devices to systems incorporating LBC and HPV triage. Existing research is insufficient for it to be safe to assume that a device offering technical advantages over its predecessors will inevitably deliver net benefit. Existing research highlights shortcomings which should be avoided in future research.

Debatably, the single most urgent priority for further effectiveness research is independent confirmation that the currently commercially available AutoPap device can achieve reductions in mean slide processing time similar to those achieved by PAPNET in the PRISMATIC trial.

Introduction

The essential feature of the evaluation of the clinical effectiveness was to build on and amplify past systematic reviews of effectiveness identified and appraised in Chapter 3. In particular, the reviews by McCrory and colleagues[12] and to a lesser extent Broadstock[62] were used as the starting point in developing a systematic review of effectiveness. The approach taken by McCrory and colleagues has been amplified and developed in five major respects:

- The searches were updated to the end of 2000.
- The searches were broadened to include foreign-language articles.
- The searches were broadened to capture studies other than those that measure sensitivity and specificity, especially historically controlled studies.
- The review specifically reported on reproducibility, impact on health outcomes and impact on process, as well as alterations in test performance.
- The review of test performance started with the strict inclusion criteria used by previous reviews and then relaxed those criteria not considered fundamental to the appraisal to ascertain whether an extended set of articles would have any impact on the conclusion.

The chapter has been divided into four parts so that the methodology and results of each of the following components may be described separately:

- Part 1: Test performance
- Part 2: Reproducibility
- Part 3: Health outcomes
- Part 4: Process.

General points and results applicable to all of these parts are first described below.

General method

Definition of published and unpublished studies

As the search for unpublished studies amounted to a separate exercise from the main review (see Chapter 7) it was important to define clearly what constituted a published and an unpublished article.

Published studies are defined as all articles that may be retrieved from the following sources:

- MEDLINE, EMBASE, the Cochrane Library, HEALTHSTAR, CINAHL, CancerLit and other major bibliographic databases
- any recognised specialist journal not indexed in the above databases; a recognised specialist journal may be defined as a journal that, by consensus agreement of specialists in a field, would be a journal that they would consult for further information in that field
- publications from any health technology assessment body or equivalent on a worldwide basis.

Those articles that may only be found in the recognised grey literature databases, such as SIGLE, are excluded.

Unpublished studies on automated cervical screening are all articles in the subject not found in the set of published studies.

From these definitions a document such as a conference proceedings would be considered an unpublished article. Furthermore, as there is no restriction placed on the completion of a study; an incomplete study would also be considered as an unpublished article.

Search strategy

There were three main objectives to the search strategy on clinical effectiveness:

- to update the search algorithm detailed in the report by McCrory and colleagues[12] for foreign-language articles (1 January 1990 to 31 December 1997, inclusive)
- to search for historically controlled trials (1 January 1990 to 31 December 1997)
- to produce a detailed broad search strategy with no restrictions on language or criteria on test performance (1 January 1998 to 31 December 2000).

The databases searched were:

- MEDLINE
- EMBASE
- Cochrane
- HealthSTAR
- CINAHL
- CancerLit.

The objectives require explanation. The work of McCrory and colleagues,[12] AHTAC, Canadian Co-ordinating Office HTA, and more recently the New Zealand HTA, has meant that the subject area of automated cervical screening systems has been trawled a number of times. From Chapter 3 it was concluded that the report by McCrory and colleagues would serve as the strongest foundation for this health technology assessment. The period of searching for their review was from 1966 to around March 1998. The results of their search and studies that met their inclusion criteria until December 1997 have been accepted, and it was considered that no value would be obtained from retrawling this period.

Previous reviews have not specifically searched for historically controlled trials, and so searching for the period 1990–1997 was re-covered, targeting such study designs; similarly, the same period was re-searched for non-English-language articles. Searching before 1990 was felt to be irrelevant as commercially available automated cervical screening systems did not exist before this date. The search for 1998–2000 was designed to capture all articles that would not have been captured by the McCrory search, all relevant historically controlled studies and all relevant non-English-language articles.

As there are essentially three devices of interest (AutoPap, PAPNET and AutoCyte SCREEN), the search strategy was based on text words and combinations of text words. Such an approach seemed likely to be highly sensitive, although lacking precision. The inclusion criteria were used to filter out the target studies.

The search strategy for the period 1998–2000 is provided in Appendix 6, which also details other search algorithms. The search algorithm used to identify foreign-language articles from 1990 to 1997 is as reported in the McCrory report,[12] with the omission of English-language restriction.

Inclusion/exclusion criteria

Inclusion criteria were devised for each of the four components of the clinical effectiveness review. Again, the theme of the approach was to use, and where necessary to adapt, those criteria applied to

TABLE 16 *Yield of searches undertaken to update and amplify the searches undertaken for the report by McCrory and colleagues*[12]

Data base	Update search 1998–2000	Historical trials 1990–1997	Foreign language 1990–1997
MEDLINE	120	65	32
EMBASE	104	50	5
COCHRANE	11	5	1
HealthSTAR (non-MEDLINE)	3	3	0
CINAHL	11	2	0
CancerLit (non-MEDLINE)	0	2	0
Total	249	127	38
Unduplicated articles	163	66	38

studies retrieved by McCrory and colleagues.[12] All articles were screened by one reviewer (BW). Those that passed step 1 were independently screened by a second reviewer (CH). Inclusion decisions were compared and any disagreements resolved by consensus.

Data extraction

A data extraction form, detailing important elements of the study, was completed for all papers that passed step 1. For the historic studies, data extraction forms were completed on included articles only. The data extraction forms were completed by one author (BW).

Yield of searches

The results from the search algorithms applied to each database are shown in *Table 16*. These results represent the number of articles retrieved from each database before removal of the duplicate articles. The total number of unduplicated articles that could potentially be included in the effectiveness review is given in the bottom row.

Part 1: Test performance

Introduction

The purpose of this component of the systematic review was to assess whether the ability of automated cervical screening image analysis to identify correctly preinvasive cervical lesions was improved relative to alternatives, particularly those similar to the current UK manual system. Eligible studies were identified from an existing systematic review covering the period up to the end of 1997, potentially relevant studies published in non-English languages for the period 1990–1997 and *de novo* searches covering the period 1998–2000. Concerning the first of these, no studies on PAPNET, AutoPap or other automated image analysis devices passed all of the inclusion criteria set in the report by McCrory and colleagues.[12] However, evidence tables were produced on those studies that came closest to being included, and studies in these were reconsidered with respect to inclusion or exclusion in this review. These amounted to six articles on AutoPap and 11 on PAPNET. Some of these articles were published in early 1998 and were located as part of the search covering 1998–2000 (two studies on AutoPap and three on PAPNET). Thus, just 12 studies considered for inclusion in the review were uniquely identified via the report by McCrory and co-workers.[12]

Inclusion/exclusion criteria

These were adapted directly from the report by McCrory and colleagues[12] and applied as a two-stage process.

(1) Step 1
   1. The study type was a primary study.
   2. The study evaluated cervical cytology as a screening test.
   3. The study evaluated an automated or a semi-automated cervical screening system.
   4. The screening system was used in a primary screening, rescreening only or primary screening with rescreening mode.

(2) Step 2
   1. The study used a two-armed design (see explanation below).
   2. The study used a reference standard for diagnosis.
   3. The reference standard used was either
      (a) histology with colposcopy or
      (b) an independent panel of two or more cytologists adjudicating differences between the conventional and the new technology.
   4. The majority of those testing positive for HSIL were verified with histology or colposcopy (studies that did not verify at least half of those with screening tests positive for HSIL were excluded).

5. The study allowed for separate analyses of sensitivity (or relative TPR) and specificity (or relative FPR).

The expression 'two-armed design' encompasses a number of points. The first is that the control test and the new test are applied to the same set of slides. For the purpose of ascertaining the new test performance, the control test should be one of the conventional Pap test, LBC or either of these with HPV triage. The second is that both arms are executed under the same test conditions. For instance, the original diagnosis obtained in routine screening being used as a control arm, compared with a new test performed subsequently, albeit on the same slides, would not be considered a two-armed design. Third, both the control test arm and the new test need to be compared against the reference standard. Finally, it was not considered necessary to insist on prospective trials; retrospective studies were considered equally valid, provided that the foregoing points were met.

It is necessary to note that the algorithm is designed hierarchically, which is particularly relevant to step 2.

All studies passing all the criteria in steps 1 and 2 were included. However, to minimise the chance that important research information was overlooked, and to acknowledge that some of the inclusion/exclusion criteria did not seem to differentiate clearly valid from invalid research, a sensitivity analysis was conducted in which a wider set of studies was considered, including those that just failed either step 2.4 or 2.5 alone.

Quality assessment

A quality assessment checklist was developed after consulting methodological papers on assessing diagnostic/screening tests[50,83,104] and appraising the approaches used in previous reviews on this topic (see Chapter 3). The following list was primarily based on Reid and colleagues' criteria,[83] amplified with others that seemed particularly pertinent to the topic in question.

1. **Recruitment of study subjects**: the preferred designs are a consecutive within-subject design where the outcome is the diagnosis of preinvasive disease, or a randomised between-subjects design where the outcome is the prevention of invasive disease. For the purpose of the report, where a case-mix had been created retrospectively this was not considered adequate.
2. **Spectrum composition**: the spectrum of the tested patients should be specified, in terms of age, gender, symptoms, level of disease and eligibility criteria of patients. Where the cytology for the whole study sample could be derived, this was accepted.
3. **Reference standard**: the study used a reference standard.
4. **Avoidance of verification bias**: in cohort studies all subjects should receive the diagnostic test and gold-standard verification, and where this was not possible should have suitable clinical follow-up as a surrogate gold (reference) standard.
5. **Avoidance of review bias**: the screening test and gold standard should be performed independent (blind) of the other's results, whichever the order.
6. **Precision of results for test accuracy**: all performance metrics (sensitivity, specificity, likelihood ratios) should have standard errors or confidence intervals calculated. This was complied with when sufficient data were disclosed in the article to calculate these values across all thresholds.
7. **Presentation of indeterminate test results**: the study should report all results, including those in which the test was inconclusive, and whether such indeterminate results were included or excluded from analyses of test accuracy. This was interpreted to mean disclosure of slides that could not be processed.
8. **Test reproducibility**: where the test requires observers or machine intervention, summary measures of observer and instrument variability should be given.
9. **Industry sponsorship**: studies should be independently funded without manufacturer involvement. Where any of the authors were connected by payroll, or in receipt of a grant from the manufacturer under investigation, then this was not accepted.

It should be clearly noted that this list is only a framework for consistently assessing and summarising the main strengths and weaknesses of the included papers. It is not intended to provide any basis for ranking or weighting the results of one included study above another. The main factor precluding this is that there is debate on which criteria are more likely to lead to bias than others, and indeed whether the implications of failings in any particular criterion are the same irrespective of the diagnostic or screening test being assessed. A particular case in point, as far as cervical screening is concerned, is that a study population compiled retrospectively (see criterion 1) may be less of a threat to validity than in other tests, because the framework from which slides are

TABLE 17 *Number of studies included and excluded in the review of new test performance*

	Update search 1998–2000	Foreign language 1990–1997	Studies considered by McCrory *et al.*[12]
Unduplicated articles	163	38	12
Step1			
Excluded	115	38	0
Remaining	43	0	12
Step 2			
Excluded	43		12
Remaining	2		0
Included	2	0	0

sampled is well defined in systems where there is population-based screening. Further, there may be associated advantages with respect to the quality of the reference standard, which may be practically difficult to achieve with a prospectively compiled test population.

Analysis

The main method of analysis was qualitative, relying on clear tabulation of study characteristics, quality and results, basing conclusions on the patterns revealed. Quantitative techniques such as meta-analysis were actively considered, but were felt to be inappropriate for the study designs encountered.

Unless otherwise stated, the results are those recalculated by the authors of this report from the raw data provided in the article, particularly from two-by-two tables. Results reported in the article are also quoted where appropriate. Confidence intervals for the sensitivity and specificity have been calculated using Wilson's technique.[105] Where there was uncertainty in the true diagnosis of the whole study sample, from a lack of verification of test negatives, relative TPR and relative FPR have been calculated.[96]

Results of main analysis: studies included and excluded

The results of the inclusion and exclusion decisions are given in *Table 17*, subdivided by the source from which the study was identified and whether the study was excluded at step 1 or 2.

Only two studies met all of the inclusion criteria.[106,107] Both were identified from the update search covering the period 1998–2000. No included foreign-language papers were identified and no studies previously excluded by the review by McCrory *et al.* were included in this review.

Brief details of inclusion and exclusion decisions on individual studies may be found in Appendix 7. In total, 153 potentially eligible studies were excluded for step 1 criteria concerning the general nature and relevance of the articles. *Tables 18* and *19* show the number of studies that failed on each criterion of step 2, subdivided by technology.

Over half of the articles that passed step 1 failed step 2.1, that is, the criterion necessitating a two-armed design. There was an even mix of reasons for this. Some of the studies that had two arms to their design did not apply both arms equally to the same set of smears. Others defined the control arm as the reference standard, or the original diagnosis.

Table 19 illustrates that there were some differences between the present interpretation of a two-armed design and that used by McCrory and colleagues. In the McCrory health technology assessment[12] most of the 12 studies in question were excluded for reasons of inadequate reference standards, that is, criterion 3. The present group did not consider using the reference standard or the original diagnosis from routine screening as a control arm, as applying the bias of hypervigilance equally to both arms. The new technology would unilaterally benefit from the test conditions.

Consistent with the findings of previous reviewers, several papers were excluded owing to the reference standard. In particular, some otherwise strong studies were excluded because of a failure to biopsy more than 50% of the high-grade lesions in the study population. In these cases, the reference standard was an independent panel review of discrepant cytological samples between the two arms of the study, the final diagnosis being the consensus view of the panel.

TABLE 18 *Studies excluded by step 2 criteria from update search 1998–2000*

	Number of articles excluded at step 2.*X*						
Device	***X* = 1**	***X* = 2**	***X* = 3**	***X* = 4**	***X* = 5**	**Included**	**Total**
AutoPap	6	0	1	2	0	0	9
PAPNET	19	1	1	1	4	2	28
AutoCyte SCREEN	0	2	4	0	0	0	6
Others	2	0	0	0	0	0	2
Totals	27	3	6	3	4	2	45

TABLE 19 *Studies excluded by step 2 criteria from 'near-miss' studies identified by McCrory and colleagues*[12]

	Number of articles excluded at step 2.*X*						
Device	***X* = 1**	***X* = 2**	***X* = 3**	***X* = 4**	***X* = 5**	**Included**	**Total**
AutoPap	4	0	0	0	0	0	4
PAPNET	3	1	2	1	1	0	8
AutoCyte SCREEN	0	0	0	0	0	0	0
Others	0	0	0	0	0	0	0
Totals	7	1	2	1	1	0	12

TABLE 20 *Scope of included studies (main analysis)*

Study	Current system					Change whose impact was evaluated					
	Nature of system				Nature of QA	Technology			Mode		
	Pap	Pap + HPV	LBC	LBC + HPV		PAP-NET	Auto-Pap	Auto Cyte SCREEN	Replaces existing primary step	Replaces existing QA step	Adds new QA step
Doornewaard *et al.*, 1999[106]	✓				None mentioned[a]	✓			✓		
Sherman *et al.*, 1998[107]	✓				None mentioned	✓			✓		

[a] Dutch system does not normally incorporate a quality assurance (QA) step.

Two excluded studies are worthy of further mention. The first, an article written in Danish by Hølund and colleagues, was excluded for not having a two-armed design.[108] However, the tables, written in English, seemed to contradict the translated text of the article, and suggest that the study may actually have satisfied the criterion. The study could also have been excluded on the nature of the reference standard because it was not explicitly described. Given the nature of the reasons for exclusion, it is possible that this study could be included should suitable clarification emerge, but none has so far been forthcoming.

The second study, by Kok and colleagues,[109] was an RCT, a type of study that was not anticipated at the outset of this review and so excluded from the assessment of test performance on the basis that it was not a two-armed design. Given its nature and potential, it was, however, appraised and assessed in full in the section on impact on health outcomes.

Results of main analysis: details, quality and results of included studies

The characteristics, quality and results of the two included studies[106,107] are summarised below.

What was investigated?

Both included studies evaluated PAPNET. *Table 20* emphasises how the included studies deal with only a small subset of the possible strategies

TABLE 21 *Details of intervention and comparator in included studies (main analysis)*

Study	Intervention description	Base comparator description	Blinding	Observer bias
Doornewaard *et al.*, 1999[106]	PAPNET scans all slides, and images reviewed on monitor. Abnormalities identified from images, given targeted microscopic review. All positive slides referred to a pathologist. Negative slides archived	Primary microscopic screening of all slides. Abnormal slides referred to same pathologist as in intervention, with negative slides archived. Screening done by same four screeners in both arms	Yes	No
Sherman *et al.*, 1998[107]	All slides scanned by PAPNET in New York, and images reviewed on monitor by a US senior screener. Abnormal cases received full microscopic review by screener. If still abnormal referred to pathologist. Negative slides archived	All slides subject to primary microscopic screening in Costa Rica and assessed by a Costa Rican pathologist	Yes	Yes

ideally needing to be considered in this project. Thus, there were no included studies on the other technologies, especially the AutoPap, and the conventional Pap test was used as the comparator in both cases. Referring back to Chapter 5, *Figures 5* and *6*, there is no information on the impact of automated image analysis compared with the Pap test + HPV and LBC ± HPV. The screening modality appraised in each case was for primary screening.

Details of the intervention and comparator

Despite both studies investigating PAPNET in a primary screening mode, it can be seen from *Table 21* that there are important differences between the systems adopted in each study. In the first study the location of abnormalities was followed by a targeted microscopic review of the slide based on the coordinates of the abnormalities identified on the PAPNET monitor. In the second study the whole slide received microscopic review following detection of abnormalities without recourse to any location information available. Such differences in the deployment of the new technologies make comparisons between studies difficult. This is the most obvious difference highlighted, but others undoubtedly exist. For instance, the first study was conducted in The Netherlands, whereas the second was conducted in Costa Rica. There could therefore be marked ethnic and relative risk differences between these two populations.

The Dutch investigators aimed to reduce the chances of interobserver bias by using the same four screeners to rescreen the case series. There is a risk of compromising blinding between the two arms in this instance, although given the number of slides, individual detail of slides is less likely to be remembered. In the second study there was potential for interobserver bias, both by the deployment of different screeners for the different arms (one of whom was an expert pathologist) and by the different interpretations of the Bethesda system between the two countries.

Study population

The performance of a diagnostic test is intrinsically linked to the spectrum of disease in the study population, and so it is important that test performance be evaluated in a population similar to that in which the test will be eventually deployed.

The design and disease spectra for both studies are tabulated in *Table 22*. As a study design, a prospective cohort is more likely to be representative of the screening population than a case-mix chosen retrospectively. In this sense the study by Sherman and colleagues[107] is closer to the ideal, illustrated by the greater similarity to the profile of slide categories actually obtained in the NHSCSP.

However, the investigation by Doornewaard and colleagues,[106] despite being retrospective, also has positive design features. There were three sources to the case-mix. With a base year of 1988, all abnormal cytology was chosen for that year, along with any negative slides that had subsequently been shown to have histological abnormalities. The final source was a random selection from all negative cytology over the years 1988–1995. This retrospective selection of the cases has the

TABLE 22 *Study design and disease spectrum in included studies (main analysis)*

Study	Design and location	No. of slides	Spectrum (%)					
			Negative	Equivocal	CIN1	CIN2	CIN3	Invasive
Doornewaard *et al.*, 1999[106]	Retrospective, case-mix, Utrecht, The Netherlands	6063	86	0	8.6		5.4	
Sherman *et al.*, 1998[107]	Prospective, cohort, Guanacaste, Costa Rica	7323	89.6	6.7	2.1	1.4		0.2
NHSCSP: % of slides in categories, negative, inadequate or borderline, mild, moderate, severe dyskaryosis		4.25 million	82	14	2.4	0.9		0.7

TABLE 23 *Reference standard used in included studies (main analysis)*

Study	Reference standard	Reviewer blinding
Doornewaard *et al.*, 1999[106]	PALGA database True-negative smears had 7 years of follow-up Abnormal cases nearly all had biopsy in 7 years of follow-up False negatives were smears initially diagnosed as normal in 1988 and shown to have abnormality in the following 7 years	NA
Sherman *et al.*, 1998[107]	Based on all pathology material available (smear, colposcopy report, biopsy and cervigrams) reviewed in USA by pathologist Colposcopic referral after positive pelvic examination, cytology or cervigram Biopsy on 93% HSIL, 100% cancer and 39% LSIL	Unlikely

NA, not applicable.

advantage of providing a strong reference standard (see below), while maintaining the true relative proportions of abnormalities, if not the negatives; 50,000 smears were processed in 1988, but only 5000 negatives were included in the study sample.

Reference standard

The reference standards used in both included studies are described in *Table 23*. A repeated problem faced by investigators of automated cervical screening systems in establishing the true level of disease has been the difficulty in verifying negative cytology, owing to ethical or logistic reasons. The study by Doornewaard and colleagues is one of the few in which negative smears were followed up over a reasonable period (7 years) to ensure that the level of true negatives could be accurately gauged. In the Dutch study the level of disease in the study population was also ascertained by 7-year follow-up, by histology in nearly all cases. The level of abnormalities was derived from two groups: those with a positive smear in the base year with confirmation by biopsy, and those with a negative smear, but in which either subsequent biopsy or cytology had been shown to be abnormal. Confirmation by later histology of the positive smears removes the false positives from this group. Owing to the long follow-up it is possible that some of the smears demonstrated genuine abnormalities, but as part of the natural course of the disease had regressed. Such abnormalities would have been reported incorrectly if a cytology reference standard had been applied.

Negative slides, which on later histology showed the presence of disease, could be a manifestation of either screening or, more probably, sampling error. They could also be due to *de novo* lesions, which is the main failing of using this reference standard. However, over a 7-year period *de novo* lesions are more likely to be of a low-grade than high-grade classification, giving a realistic estimate of the number of high-grade lesions missed by the original screening programme.

In contrast, the Costa Rican study used all the pathology material available to the reviewing

TABLE 24 *Study quality in included studies (main analysis)*

	1	2	3	4	5	6	7	8	9	Weaknesses identified	Comment
Doornewaard *et al.*, 1999[106]	×	✓	✓	✓	✓	×	✓	✓	✓	Retrospective recruitment No confidence intervals	Neither weakness felt to constitute a major threat to validity
Sherman *et al.*, 1998[107]	✓	✓	✓	×	×	✓	✓	×	×	Possible verification bias Possible review bias No information on test reproducibility Industry sponsored	Possibility of review bias felt to constitute an important threat to validity

1, Study subjects recruitment; 2, spectrum composition; 3, reference standard; 4, avoidance of verification bias; 5, avoidance of review bias; 6, precision of results for test accuracy; 7, presentation of indeterminate test results; 8, test reproducibility; 9, industry sponsorship.

pathologist to diagnose the study population. Since the sole objective of applying a reference standard is to obtain the correct classification of the study population, the more evidence on each case the better. Thus, the information recorded from the smear, colposcopy report, cervigrams and the biopsy all provide evidence on the cases. Unfortunately, there are two shortcomings to this study. The first is that all negatives were not verified and, therefore, the values of the true and false negatives cannot be estimated. The second, which also applies to the Dutch study, is that the reference diagnostic testing in both cases was unlikely to be applied independently of the test results. This means that verification bias cannot be eliminated. A particular concern about the study of Sherman and colleagues is the possibility that the pathologist overseeing the PAPNET arm appears to have had overall say in the reference diagnosis. This would compromise his independence and increase the risk of diagnostic review bias.

Quality of included studies

These are summarised in *Table 24*.

Most of the items have been introduced in the preceding sections. The following points are particularly worth highlighting. The study by Doornewaard and colleagues has two main failings according to these criteria. The first is that the study population is not a natural cohort that would be encountered in the normal screening setting, and the second is that insufficient data were given to be able to calculate the sensitivity and specificity across all thresholds. However, these were not felt to constitute major threats to validity.

In contrast, the shortcomings of the Sherman study were of some concern, particularly that not all the test subjects had the reference standard applied and that there was a potential for reviewer bias.

Test performance results in included studies

Table 25 summarises the performance characteristics of the PAPNET from the data provided by the two included studies over a range of test and disease thresholds. The main threshold for disease was taken to be HSIL+ or moderate+ (where based on cytology) or CIN3+ (where based on histology), as diagnosed by the reference standard, since these are the preinvasive states most likely to progress to invasive cancer. However, to ensure that possible advantages of the new technology over conventional screening at the low end of the spectrum are not missed, the lower row defines disease to be an abnormality of severity equal to or greater than ASCUS or borderline using a cytological reference standard. Cells marked with an asterisk represent where the difference between the metrics was statistically significant with $p < 0.05$, as judged by the 95% CI not overlapping.

Both included studies assessed PAPNET in primary screening mode with the conventional Pap test as a comparator.

In terms of sensitivity, only one comparative result showed a statistically significant difference. With threshold of disease set at HSIL+, and a test threshold set at a similar level, Sherman and colleagues reported the sensitivity of conventional screening as over 18% better than the PAPNET. In contrast, at the same thresholds the Dutch study demonstrated a slight improvement of PAPNET over conventional screening, but this was not statistically significant. There were sufficient data

to make six comparisons of sensitivity between the automated screening and the conventional screening (four from Sherman and colleagues, two from Doornewaard and colleagues), and in only one case was the difference significant. Although far from conclusive, what evidence there is in the main analysis is slightly more in favour of use of PAPNET in a primary screening mode being associated with a reduction in sensitivity rather than an improvement.

With respect to specificity, statistically significant gains in favour of PAPNET were found in the Costa Rican study at all disease and test thresholds analysed. The increases ranged from +0.7% to +2.7%. This compares with the Dutch study, where the difference was slightly in favour of conventional screening, but was not statistically significant in either instance. Again, although far from conclusive, the balance of evidence available in the main analysis slightly favours use of PAPNET in primary screening being associated with an improvement in specificity.

The pattern of improved specificity at the expense of sensitivity is surprising given that it has generally considered that automated technology would be most likely to improve on the sensitivity of conventional screening, but possibly at the expense of the specificity.

The absolute levels of sensitivity in the included studies are also noteworthy, although it should be noted that with the study designs in question, the absolute levels of sensitivity (and specificity) are much less trustworthy than the comparisons between them. In the Dutch study the sensitivities at the high-grade threshold were much lower than the Costa Rican study. The authors cite diminishing quality of the dyes and increasing artefacts over the period in which they were originally taken (1988) and when they were rescreened (1995). They also mention the possibility of *de novo* lesions, which may have occurred over the protracted follow-up. Both points are probably true, but it must be borne in mind that the sensitivities reported for the study by Doornewaard and colleagues are in line with the results of recent published meta-analyses assessing the sensitivity of conventional Pap testing as being much lower than traditionally accepted.[21,65] It is true there are likely to be *de novo* lesions, as discussed earlier, but equally, there are likely to be lesions that the Pap test will always miss, which were identified by the reference standard. Neither type of lesion will be reflected by the test sensitivity if verification is by cytology and, to some extent, histology in short-term follow-up (as this test is not perfect either). Therefore, the sensitivity measured here is more likely to reflect programme sensitivity than test sensitivity (the more often quoted parameter).

The included studies provide no evidence to make assertions about the performance of automated cervical screening in a rescreening mode, or in a combined primary and rescreening mode.

Provisos concerning interpretation of pattern of results across included studies

There is an important caveat to the analysis of the results of the included studies as they relate to PAPNET used in primary screening. There is no clear pattern and there is variation in the results in these two studies in both the sensitivity and the specificity. This could be due to the expected statistical random variation, but it could also be an effect in the quality of the studies, as there were noted shortcomings in the design and there were important clinical differences, particularly with respect to how the automated system was deployed and the nature of the population. Thus, statistical, methodological and clinical heterogeneity ideally needs to be taken into account in interpreting these results, but the very limited number of studies precludes further meaningful investigation. Inevitably, the certainty with which any overall conclusions can be made is in consequence greatly limited.

Conclusions of main analysis

No conclusions can be drawn on the impact of automation used in a quality assurance mode, and no conclusions can be drawn on what the impact of introducing automation would be if the current manual system incorporated LBC or HPV triage.

The included studies provide some information on the performance of an automated device (PAPNET) compared with the conventional Pap test in a primary screening setting. The data are limited and need to be interpreted with extreme caution. With this proviso the balance of evidence slightly favours PAPNET in primary screening mode leading to a small improvement in specificity at the expense of loss of sensitivity. Any changes cannot be precisely quantified from the data available.

Sensitivity analysis: method

Because of concerns that overly strict inclusion criteria may have given a false impression of the limited evidence base underpinning the use of automation, and a concern that failure to meet

TABLE 25 *Summary of the test performance over different thresholds of disease and thresholds for test positive (main analysis)*

Study	Technology/ comparator (disease threshold)	*n*	Sensitivity: technology (95% CI)	Sensitivity: comparator (95% CI)	Sensitivity gain (technology – comparator)	Specificity: technology (95% CI)	Specificity: comparator (95% CI)	Specificity gain (technology – comparator)
Threshold for: test = HSIL+ or moderate+; disease = HSIL+ or moderate+ or CIN3+								
Doornewaard *et al.*, 1999[106]	PAPNET primary screening/manual (CIN3+/7-year follow-up)	6063 (325:5738)	19.7% (15.7 to 24.4) (HSIL+)	14.5% (11.1 to 18.7)	+5.2% ns	99.5% (99.3 to 99.7) (HSIL+)	99.6% (99.4 to 99.7)	–0.1% ns
Sherman *et al.*, 1998[107]	PAPNET primary screening/manual (HSIL+)	7323 (114:7209)	49.1% (40.1, 58.2) (HSIL+)	67.5% (58.5 to 75.4)	–18.4%*	99.9% (99.7 to 99.9) (HSIL+)	99.2% (98.9 to 99.4)	+0.7%*
Threshold for: test = LSIL+ or mild+; disease = HSIL+ or moderate+ or CIN3+								
Doornewaard *et al.*, 1999[106]	PAPNET primary screening/manual (CIN3+/7-year follow-up)	6063 (325:5738)	61.5% (56.1 to 66.7) (LSIL+)	63.4% (58.0 to 68.4)	–1.9% ns	91.8% (91.1 to 92.5) (LSIL+)	92.0% (91.2 to 92.7)	–0.2% ns
Sherman *et al.*, 1998[107]	PAPNET primary screening/manual (HSIL+)	7323 (114:7209)	60.5% (51.4 to 69.0) (LSIL+)	76.3% (67.7 to 83.2)	–15.8% ns	99.0% (98.7 to 99.2) (LSIL+)	96.3% (95.8 to 96.7)	+2.7%*
Threshold for: test = ASCUS+ or borderline+; disease = HSIL+ or moderate+ or CIN3+								
Doornewaard *et al.*, 1999[106]	PAPNET primary screening/manual (CIN3+/7-year follow-up)	6063 (325:5738)	Unable to calculate sensitivity and specificity at this range of test threshold levels					
Sherman *et al.*, 1998[107]	PAPNET primary screening/manual (HSIL+)	7323 (114:7209)	86.0% (78.4 to 91.2) (ASCUS+)	79.8% (71.5 to 86.2)	+6.2% ns	97.0% (96.6 to 97.4) (ASCUS+)	94.6% (94.1 to 95.1)	+2.4%*
Threshold for: test =ASCUS+ or borderline+; disease =ASCUS+ or borderline+								
Doornewaard *et al.*, 1999[106]	PAPNET primary screening/manual (ASCUS+)	6063	Unable to calculate sensitivity and specificity at this range of test threshold levels					
Sherman *et al.*, 1998[107]	PAPNET primary screening/manual (ASCUS+)	7323 (269:7054)	66.5% (60.7 to 71.9) (ASCUS+)	69.5% (63.8 to 74.7)	–3.0% ns	98.1 % (97.8 to 98.4) (ASCUS+)	95.9% (95.4 to 96.3)	+2.2%*

* $p < 0.05$; ns, not significant.

TABLE 26 *Sensitivity analysis: scope of included studies*

Study	Current system					Change whose impact was evaluated					
	Nature of system				Nature of QA	Technology			Mode		
	Pap	Pap + HPV	LBC	LBC + HPV		PAP-NET	Auto-Pap	Auto Cyte SCREEN	Replaces existing primary step	Replaces existing QA step	Adds new QA step
Doornewaard *et al.*, 1999[106]	✓				None mentioned[a]	✓			✓		
Sherman *et al.*, 1998[107]	✓				None mentioned	✓			✓		
Duggan and Brasher, 1997[110]	✓				None mentioned	✓			✓		
Lerma *et al.*, 1998[111]	✓				Full manual[b] rescreening if ASCUS on primary screen	✓				✓[b]	
PRISMATIC, 1999[112]	✓				Rapid review of all negatives and inadequates[c]	✓			✓		
Veneti *et al.*, 1999[113]	✓				Full manual rescreening if primarily negative on primary screen and then developed precancer or cancer[b]	✓				✓[b]	
Wilbur *et al.*, 1998, 1999[114,115]	✓				10% random rescreen of "within normal limits of slides"		✓[d]		✓[e]	✓[f]	

[a] Dutch system does not normally incorporate a QA step.
[b] Debatable whether study truly represents the impact of automation introduced to replace existing QA in a whole system, as study was confined to a subsection of slides likely to be encountered.
[c] Detailed results only available for the primary screening step, although the study did consider the test performance of the whole system too.
[d] AutoPap Primary Screening System.
[e] Amplifies rather than replaces the primary screening step. AutoPap first analyses slide and designates 'no review', 'review' or 'process review, or rerun'. Primary manual screen only undertaken if designated for 'review' or 'process review, or rerun'. Further, primary manual screen of these slides was undertaken with the benefit of 'ranking' by AutoPap.
[f] Amplifies rather than replaces QA step. Slides on which QA rescreening is undertaken were selected on the basis of the 15% highest ranking slides in the AutoPap 'review' category as opposed to a 10% random sample of 'within normal limit' slides.

some of the criteria may not greatly compromise validity of the results, the review of evidence on test performance incorporated a sensitivity analysis of the primary conclusions. The method was exactly as for the primary analysis, except that two criteria felt to have least bearing on validity were relaxed and conclusions on test performance redrawn on the resulting wider pool of studies included. The two criteria step 2.5 (the requirement of sensitivity/specificity or relative TPR/FPR) and step 2.4 (the criterion of biopsy confirmation on a simple majority of high-grade lesions) were relaxed. This resulted in six further papers, reporting five studies, being considered.[110–115] All the tables produced in the last section have been reproduced with the

additional studies. The two original studies included appear at the top of all the tables.

Results of sensitivity analysis: details, quality and results of included studies

What was investigated?

Of the six further studies, four were on PAPNET[110–113] and one was on AutoPap, this being reported in two separate articles.[114,115] As in the main analysis, the comparator in all cases was the conventional Pap test. Thus, even after the sensitivity analysis no direct information was identified on what the impact of automated screening might be if introduced to a system incorporating LBC and/or HPV triage. *Table 26* illustrates how the additional studies provide evidence on some of the strategies not addressed in the main analysis. However, there remains an absence of evidence on an equally large number of strategies.

Other types of screening modality were assessed with the additional studies. Two studies[111,113] apparently assessed the impact of automation replacing an existing quality assurance step. However, this classification is open to debate. As will be seen, both of these studies involved the screening of a case set with a high prevalence of disease, which does not reflect the type of population faced by a screening programme either in a primary screening mode or in a quality control mode. The three other additional studies provided information on automation replacing manual primary screening[110,112] or automation focusing on the targeting of the primary manual screen and altering the selection of the slides examined for quality assurance purposes.[114,115]

Details of the intervention and comparator

In the corresponding section of the main analysis it was noted that making comparisons between studies was made difficult owing to there being different methods of deploying the same technology in the same screening mode. The types of screening mode are used to classify the different studies. The details are described in *Table 27*.

Primary screening mode

There were two additional studies, four in total, which evaluated automated screening systems in a purely primary screening mode. Both assessed PAPNET.

One additional study, by Duggan and Brasher,[110] used a protocol that resembled that of Doornewaard and colleagues, using a targeted screening of the slides following identification of abnormalities on the monitor. Unlike the earlier Dutch study, there appears to be potential for interobserver bias between the two arms.

The other additional study was the PRISMATIC trial.[112] This was the only identified study in any category directly relevant to the UK NHSCSP. It was based in south-east England and was designed to evaluate PAPNET as a primary screening system against the conventional screening system, where quality control in both arms was provided by the rapid rescreening of negative slides. NHSCSP protocols were used, and as such it is likely to be the truest reflection of the impact of introducing one type of automation to the national programme in the UK, provided that the results are internally valid. Unfortunately, there is a possible weakness in the PRISMATIC trial. Blinding between the trial arms is so important that a lack of it would reduce the validity of the study results, and its reporting in this study is unclear. It is quite probable that blinding did take place, but was overlooked or edited out in the published article. A quote from the study's abstract supports a view that blinding did occur, stating that, "The study complied with international standards for assessment of automated cervical screening systems". Unfortunately, these standards were not made explicit, and in all other studies in the sensitivity analysis blinding was clearly reported.

Finally, it should be noted that a third additional study, reported in two articles,[114,115] provides some information on the impact of introducing an older version of AutoPap into the primary screening step, but that automation was also used simultaneously to amplify the quality assurance step. It is thus described in a separate subsection, 'Combined primary screening and quality control' (p. 74).

Rescreening mode

The two studies identified in the sensitivity analysis[111,113] add little to the evaluation of automated screening systems used in quality control. Both studies take the final diagnosis of the PAPNET to be that obtained after reviewing the monitor only. Thus, any abnormalities or suspicious lesions identified after reviewing the monitor were not subject to microscopic review. Only one of the studies was used in the rescreening of negatives,[113] albeit false negatives, which would be unrealistic in a normal screening setting. It is thus highly debatable whether either study gives a true impression of what the impact would be of introducing automation as an alternative or additional quality assurance step.

TABLE 27 *Sensitivity analysis: details of intervention and comparator in included studies*

Study	Intervention description	Base comparator description	Blinding	Observer bias
Doornewaard *et al.*, 1999[106]	PAPNET scans all slides, and images reviewed on monitor. Abnormalities identified from images, given targeted microscopic review. All positive slides referred to a pathologist. Negative slides archived	Primary microscopic screening of all slides. Abnormal slides referred to same pathologist as in intervention, with negative slides archived. Screening done by same four screeners in both arms	Yes	No
Sherman *et al.*, 1998[107]	PAPNET scans all slides in New York, and images reviewed on monitor by a US senior screener. Abnormal cases received full microscopic review by screener. If still abnormal referred to pathologist. Negative slides archived	Primary microscopic screening of all slides in Costa Rica and diagnosed by a Costa Rican pathologist	Yes	Yes
Duggan and Brasher, 1997[110]	PAPNET scans all slides, and images reviewed on monitor by one of two experienced (>10 years) screeners. Abnormal cases received a targeted microscopic review by screener. If still abnormal referred to one of four pathologists. All negative slides archived	Primary microscopic screen of all slides by one of three screeners. All abnormal slides referred to one of four pathologists. Pathologists blind to cytology results. All negative slides archived	Yes	No
Lerma *et al.*, 1998[111]	PAPNET scanning of case set, and images reviewed on a monitor. Diagnosis appears to be based on the images and no further microscopic review	Microscopic rescreening of case set. No protocol for referral given	Yes	Who the screeners were or their relative experience not made explicit
PRISMATIC, 1999[112]	PAPNET scans all slides, and images reviewed on monitor. Abnormal cases received full microscopic review by screener. If still abnormal referred to pathologist. All negative slides only underwent rapid review	Primary microscopic screen of all slides by ? screeners. All abnormal slides referred to pathologist. All negative and inadequate slides underwent rapid review	Presumed	Not clear. Trained two or three screeners from each centre for PAPNET. Number of screeners per centre in the control arm not known
Veneti *et al.*, 1999[113]	PAPNET scanning of case set, and images reviewed on a monitor by two cytologists. Diagnosis based on the consensus opinion of the two cytologists after review of the images. No further microscopic review	Microscopic rescreening of case set by one screener. No protocol for referral given	Yes	No
Wilbur *et al.*, 1998, 1999[114,115]	AutoPap scanning of all slides. Slides classified for review, reviewed by screener. Those review slides passed as normal, but classified as QC review by AutoPap received further rescreening (15%). Final diagnosis based on this. No pathologist referral	Primary microscopic screen of all slides by screeners. Random selection of 10% of negative slides receive full rescreening by ? screener. Final diagnosis was that arrived at by the screeners, no pathologist referral	Yes	Probable, as implied by "at no time did the same c/t review a slide from both arms"

As above, one study did give information on the potential impact of an older version of AutoPap on the quality assurance step most often used in the USA. However, this was combined with the amplification of the primary screening step, and is discussed in detail in the next section.

Combined primary screening and quality control
One study, reported as two articles,[114,115] evaluated the impact of introducing automation simultaneously to both the primary screening stage and the quality assurance step. It should be noted, however, that the quality assurance step used in the control arm and modified in the automation arm is that typically used in the USA, namely a full manual rescreen of a 10% random selection of slides designated as 'within normal limits' in the primary screen. Such practice is not used in the UK cervical screening programme, reducing the potential applicability of any results to the NHSCSP. However, the study is notable in that it is the only study considered that provides information on the potential impact of AutoPap, albeit an early predecessor to the version currently being marketed.

Comparison of the studies on AutoPap by Wilbur and colleagues[114,115] with those assessing PAPNET helps to contrast the way in which the two automated devices operate. PAPNET, as used for instance in the PRISMATIC trial, generates a computer summary of the slide as data tapes containing computer images and locations on the original slide of the 128 areas 'of least normal appearance'. The automated system makes no designation as to whether the slide is normal or abnormal. This is done by a screener viewing the 128 selected images on the data tape, viewed via a personal computer. The screener may refer back to the original slide viewed via a traditional light microscope using the location information. In the PRISMATIC trial (in the context of the UK system), slides judged as normal were then subject to manual rapid review and slides designated as abnormal were subject to full manual review before reporting.

In contrast, AutoPap essentially separates slides into those with a very low probability of containing an abnormality ('no review') and those where the probability of abnormality is sufficiently high for a primary manual screening to be required ('review'). 'Review' slides are further ranked according to the likelihood of abnormality. The probability of abnormality in 'no review' slides is deemed to be so low that such slides are recommended to be reported as normal and archived without any further manual examination. In a system similar to that used in the USA, the ranking of the 'review' slides can be used to select preferentially those slides where the likelihood of abnormality is detected to be highest (where abnormality has not already been detected on the primary manual screening), which are then subject to full manual rescreening as part of quality assurance. This is in contrast to the usual means, which is a 10% random selection of slides designated as 'within normal limits' after the primary manual screen. In the studies by Wilbur and colleagues, the AutoPap Primary Screening System (the successor to the AutoPap 300 QC) is used in these ways to amplify both the primary screening step and the quality assurance step. However, referral back to *Figure 3* emphasises that this system, evaluated in the studies by Wilbur, still differs significantly from the AutoPap GS System currently being marketed, particularly with respect to the incorporation of location guidance systems.

Study population
As shown in *Table 28*, four studies[107,110,112,114] met the ideal of being a prospective cohort design. These studies should be closer to the case-mix actually experienced in the respective local screening population, a contention supported by their greater similarity to the profile of cytological diagnoses obtained in the NHSCSP, than those compiled retrospectively[106,110,113] from artificial case-mixes retrieved from the archives.

There are several issues of interest. First, the studies come from a diverse range of locations and so the proportions of preinvasive lesions may be expected to vary with the different population sources. Second, the prevalence of disease varies dramatically across the studies, CIN 1+ ranging from 1 to 38%. Even confining the analysis to the four studies of prospective cohort design, the range for CIN 1+ was from 1.2 to 6.3%. The prevalence of invasive disease also varied widely. Ignoring the special case of Veneti and colleagues, where the study sample was only 24,[113] the proportion of invasive lesions ranged from 0.008 to 0.2%, nearly 30 times more. The reasons for such variations do not just relate to the differing population characteristics. As well as there being a number of classification systems used here, there is the problem of ascertaining the true level of disease, as the reference standard has not been applied to the entire population in all but one case. The cytology classifications from the study arms have been used as a surrogate reference standard for a number of the results in the table.

TABLE 28 *Sensitivity analysis: study design and disease spectrum in included studies*

Study	Design and location	No. of slides	Spectrum (%)					
			Negative	Equivocal	CIN1	CIN2	CIN3	Invasive
Doornewaard *et al.*, 1999[106]	Retrospective, case-mix, Utrecht, The Netherlands	6,063	86	0	8.6			5.4
Sherman *et al.*, 1998[107]	Prospective, cohort Guanacaste, Costa Rica	7,323	89.6	6.7	2.1	1.4		0.2
Duggan and Brasher, 1997[110]	Prospective, cohort Alberta, Canada[a]	5,037	92.0	6.8[b]	1.0	0.2		0
Lerma *et al.*, 1998[111]	Retrospective, case-mix, Barcelona, Spain	163	77.3	9.2[c]	6.7	6.7		0
PRISMATIC, 1999[112]	Prospective, cohort, SE England, UK[d]	20,008	82.3	11.4[e]	3.8	1.5	0.8	0.2
Veneti *et al.*, 1999[113]	Retrospective, case-mix, Athens, Greece	24	0	62.5[f]	20.8	0	8.3	8.3
Wilbur *et al.*, 1998, 1999[114,115]	Prospective, cohort, New York, USA[g]	25,124	93.7	4.9[h]	1.1	0.27		0[i]
NHSCSP: % of slides in categories, negative, inadequate or borderline, mild, moderate, severe dyskaryosis		4.25 million	82	14	2.4	0.9		0.7

[a] Percentages for spectrum of disease estimated from study data. Based on the average number for each cytology classification diagnosed by both the PAPNET arm and the control arm.
[b] Equivocal: unsatisfactory (0.8%), ASCUS/AGUS (3.6%), benign cellular changes (3.0%), miscellaneous (0.2%).
[c] Equivocal: ASCUS (1.2%), koilocytosis (8.0%).
[d] Percentages for spectrum of disease based on the average number for each cytology classification diagnosed by both the PAPNET arm and the control arm.
[e] Equivocal: inadequate (6.9%), borderline (4.5%).
[f] Equivocal: HPV diagnosed on biopsy.
[g] Information on spectrum from article and product insert for AutoPap.
[h] Equivocal: unsatisfactory (0.7%), ASCUS (4.0%), AGUS (0.2%).
[i] Number of invasives: 3/25124 (0.008%).

Therefore, these provide estimates only. Finally, sample sizes also vary, from 24 to 25,000.

All of these observations predict that for apparently similar screening systems there are likely to be differences in the absolute levels of test performance measured from one study to the next, as they are being performed on markedly different populations.

Reference standard

The reference standard for each of the studies is detailed in *Table 29*. The quality of the reference standard chosen by each of the studies included in the sensitivity analysis varied. In addition to the two studies included in the main analysis which used histology to identify the correct diagnosis, two other studies used this too.[111,113] The study by Lerma and colleagues[111] had little wrong with its design and deserves a mention for that reason. However, owing to inadequate reporting, it was impossible to calculate directly the test sensitivity and specificity, and this was the only reason it was not included in the main analysis. The study by Veneti and colleagues[113] had biopsy results for the whole study sample (24 cases), but some were taken up to 2 years after the smear, raising the possibility of low-grade *de novo* lesions. As for the study by Lerma, the only reason it was not included in the main analysis was that sensitivity and specificity could not be directly calculated.

The remaining three additional included studies[110,112,114] generally relied on independently obtained consensus on cytology results where diagnosis was discrepant between one arm and the other. Concern has been expressed about the validity of estimates of test performance based on such reference standards, particularly because there is little or no information on the extent to which both tests have misclassified an abnormal slide as normal.[97] Unfortunately, apparently 'true-

TABLE 29 *Sensitivity analysis: reference standard used in included studies*

Study	Reference standard	Reviewer blinding
Doornewaard *et al.*, 1999[106]	PALGA database True-negative smears had 7 years of follow-up Abnormal cases nearly all had biopsy in 7 years of follow-up False negatives were smears initially diagnosed as normal in 1988 and shown to have abnormality in the following 7 years	NA
Sherman *et al.*, 1998[107]	Based on all pathology material available (smear, colposcopy report, biopsy and cervigrams) reviewed in USA by pathologist Colposcopic referral after positive pelvic examination, cytology or cervigram Biopsy on 93% HSIL, 100% cancer and 39% LSIL	Unlikely
Duggan and Brasher, 1997[110]	Discrepant cases only (no absolute correlation) reviewed by a panel of two cytopathologists and a consensus opinion reached	Yes
Lerma *et al.*, 1998[111]	All 163 patients had colposcopy 111 had lesions at colposcopy and had biopsy. The remaining 52 had no lesion at colposcopy, and were followed up annually by smears	Unlikely
PRISMATIC, 1999[112]	All of the abnormal smears and 298 randomly selected negative smears were independently reviewed by one expert cytopathologist who was blind to the results in the two study arms Discrepancy of slides between the two study arms of one or more grades was reviewed by two other experts and a consensus diagnosis was reached	Yes
Veneti *et al.*, 1999[113]	Biopsy of all 24 cases 11/24 were biopsied up to 2 years after the smear, 7/12 were taken 1 year afterwards and 6/24 less than 1 year afterwards	NA
Wilbur *et al.*, 1998, 1999[114,115]	Concordant results received no further review Discrepant diagnoses had blinded adjudication by an expert panel of three cytopathologists. Majority decision taken as diagnosis; if not achieved, further review at a multiheaded microscope until consensus was reached Biopsy/cytology follow-up of all HSIL; 27/70 HSIL had biopsy	Yes

negative' slides often remain unverified (for legitimate reasons in the case of biopsy confirmation) and result in optimistic estimates of sensitivity and specificity. An alternative measure of test performance, relative TPR and relative FPR, can be used in studies relying on scrutiny of discordant results.[96] They do not require verification of apparently 'true negatives', but do require verification of all slides identified as abnormal by one test, the other or both tests. Even where relative TPR and FPR can be calculated, caution needs to be applied in interpreting test performance results based on a reference standard relying principally on analysis of discordant cytological findings. This caution needs to be extremely great when there is verification of neither apparently 'true negatives' nor apparently 'true positives'. In this situation it is likely that one can only rely on the direction of the difference between the test performance in one arm and the other; the magnitude of the difference and the absolute values of test performance are likely to be erroneous.

Of the three studies that adopted cytology as the reference standard, only one study verified all of the abnormal slides independently.[112] In the PRISMATIC trial, after normal programme protocol had been followed, abnormal slides were further independently assessed in a masked fashion, by one cytopathologist. Discordant cases between the two study arms were subject to panel review, by two independent cytopathologists. A small, random sample of negative cases was also verified. Such an approach allows for more accurate estimates of the true and false negatives to be made. Unfortunately, the sample size of 298 was too small to obtain a precise estimate of these. For instance, it is known from meta-analyses that the sensitivity of the Pap test can be as low as 50%.[21,65] If an optimistic estimate of 75% is anticipated, this would suggest a prevalence of disease of 3% in test negative subjects. At this prevalence, for 1% precision, the sample size would have to be greater than 1000. At lower anticipated levels of sensitivity, the sample size would need to be greater still.

The studies by Duggan and Brasher[110] and Wilbur and colleagues[114] only verified slides with discordant results between the two tests. Concordant results, normal or abnormal, remained unverified, although the effect of failure to verify concordant abnormals is mitigated by the fact that test results not agreeing as to the specific type of abnormality would be considered discordant.

Sensitivity analysis: quality of included studies

Shown in *Table 30* is a summary of the quality of the two original studies and five additional studies. Again, it must be emphasised that studies have not been scored. The checklist provides a standard framework against which the strengths and weakness of each study included in the sensitivity analysis can be assessed. The bottom row demonstrates the overall frequency with which particular criteria were met.

As with the studies included in the main analysis, all of the additional studies included in the sensitivity analysis had clearly identifiable weaknesses, causing concern about the interpretation of the test performance results reported. Arguably, however, in general terms these were of no greater magnitude than the concerns encountered in the interpretation of the studies in the main analysis. In this respect, the two-arm plus reference standard design is reassuring. In this case even where sources of bias can be identified, provided one is reasonably confident that the testing in each arm has been conducted in an independent manner, and that the reference standard is reasonably independent of (or equally dependent on) the information provided from each test, these can be assumed to be operating equally between each arm. A proviso to this, arising from an issue only encountered among the additional studies in the sensitivity analysis, was the use of analysis of discordant slides as the basis for the reference standard. However, although much greater caution is dictated concerning the interpretation of absolute values of test performance and the magnitude of their differences between the automated and current practice arms of the study, the two-armed designs again provide reassurance because, qualitatively at least, the performance of each arm is affected in the same way. When discordant analysis alone is performed (i.e. without concordant verification) it may be shown (see Appendix 8) that if $\text{sensitivity}_{\text{test1}}$ is greater than $\text{sensitivity}_{\text{test2}}$ using discordant analysis, then true $\text{sensitivity}_{\text{test1}}$ is greater than true $\text{sensitivity}_{\text{test2}}$ in absolute terms, although statistical significance does not follow.

Thus, it seems reasonable at least to explore whether conclusions based on the main analysis can be confirmed or extended using the additional studies included in the sensitivity analysis. In this respect, the inclusion of a reasonably robust study on AutoPap is important, whereas none was present in the main analysis.

Unfortunately, the promise of additional information on automation used in a rescreening mode offered by the two included studies by Lerma and Veneti was not realised because of limitations on their external generalisability and limited reporting of the results.[111,113]

It is also worth examining the general strengths and weaknesses across all the studies included in the sensitivity analysis. An important general strength in relation to the interpretation of two-armed study designs alluded to above is that all but one included study,[107] was felt to be relatively free from review bias (criterion 5). Other universal strengths were the inclusion of a defined reference standard (although this was implicit in the inclusion criteria) and definition of the spectrum of 'disease' in the population whose slides were included in the studies.

With respect to weaknesses, some, such as the fact that four of the seven studies had investigators who had some involvement with the manufacturer of the device under study, and that only three avoided verification bias, were issues of concern. The implications of the latter are discussed above. The former has already been alluded to earlier in the report, the issue being not so much that there should not be industry-sponsored research, but that there should be a better balance between industry-funded and independent, publicly funded research, particularly where the technology has reached a stage where wide-scale implementation is being considered. Other general failings, such as failure to report reproducibility, are perhaps as much of a challenge to the validity of the checklist as they are to the validity of the conclusions based on the included studies. Failure to report information on reproducibility is a case in point. Although it is surprising that so few evaluations of test performance seem to consider the importance of interpreting their results in the light of knowledge about reproducibility, it does not fundamentally undermine the internal validity of the research findings reported. Thus, although the inclusion of the need to have information on reproducibility in the quality assessment checklist may be useful to remind reviewers that this is essential information and that they should review

TABLE 30 *Sensitivity analysis: study quality in included studies*

Study	1	2	3	4	5	6	7	8	9	Weaknesses identified	Comment
Doornewaard *et al.*, 1999[106]	×	✓	✓	✓	✓	×	✓	✓	✓	Retrospective recruitment No confidence intervals	Neither weakness felt to constitute a major threat to validity
Sherman *et al.*, 1998[107]	✓	✓	✓	×	×	✓	✓	×	×	Possible verification bias Possible review bias No information on test reproducibility Industry sponsored	Possibility of review bias felt to constitute an important threat to validity
Duggan and Brasher, 1997[110]	✓	✓	✓	×	✓	×	✓	×	✓	Definite verification bias (intrinsic to reference standard used) No confidence intervals No information on test reproducibility	Reliance on reference standard based on analysis of slides with discordant results, with no verification of concordant normals or abnormals, dictates extreme caution required in interpretation of the results
Lerma *et al.*, 1998[111]	×	✓	✓	✓	✓	×	×	×	✓	Retrospective recruitment: artificial case-mix No confidence intervals No indeterminate result information No information on test reproducibility	Main concern relates to poor generalisability of results. Value of study also greatly limited by inadequate presentation of results
PRISMATIC, 1999[112]	✓	✓	✓	×	✓	✓	✓	×	✓	Definite verification bias (intrinsic to reference standard used) No information on test reproducibility	Reliance on reference standard based on analysis of slides with discordant results dictates caution required in interpretation of the results
Veneti *et al.*, 1999[113]	×	✓	✓	✓	✓	✓	×	×	×	Retrospective recruitment: highly artificial case-mix No indeterminate result information No information on test reproducibility Industry sponsored	Main concern relates to poor generalisability of results. Value of study also greatly limited by poor presentation of results
Wilbur *et al.*, 1998, 1999[114,115]	✓	✓	✓	×	✓	✓	✓	×	×	Definite verification bias (intrinsic to reference standard used) No information on test reproducibility Industry sponsored	Reliance on reference standard based on analysis of slides with discordant results, with no verification of concordant normals or abnormals, dictates extreme caution required in interpretation of the results
Totals	4	7	7	3	6	4	5	1	4		

1, Study subjects recruitment; 2, spectrum composition; 3, reference standard; 4, avoidance of verification bias; 5, avoidance of review bias; 6, precision of results for test accuracy; 7, presentation of indeterminate test results; 8, test reproducibility; 9, industry sponsorship.

it, it is unclear how it should alter the overall conclusions on what the impact on test performance should be. Implications for the review of failure to give the precision of estimates of test performance are similarly unclear. Where most reviews would recalculate confidence intervals, it is difficult to see why their absence in the original studies has a bearing on their internal validity, provided the raw data are clearly presented.

Sensitivity analysis: test performance results in included studies

The results of seven studies included in the sensitivity analysis [two from the main analysis (in the top two rows) and five additional] are shown in *Tables 31–34*. As already stated, even with the additional studies there were none that evaluated the alternative technologies of LBC with or without HPV testing as a comparator to automated screening. The comparisons below are divided according to the disease and test threshold. Results are presented subdivided into four groupings of disease and test thresholds.

Disease: HSIL+ or moderate+ or CIN3+; test: HSIL+ or moderate+

Results for this grouping are presented in *Table 31*. From the five extra studies included, four additional estimates of sensitivity gain and four additional estimates of specificity gain could be obtained. For the fifth study, by Duggan and Brasher, the reference diagnosis could not be derived for the whole population, and therefore it was not possible to calculate the sensitivity and specificity in any of the threshold groupings.[110]

Across all the included studies there was only one statistically significant change in sensitivity and one in specificity. These both occurred in one of the studies in the main analysis,[107] in which sensitivity was reduced and specificity increased. The pattern of results in the additional studies included in the sensitivity analysis was more typical of the other study included originally,[106] with small, non-statistically significant gains in sensitivity and unchanged specificity.

Considering the differing impact of automation used in primary screening, quality assurance or both, there is little evidence from the included studies that the impact of automation differed strikingly between primary screening mode (three contributing studies), quality assurance mode (two studies) and combined mode (one study). It should, however, be reinforced that the nature of the two studies that used automation in rescreening means that they are likely to give a poor indication of what the impact of introducing automation in such a mode may actually be. In consequence, no further comments are made about whether the impact of introducing automation to the quality assurance step may be systematically different to that of introducing it in the primary screening step.

Similarly, there is little evidence that the pattern of impact on test performance is any different for the majority of studies that considered PAPNET to the single included study that used AutoPap, albeit an older version.

A very notable feature of the results is the enormous variation in the absolute levels of sensitivity seen across the included studies, ranging from 0 to 97% in the automation arms, and 0 to 93% in the manual screening arms.

Disease: HSIL+ or moderate+ or CIN3+; test: LSIL+ or Mild+

Results for this grouping are presented in *Table 32*. From the five extra studies included, four additional estimates of sensitivity gain and four additional estimates of specificity gain could be obtained. For the fifth study, by Duggan and Brasher, the reference diagnosis could not be derived for the whole population, and therefore it was not possible to calculate the sensitivity and specificity in any of the threshold groupings.

Across all the included studies there were no statistically significant changes in sensitivity and two in specificity, both increases. One occurred in one of the studies in the main analysis,[107] and as for the result for the previous threshold grouping for this study the statistically significant increase in specificity was associated with a loss in sensitivity. In the second study showing a statistically significant increase in specificity[112] there was a simultaneous small improvement in sensitivity. The pattern of results in the remaining studies was generally one of improved sensitivity (not statistically significant) associated with slightly worsened specificity (again not statistically significant). An exception was one of the studies from the main analysis,[106] which showed a small worsening of sensitivity associated with a small deterioration in specificity (neither statistically significant). There was thus little consistency in results at this test/disease threshold grouping.

The study that evaluated combined use of automation at the primary screening step with use in quality assurance, which is also the study that

TABLE 31 *Sensitivity analysis: test performance summary where threshold for test = moderate+ or HSIL+ and threshold for disease = HSIL+ or moderate+ or CIN3+*

Study	Technology/ comparator (disease threshold)	*n*	Sensitivity: technology (95% CI)	Sensitivity: comparator (95% CI)	Sensitivity gain (technology – comparator)	Specificity: technology (95% CI)	Specificity: comparator (95% CI)	Specificity gain (technology – comparator)
Doornewaard *et al.*, 1999[106]	PAPNET primary screening/manual (CIN3+/7-year follow-up)	6,063 (325: 5,738)	19.7% (15.7 to 24.4) (HSIL+)	14.5% (11.1 to 18.7)	+5.2% ns	99.5% (99.3 to 99.7) (HSIL+)	99.6% (99.4 to 99.7)	–0.1% ns
Sherman *et al.*, 1998[107]	PAPNET primary screening/manual (HSIL+)	7,323 (114: 7,209)	49.1% (40.1 to 58.2) (HSIL+)	67.5% (58.5 to 75.4)	–18.4%*	99.9% (99.7 to 99.9) (HSIL+)	99.2% (98.9 to 99.4)	+0.7%*
Duggan and Brasher, 1997[110]	PAPNET primary screening/manual (HSIL+)	5,037	Unable to calculate sensitivity at any threshold level			Unable to calculate specificity at any threshold level		
Lerma *et al.*, 1998[111]	PAPNET rescreening/manual (HSIL+)	163 (11:152)	63.6% (35.4 to 84.8) (HSIL+)	45.5% (21.3 to 72.0)	+18.1% ns	100% (97.5 to 100) (HSIL+)	100% (97.5 to 100)	0% ns
PRISMATIC, 1999[112a]	PAPNET primary screening/manual (Moderate+)	20,008 (582: 19,426)	65.9% (62.0 to 69.6) (Moderate+)	64.3% (60.4 to 68.1)	+1.6% ns	99.4% (99.3 to 99.5) (Moderate+)	99.3% (99.2 to 99.4)	+0.1% ns
Veneti *et al.*, 1999[113]	PAPNET rescreening/manual (HSIL+)	24 (4:20)	0% (0 to 65.8) (HSIL+)	0% (0 to 65.8)	0% ns	100% (85 to 1,100) (HSIL+)	100% (85.1 to100)	0% ns
Wilbur *et al.*, 1998, 1999[114,115]	AutoPap primary + QC screening/manual (HSIL+)	25,124 (70: 25,054)	97.1% (90.2 to 99.2) (HSIL+)	92.8% (84.3 to 96.9)	+4.3% ns	100% (99.97 to 100) (HSIL+)	100% (99.98 to 100)	0% ns

[a] Calculations of sensitivity and specificity have been modified to include false negatives at a rate of 3/298 (with one borderline, one mild and one moderate+), as reported in the article. Inadequate smears have been excluded, as analysis of actual diagnosis not given in the article.

* $p < 0.05$.

TABLE 32 *Sensitivity analysis: test performance summary where threshold for test = mild+ or LSIL+ and threshold for disease = HSIL+ or moderate+ or CIN3+*

Study	Technology/ comparator (disease threshold)	*n*	Sensitivity: technology (95% CI)	Sensitivity: comparator (95% CI)	Sensitivity gain (technology – comparator)	Specificity: technology (95% CI)	Specificity: comparator (95% CI)	Specificity gain (technology – comparator)
Doornewaard *et al.*, 1999[106]	PAPNET primary screening/manual (CIN3+/7-year follow-up)	6,063 (325: 5,738)	61.5% (56.1 to 66.7) (LSIL+)	63.4% (58.0 to 68.4)	–1.9% ns	91.8% (91.1 to 92.5) (LSIL+)	92.0% (91.2 to 92.7)	–0.2% ns
Sherman *et al.*, 1998[107]	PAPNET primary screening/manual (HSIL+)	7323 (114: 7,209)	60.5% (51.4 to 69.0) (LSIL+)	76.3% (67.7 to 83.2)	–15.8% ns	99.0% (98.7 to 99.2) (LSIL+)	96.3% (95.8 to 96.7)	+2.7%*
Duggan and Brasher, 1997[110]	PAPNET primary screening/manual (HSIL+)	5,037	Unable to calculate sensitivity at any threshold level			Unable to calculate specificity at any threshold level		
Lerma *et al.*, 1998[111]	PAPNET rescreening/manual (HSIL+)	163 (11:152)	63.6% (35.4 to 84.8) (LSIL+)	45.5% (21.3 to 72.0)	+18.1% ns	92.8% (87.5 to 95.9) (LSIL+)	97.4% (93.4 to 99.0)	–4.6% ns
PRISMATIC, 1999[112a]	PAPNET primary screening/manual (Moderate+)	20,008 (582: 19,426)	83.2% (80.0 to 86.0) (Mild+)	82.5% (79.3 to 85.3)	+0.7% ns	96.2% (95.9 to 96.5) (Mild+)	95.4% (95.0 to 95.6)	+0.8%*
Veneti *et al.*, 1999[113]	PAPNET rescreening/manual (HSIL+)	24 (4: 20)	0% (0 to 65.8) (LSIL+)	0% (0 to 65.8)	0% ns	95% (76.4 to 99.1) (LSIL+)	100% (85.1 to 100)	–5% ns
Wilbur *et al.*, 1998, 1999[114,115]	AutoPap primary + QC screening/manual (HSIL+)	25,124 (70: 25,054)	97.1% (90.2 to 99.2) (LSIL+)	92.8% (84.3 to 96.9)	+4.3% ns	99.0% (98.9 to 99.1) (LSIL+)	99.1% (98.9 to 99.2)	–0.1% ns

[a] Calculations of sensitivity and specificity have been modified to include false negatives at a rate of 3/298 (with one borderline, one mild and one moderate+), as reported in the article. Inadequate smears have been excluded, as analysis of actual diagnosis not given in the article.

* $p < 0.05$.

used AutoPap (as opposed to PAPNET),[114,115] showed a non-statistically significant gain in sensitivity associated with a small, but non-statistically significant loss in specificity. As such, at this test/disease threshold grouping, again there was neither evidence that automation applied in a combined mode has a substantially different impact to other modes, nor evidence that AutoPap was substantially different to PAPNET.

Again, there was very wide variation in the absolute levels of sensitivity seen across the included studies, ranging from 0 to 97% in the automation arms, and from 0 to 93% in the manual screening arms. As would be expected from the lowering of the test threshold with the disease threshold remaining constant, the general level of sensitivity was improved and specificity was reduced relative to the general level in the previous test/disease threshold grouping.

Disease: HSIL+ or moderate+ or CIN3+; test: ASCUS+ or borderline+
Results for this grouping are presented in *Table 33*. In the main analysis only one set of sensitivity and specificity changes was available, data not being available to calculate these changes for the study by Doornewaard and colleagues.[106] From the five extra studies included, only three additional estimates of sensitivity gain and specificity gain could be obtained. As before, in the study by Duggan and Brasher[110] the reference diagnosis could not be derived for the whole population, and in the study by Lerma and colleagues[111] data were not available to calculate sensitivity and specificity in both arms.

Across the four included studies providing information on sensitivity and specificity change, there was little consistency. Two studies showed statistically significant improvements in specificity. However, in one case this was associated with a non-statistically significant improvement in sensitivity,[107] and in the other with a very small and non-statistically significant deterioration in sensitivity.[112] In the other two studies an unchanged sensitivity was associated with a deterioration in specificity[113] and an improved sensitivity associated with a small deterioration in specificity,[114,115] none of these changes being statistically significant.

The results of the studies by Wilbur and colleagues,[114,115] which evaluated AutoPap in a combined mode, are not markedly better or worse than others available for this test/disease threshold grouping, so again analysis provides no evidence to suggest an advantage of combined mode relative to primary screening mode, or AutoPap relative to PAPNET.

Disease: ASCUS+ or borderline+; test: ASCUS+ or borderline+
Results for this grouping are presented in *Table 34*. In the main analysis only one set of sensitivity and specificity changes was available, data not being available to calculate this for the study by Doornewaard and colleagues.[106] From the five extra studies included, only three additional estimates of sensitivity gain and specificity gain could be obtained. In the study by Duggan and Brasher[110] the reference diagnosis could not be derived for the whole population, and in the study by Lerma and colleagues[111] data were not available to calculate sensitivity and specificity in both arms.

Across the four included studies providing information on sensitivity and specificity change, there were two general patterns. In two studies[107,112] a statistically significant improvement in specificity was associated with a small, but non-statistically significant deterioration in specificity. In the other two studies[113–115] improvements in sensitivity (one statistically significant, the other not) were associated with unchanged specificities.

Again, consideration of whether the results of the studies by Wilbur and colleagues,[114,115] which evaluated AutoPap in a combined mode, are typical or not of others available for this test/disease threshold grouping, suggests that there is no evidence for a markedly different impact for combined mode relative to primary screening mode, or AutoPap relative to PAPNET.

Finally, notwithstanding the fact that there were only four studies contributing information there was wide variation in the absolute levels of sensitivity, which ranged from 11 to 91% in the automation arms, and from 0 to 91% in the manual screening arms.

Summary across disease/test threshold groupings
The findings from the studies included in the sensitivity analysis are summarised in *Table 36*. In this table judgements are made about the implications of particular findings according to whether sensitivity or specificity improves or deteriorates in a statistically significant manner. The nature of these judgements is made explicit in *Table 35*.

TABLE 33 *Sensitivity analysis: test performance summary where threshold for test = ASCUS+ or borderline+ and threshold for disease = HSIL+ or moderate+ or CIN3+*

Study	Technology/ comparator (disease threshold)	*n*	Sensitivity: technology (95% CI)	Sensitivity: comparator (95% CI)	Sensitivity gain (technology – comparator)	Specificity: technology (95% CI)	Specificity: comparator (95% CI)	Specificity gain (technology – comparator)
Doornewaard *et al.*, 1999[106]	PAPNET primary screening/manual (CIN3+/7-year follow-up)	6,063 (325: 5,738)	Unable to calculate sensitivity and specificity at this range of test threshold levels					
Sherman *et al.*, 1998[107]	PAPNET primary screening/manual (HSIL+)	7,323 (114: 7,209)	86.0% (78.4 to 91.2) (ASCUS+)	79.8% (71.5 to 86.2)	+6.2% ns	97.0% (96.6 to 97.4) (ASCUS+)	94.6% (94.1 to 95.1)	+2.4%*
Duggan and Brasher, 1997[110]	PAPNET primary screening/manual (HSIL+)	5,037	Unable to calculate sensitivity at any threshold level			Unable to calculate specificity at any threshold level		
Lerma *et al.*, 1998[111]	PAPNET rescreening/manual (HSIL+)	163 (11:152)	100% (74.1 to 100) (ASCUS+)	Unable to calculate sensitivity at this threshold		37.5% (30.2 to 45.4) (ASCUS+)	Unable to calculate specificity at this threshold	
PRISMATIC, 1999[112 a]	PAPNET primary screening/manual (Moderate+)	20,008 (582: 19,426)	90.7% (88.0 to 92.6) (Borderline+)	90.8% (88.3 to 92.9)	–0.1% ns	92.2% (91.8 to 92.6) (Borderline+)	90.1% (89.6 to 90.5)	+2.1%*
Veneti *et al.*, 1999[113]	PAPNET rescreening/manual (HSIL+)	24 (4:20)	0% (0 to 65.8) (ASCUS+)	0% (0 to 65.8)	0% ns	95% (76.4 to 99.1) (ASCUS+)	100% (85.1 to100)	–5% ns
Wilbur *et al.*, 1998, 1999[114,115]	AutoPap primary + QC screening/manual (HSIL+)	25,124 (70: 25,054)	97.1% (90.2 to 99.2) (ASCUS+)	92.8% (84.3 to 96.9)	+4.3% ns	95.5% (95.2 to 95.7) (ASCUS+)	95.8% (95.6 to 96.1)	–0.3% ns

[a] Calculations of sensitivity and specificity have been modified to include false negatives at a rate of 3/298 (with one borderline, one mild and one moderate+), as reported in the article. Inadequate smears have been excluded, as analysis of actual diagnosis not given in the article.

* $p < 0.05$.

TABLE 34 *Sensitivity analysis: test performance summary where threshold for test = ASCUS+ or borderline+ and threshold for disease = ASCUS+ or borderline+*

Study	Technology/ comparator (disease threshold)	*n*	Sensitivity: technology (95% CI)	Sensitivity: comparator (95% CI)	Sensitivity gain (technology – comparator)	Specificity: technology (95% CI)	Specificity: comparator (95% CI)	Specificity gain (technology – comparator)
Doornewaard *et al.*, 1999[106]	PAPNET primary screening/manual (ASCUS+)	6,063	Unable to calculate sensitivity and specificity at this range of test threshold levels					
Sherman *et al.*, 1998[107]	PAPNET primary screening/manual (ASCUS+)	7,323 (269: 7,054)	66.5% (60.7 to 71.9) (ASCUS+)	69.5% (63.8 to 74.7)	–3.0% ns	98.1 % (97.8 to 98.4) (ASCUS+)	95.9% (95.4 to 96.3)	+2.2%*
Duggan and Brasher, 1997[110]	PAPNET primary screening/manual (ASCUS+)	5,037	Unable to calculate sensitivity at any threshold level			Unable to calculate specificity at any threshold level		
Lerma *et al.*, 1998[111]	PAPNET rescreening/manual (ASCUS+)	163 (37:126)	100% (90.6 to 100) (ASCUS+)	Unable to calculate sensitivity at this threshold		45.2% (36.8 to 53.9) (ASCUS+)	Unable to calculate specificity at this threshold	
PRISMATIC, 1999[112 a]	PAPNET primary screening/manual (Borderline+)	18,477 (1,877: 16,600) and 18,715 (1,921: 16,794)	91.1% (89.8 to 92.3) (Borderline+)	91.4% (90.1 to 92.6)	–0.3% ns	98.6% (98.4 to 98.8) (Borderline+)	96.5% (96.2 to 96.8)	+2.1%*
Veneti *et al.*, 1999[113]	PAPNET rescreening/manual (ASCUS+)	24 (9:15)	11.1% (2.0 to 43.5) (ASCUS+)	0% (0 to 29.9)	+11.1% ns	100% (79.6 to 100) (ASCUS+)	100% (79.6 to 100)	0% ns
Wilbur *et al.*, 1998, 1999[114,115]	AutoPap primary + QC screening/manual (ASCUS+)	25,124 (1,397: 23,727)	85.8% (83.9 to 87.6) (ASCUS+)	79.2% (77.0 to 81.2)	+6.6%*	100% (99.98 to 100) (ASCUS+)	100% (99.98 to 100)	0% ns

[a] Calculations of sensitivity and specificity have been modified to include false negatives at a rate of 3/298 (with one borderline, one mild and one moderate+), as reported in the article. Inadequate smears have been excluded, as analysis of actual diagnosis not given in the article.

* $p < 0.05$.

TABLE 35 *Sensitivity analysis: implications of particular combinations of sensitivity and specificity results*

Results: automation relative to manual system	**Sensitivity**		
Specificity	**Statistically significant improvement**	**No statistically significant change**	**Statistically significant deterioration**
Statistically significant improvement	Definite advantage	Probable advantage (specificity)	Neutral (trade-off)
No statistically significant change	Probable advantage (sensitivity)	Neutral	Probable disadvantage (sensitivity)
Statistically significant deterioration	Neutral (trade-off)	Probable disadvantage (specificity)	Definite disadvantage
Statistically significant defined as no overlapping of 95% CI.			

For each disease/test threshold grouping *Table 36* clarifies that the pattern of results is predominantly neutral with respect to definite conclusions as to whether automation offers an advantage with respect to test performance over manual systems. Unexpectedly, some advantage with respect to specificity was apparent in two studies. One study at the lowest disease and test thresholds suggested an advantage with respect to sensitivity.

Beyond the general pattern, drawing conclusions about subgroups, with respect to either the mode in which automation was deployed or the type of screening, is extremely difficult where the number of included studies is so small. With respect to drawing definite conclusions, *Table 36* makes clear that there is no strong evidence that either the mode of application of automation or the type of device is associated with a different pattern of results to that described overall. However, neither does it exclude an important difference in pattern.

From the point of view of generating and testing further hypotheses, the observation that the single study showing a probable advantage of automation with respect to sensitivity was the only study where the automated device was AutoPap, and the only study where automation was applied to both the primary and quality assurance steps is of importance.

Sensitivity analysis: provisos concerning interpretation of pattern of results across included studies

As for the main analysis, there is an important caveat to the above analysis. It arises from the qualitative method of summary used. Great care must be applied when using vote-counting approaches, and to be conclusive, high thresholds for a 'vote' to be registered for or against an intervention are an inevitable consequence of necessary caution. Thus, in this analysis a statistically significant improvement or decrease in both sensitivity and specificity was required to provide clear evidence of benefit or disbenefit. No such results were apparent, although there were two results showing a statistically significant improvement in specificity in three of the four disease/test threshold groupings, and one statistically significant improvement in sensitivity in one of the four disease/test threshold groupings.

The vast majority of findings were neutral, yet it is possible that these results may conceal important patterns confirming benefit and disbenefit. Quantitative analysis, meta-analysis in particular, could achieve this. Unfortunately, although techniques are available to achieve this, the demonstrated level of clinical and methodological heterogeneity, confirmed by extreme variability in sensitivity noted in many disease/test threshold groupings, challenges the validity of attempting to summarise the results quantitatively. Quantitative techniques could be of value to explore the relationship between the results and the many factors varying between the included studies (as a minimum, nature of intervention, nature of comparator, nature of reference standard, population and study quality). However, given the number of variables relative to the number of studies, this exercise was not attempted. A similar argument holds for why conclusions concerning subgroups, particularly with respect to whether AutoPap is more or less effective than PAPNET, could not be more definitive.

The fundamental problem with respect to the evidence base for automated image analysis is that the current number of studies of reasonable

TABLE 36 *Sensitivity analysis: summary of key test performance findings*

	Thresholds		*n*	Advantage			Neutral	Disadvantage		
	Disease	Test		Definite	Probable			Probable		Definite
					Se	Sp		Sp	Se	
All	HSIL+	HSIL+	6				6[a]			
	HSIL+	LSIL+	6			2	4			
	HSIL+	ASCUS+	4			2	2			
	ASCUS+	ASCUS+	4		1	2	1			
Primary	HSIL+	HSIL+	3				3[a]			
	HSIL+	LSIL+	3			2	1			
	HSIL+	ASCUS+	2			2				
	ASCUS+	ASCUS+	2			2				
Rescreen				Results are very unlikely to reflect the impact of introducing automation as part of the QA screening step						
Combined	HSIL+	HSIL+	1				1			
	HSIL+	LSIL+	1				1			
	HSIL+	ASCUS+	1				1			
	ASCUS+	ASCUS+	1		1					
PAPNET	HSIL+	HSIL+	5				5[a]			
	HSIL+	LSIL+	5			2	3			
	HSIL+	ASCUS+	3			2	1			
	ASCUS+	ASCUS+	3			2	1			
AutoPap	HSIL+	HSIL+	1				1			
	HSIL+	LSIL+	1				1			
	HSIL+	ASCUS+	1				1			
	ASCUS+	ASCUS+	1		1					

[a] One study was a trade-off between a statistically significant decrease in sensitivity and a statistically significant increase in specificity.

Se, sensitivity; Sp, specificity; HSIL+, high-grade squamous intraepithelial lesion or worse (invasive); LSIL+, low-grade squamous intraepithelial lesion or worse (HSIL or invasive); ASCUS+, atypical squamous cells of unknown significance or worse (LSIL, HSIL or invasive).

validity is wholly inadequate to answer an evaluative problem as complex as the one in question.

Overall conclusions from main analysis and sensitivity analysis on test performance

The main analysis concluded that the impact of automation used in a quality assurance mode, and the impact of introducing automation if the current manual system incorporated LBC or HPV triage, were unknown. These conclusions are unchanged by the sensitivity analysis. The absence of research on particular areas is important and has implications for future research. The need for greater quantities of research to provide adequate answers to an evaluative question of the complexity of the one under consideration also needs to be recognised.

The main analysis also tentatively concluded that PAPNET used in primary screening mode may lead to a small improvement in specificity at the expense of loss of sensitivity. The additional studies in the sensitivity analysis were mainly neutral with respect to advantages and disadvantages being offered by automation, suggesting that this should be the main conclusion overall.

As such, the evidence on test performance identified does not support implementation of automation, but nor does it conclusively demonstrate that it is ineffective in this respect and that further development of the technology and research on it is not justified. Observations arising from the sensitivity analysis potentially deserving further testing were the suggestion that specificity may be improved without compromising

sensitivity. The possibility that an older version of AutoPap, used to enhance both the primary and the quality assurance screening steps in combination, may improve sensitivity is also of interest. However, it needs to be considered that this was only demonstrated where the thresholds for disease and test were low (ASCUS+), and that any improvement relates to identification of disease of uncertain significance with respect to whether a women does or does not develop invasive disease in the context of a screening programme, rather than an individual cycle.

Unfortunately, many of these conclusions are probably negated as the basis for further action and research by the fact that the main automated device on which research has been conducted, PAPNET, is no longer commercially available. Further, the single reasonably valid study encountered on the automated device that is commercially available is on a version of AutoPap that is over 5 years out of date and has clearly been superseded by a version with important differences which may affect test performance. Further publicly funded primary research on the impact of currently commercially available versions of automated devices is required. Such a study needs to anticipate the possibility of LBC and HPV triage, by assessing the impact of automation when added to systems incorporating these elements. Problems encountered with interpretation of studies included in this review need to be taken into account in the design of new studies, and it should be ensured that they are large enough to ensure sufficient precision to confirm or refute improvements in sensitivity and/or specificity that will confer an important advantage for automation and are agreed in advance. Conducting the study in the context of the system where it is to be implemented has definite advantages as far as the generalisability of any new research is concerned; in this topic there are important issues that may make application of a piece of research conducted in one country or setting to another more problematic than evaluations of other new technologies.

In the context of the health technology assessment project as a whole, careful consideration needs to be given as to how the empirical observations about potential impact of automation on test performance are reflected in the model of cost-effectiveness. The impact most compatible with the available data is that sensitivity and specificity are not altered. However, this denies the possibility, which has not been excluded statistically, that clinically important improvements in sensitivity, specificity or both may occur. Modelling the impact of upper ranges of sensitivity and specificity changes encountered may be useful for the purposes of providing optimistic illustrations of cost-effectiveness. Suggested combinations of improved sensitivity and specificity that may fulfil this purpose and be compatible with the results encountered are suggested as:

- sensitivity +5% or +10%; specificity +0%
- sensitivity +0%; specificity +1% or +2%
- sensitivity +5% or +10%; specificity +1% or +2%.

Given that the available data may be reasonably argued to be barely generalisable to the current device in the current circumstances, an alternative approach to be carried forward to the modelling could be to make no assumptions about the likely impact on test performance, and instead to use the model to indicate what level of improvement in sensitivity alone, specificity alone or combined sensitivity and specificity gain may be required to make automation cost-effective. Such information could then be used as the basis for power calculations for future research.

As well as suggesting the need for further primary research, the need for further methodological research on the evaluation of test performance of screening tests, their appraisal and methods to summarise their results quantitatively is also indicated (see Chapter 10).

Part 2: Reproducibility

Introduction

The purpose of this part of the review was to assess the ability of automated machines to provide the same or similar results when the same slide is presented on two occasions, or the same slide is presented to two machines of the same type. This was achieved by systematically reviewing research assessing the reproducibility of automated image analysis devices. Reproducibility is not only an important component of effectiveness in its own right, but if deficient it will contribute to disappointing test performance. As already alluded to, knowledge about reproducibility is thus important in considering results on test performance.

Methods

The search was effectively the same as that for test performance. The studies, which passed step 1 of the inclusion criteria from Part 1, were screened

TABLE 37 *Number of studies included and excluded in review of reproducibility*

	Update search 1998–2000	Foreign language 1990–1997	Studies considered by McCrory *et al.*[12]
Unduplicated articles originally considered	163	38	12
Step 1:			
Remaining	45	0	12
Step 2:			
Excluded	40		11
Remaining	5		1
Included	5	0	1

using a truncated version of step 2. For a study to be included for this part of the appraisal it needed to satisfy the following two criteria.

1. The study used a two-armed design.
2. The study attempted to evaluate one of the types of variation in performance of the machine, as defined earlier.

As before, a two-armed design implies that both arms are tested on the same set of slides, but the nature of the study arms was different. For a given type of technology, the design may be one of machine versus machine or observer versus observer. In the former situation the study will be evaluating intramachine or intermachine variability, and in the latter instance the objective would be to ascertain intraobserver or interobserver variability in using the technology.

As consistency of performance was the objective, not accuracy, insistence on a reference standard was not considered important for this part of the appraisal.

Data were abstracted from included studies using a proforma. To measure the level of inter- and intra-rater agreement, weighted kappa scores (using methods described by Altman[105]) have been calculated based on the data reported in the article, and 95% confidence intervals have been calculated using formulae derived by Fleiss.[86] Where this was not possible, the outcomes reported by the study authors have been quoted. Qualitative summary of results was used, with conclusions being based on the patterns revealed in clearly tabulated results. Interpretation of the weighted kappa scores was in accordance with advice from Fleiss,[86] which is reproduced in *Table 41*.

Results: number of studies included

Six papers in total were included.[116–121] The component of the search from which these were identified is indicated in *Table 37*.

Results: details, quality and results of included studies

What was investigated?

All six studies evaluated the consistency of performance of the PAPNET. As for test performance, only a small subset of the possible issues needing to be considered in this project was covered by the included studies (*Table 38*). Unsurprisingly, given the results for test performance, no included studies assessed reproducibility of automation as part of a manual system incorporating LBC or HPV triage.

Both observer and machine variation were evaluated by these studies. However, consistently throughout the studies that investigated machine variation there was a problem in determining the type of machine variation being analysed; intermachine or intramachine. This appears to be a reflection of the slides being scanned at a central point, rather than at the site of investigation. Subsequently, the investigators are unable to specify whether it was the same machine doing the scanning on each occasion.

Details of the intervention

Descriptions of the automated systems for which reproducibility was assessed are given in *Table 39*. The issues raised are explored below.

Only one study compared the inter-rater reliability of PAPNET with a simultaneously measured inter-rater reliability in a conventional arm. Such a study would probably represent the most equitable evaluation of a new technology; however, all other studies confined the measurement of reproducibility to the automated system only.

As has been raised already, it was often not clear whether machine variation was investigated, and if so whether this was inter (between) or intra (within) variation. This is expanded upon in the next section. Similarly, when observer variation was being evaluated there was too much scope for

TABLE 38 *Reproducibility: scope of included studies*

Study	Automation type			Context				Type of variation[a]			
	PAPNET	AutoPap	Auto Cyte SCREEN	Pap test	Pap + HPV	LBC	LBC+ HPV	Observer		Machine	
								Intra	Inter	Intra	Inter
Doornewaard *et al.*, 2000[116]	✓			✓				✓	✓		
Doornewaard *et al.*, 1999[117]	✓			✓					✓		✓
Jenny *et al.*, 1997[118]	✓			✓					✓		✓
Mitchell and Medley, 1998[119]	✓			✓					✓		
Mitchell and Medley, 1998[120]	✓			✓					✓		✓
Mitchell and Medley, 1998[121]	✓			✓					✓		

[a] When the type of variation (inter or intra) evaluated in the study was not clear, the tick appears between the two cells.
[b] Inter-rater reliability of conventional screening tested simultaneously.

interpretation on whether all components of machine variation had been eliminated. This point is illustrated by the study by Doornewaard and colleagues,[116] where in an otherwise excellent methodology it is not made clear whether rescanning of the slides took place.

A problem particular to using the PAPNET system is that essentially it is a two-stage process consisting of scanning the slides, followed by reviewing the digital images on monitors, before triaging. However, it is uncertain whether the same or different machines were used in the scanning, as each location would probably have had more than one machine. If the slides were not rescanned by PAPNET, then the tapes of the images from the original scan could be used as the source, and the observer variation would be measured by assigning the images to different observers, or the same observer at different times, machine variation being controlled for in each case. Rescanning of the slides, however, introduces the possibility of machine variation. In such an instance whole-system variation is being measured; that is, the objective being addressed is determining whether the system as a whole is able to reproduce the same results, such as in the study by Jenny and colleagues.[118]

Study population

As in the evaluation of test performance, how consistent or inconsistent a new technology is should be assessed in a normal screening setting. Furthermore, given that the main metric for quantifying the intra-rater/inter-rater agreement is the kappa (standard or weighted) score, a statistic that varies with the prevalence of disease, it is important that the spectrum of disease tested reflects the screening population.

Table 40 shows that there were three study locations. It should be noted that the study locations were not necessarily where the PAPNET scanning was performed, as this is generally centralised. It is probable that only two scanning centres were used in the six studies: Amsterdam and Hong Kong. In the article reported by Doornewaard and colleagues, based in Utrecht,[117] they indicate that only since 1998 has the scanning station been able to be installed at participating laboratories, and that before then the scanners were centralised. Thus, in this study the slides appear to have been sent to the NSI location in Amsterdam. To re-emphasise, when slides are rescanned this poses the problem of knowing whether the same or a different machine was used in the scanning.

In general terms, the prevalence of disease in all save one study was much higher than would be expected in a normal screening population. The first study by Doornewaard and colleagues[116] is a continuation of the included study in Part 1,[106] where 6063 slides were scanned, but a subset of 1996 slides chosen at random was used for analysis. The authors claimed that the prevalence was chosen to be deliberately high; the disease spectrum consisted of 19% abnormal smears

TABLE 39 *Reproducibility: details of the intervention in included studies*

Study	Intervention description	Control for other types of variation	Blinding of study arms
Doornewaard *et al.*, 2000[116]	Screening of the same case set twice by the PAPNET 6 months apart. Cases randomly assigned to screeners for review. Same process carried out for the conventional system Intraobserver variation measured by those smears that had by chance been reassigned to the same cytotechnologist	Unclear whether machine variation controlled for	Yes
Doornewaard *et al.*, 1999[117]	Scanning of the same case set twice by PAPNET 2 months apart. Cases reviewed on each occasion by the same cytopathologist	Unclear whether same machine used in each scanning Interobserver variation controlled for	Yes
Jenny *et al.*, 1997[118]	Scanning of the same case set twice by PAPNET (but unclear whether the same or a different scanner) and then images reviewed by an independent but different cytopathologist Whole-system agreement being measured	Interobserver variation not controlled for	Yes
Mitchell and Medley, 1998[119]	Scanning of a case set once by PAPNET and the images reviewed independently by three cytotechnologists Each tile with an abnormal cell was recorded after microscopic review of the slide	Inter- and intramachine variation controlled for	Yes
Mitchell and Medley, 1998[120]	Scanning of the same case set twice by PAPNET 6 months apart. Cases reviewed by two different cytopathologists independently	Interobserver variation measured, but not clear whether machine variation controlled for Machine variation also measured	Yes
Mitchell and Medley, 1998[121]	PAPNET screening of whole case set Image tiles to all slides were reviewed independently by each of three screeners. The review also included microscopic viewing of some of the slides	Machine variation controlled for	Yes

(defined as at least LSIL), compared with an average of 2–3% abnormal smears in their routine practice.

The one exception where the prevalence was probably not higher than the locally screened population is Mitchell and Medley's study, where CIN1+ was 1.7%.[120] This is lower than would be experienced in the UK.

Results of included studies

Results, which were generally weighted kappa scores calculated from the original data, are summarised in *Tables 42–44*. These deal in turn with observer variation, machine variation and whole-system variation, and the text is subdivided in the same way.

Observer variation

Only one study allowed the evaluation of intraobserver variation using PAPNET.[116] The reported weighted kappa scores for the four screeners ranged from 0.59 to 0.78 for conventional screening and from 0.69 to 0.77 for PAPNET-assisted screening. Two of the screeners showed increased agreement with PAPNET and two decreased. The strength of the analysis is that both the reproducibility of the new technology and

TABLE 40 *Reproducibility: disease spectrum of included studies*

Study	Location	*n*	Spectrum (%)					
			Negative	Equivocal	CIN1	CIN2	CIN3	Invasive
Doornewaard *et al.*, 2000[116a]	Utrecht, The Netherlands	1996	79.7	3.1		15.1	2.1	
Doornewaard *et al.*, 1999[117]	Utrecht, The Netherlands	196	68.4	0		31.6		
Jenny *et al.*, 1997[118]	Zurich, Switzerland	1200	47.5	9.5	1.8	9.6	17.4	14.3
Mitchell and Medley, 1998[119]	Victoria, Australia	164	0	0	0	0	100	0
Mitchell and Medley, 1998[120b]	Victoria, Australia	2690	96	2.4[c]		1.7		
Mitchell and Medley, 1998[121]	Victoria, Australia	188	19.1	0		80.9		
NHSCSP: % of slides in categories, negative, inadequate or borderline, mild, moderate, severe dyskaryosis		4.25 million	82	14	2.4	0.9	0.7	

[a] Percentages for spectrum of disease based on the average number for each of four cytology classifications twice by PAPNET, and twice by conventional. Note that the quoted prevalence for LSIL+ was 18.7% (374/1996), estimated here at 17.2%.
[b] Percentages for spectrum of disease based on the average number for each of two cytology classifications twice by PAPNET. Note that the quoted prevalence for abnormal ranged between 1.49 and 1.86%.
[c] Equivocal: unsatisfactory slides.

TABLE 41 *Fleiss' guide to the interpretation of Kappa scores[86]*

Kappa score, κ	Level of agreement
$\kappa < 0.4$	Poor
$0.4 \leq \kappa < 0.6$	Fair
$0.6 \leq \kappa \leq 0.75$	Good
$0.75 < \kappa \leq 1$	Excellent

the conventional screening system were considered. It can be seen that on the whole the agreement for both modalities was good, and there was no statistically significant difference in the intraobserver agreement across all four observers. Unfortunately, insufficient data were reported in the article to verify independently the weighted kappa scores and confidence intervals.

In contrast, four studies provided evidence on the interobserver variation associated with using the PAPNET.[116,119–121] In two of the studies[116,120] it was unclear whether machine variation had been controlled, but for the purpose of analysis it has been assumed that machine variation did not feature in the trial.

A strength of the Dutch study[116] over the other studies was that it compared conventional screening with PAPNET screening. The weighted kappa scores for interobserver variation for these screening types were calculated to be 0.72 (95% CI 0.68 to 0.75) and 0.69 (95% CI 0.66 to 0.73), respectively. Thus, there was no statistically significant difference between the ability of both systems to perform consistently.

The other three studies,[119–121] all by Mitchell and Medley, had kappa scores of 0.41, 0.56 and 0.62. Although the latter two values are standard kappa scores, the consistency in performance implied ranges from fair to good. Of some concern is that the worst agreement occurred in a slide-mix most typical of that likely to be encountered in practice.[120] It is also important to note that one of the studies by Mitchell and Medley involved the use of multiple observers. Insufficient data were available to calculate weighted kappa scores, so only a standard kappa score for multiple observers could be calculated ($\kappa = 0.56$).[121]

Machine variation

Two studies,[116,120] measured the reproducibility of the PAPNET, using system-generated codes as outcomes. In this set-up there would be no potential for observer variation. Unfortunately, it was not clear whether the same or a different machine had been used in both arms, so the type of machine variation assessed cannot be specified.

The level of agreement depends on the outcome used for comparison. The only outcomes that were comparable between the studies were the technical codes. The technical codes used in both instances were broadly similar, yet the level of agreement varied from poor to good. It is difficult to make

TABLE 42 *Reproducibility: observer variation in included studies*

Study	n^a	Intraobserver variation				Interobserver variation			
		New technology		Conventional		New technology		Conventional	
		κ_w (95% CI)	Agreement	κ_w (95% CI)	Agreement	κ_w (95% CI)	Agreement	κ_w (95% CI)	Agreement
Doornewaard *et al.*, 2000[116]									
Overall	1996					0.69 (0.66 to 0.73)	Good	0.72 (0.68 to 0.75)	Good
Observer A	110; 144	0.77[b] (0.68 to 0.87)[b]	Excellent	0.64[b] (0.49 to 0.79)[b]	Good				
Observer B	93; 53	0.69[b] (0.47 to 0.92)[b]	Good	0.74[b] (0.63 to 0.84)[b]	Good				
Observer C	69; 140	0.72[b] (0.62 to 0.82)[b]	Good	0.78[b] (0.65 to 0.90)[b]	Excellent				
Observer D	166; 129	0.72[b] (0.63 to 0.80)[b]	Good	0.59[b] (0.49 to 0.68)[b]	Fair				
Mitchell and Medley, 1998[119]	164					0.62[bc] (overall)	Good		
Mitchell and Medley, 1998[120]	2681					0.41[b] (0.32 to 0.50)	Fair		
Mitchell and Medley, 1998[121]	188					0.56[c]	Fair		

[a] Sample size analysed. Where values are split, e.g. 110; 144, the 110 refers to sample size of the new technology, and the 144 refers to the sample size of the conventional screen.
[b] Quoted values from article. Insufficient data to verify here.
[c] Standard kappa value, i.e. not weighted.

TABLE 43 *Reproducibility: machine variation in included studies*

Study	*n*	Machine variation					
		% air bubbles		Technical codes[a]		Decisions on slides	
		κ_w (95% CI)	Agreement	κ_w (95% CI)	Agreement	κ_w (95% CI)	Agreement
Doornewaard *et al.*, 1999[117]	196	0.59 (0.51 to 0.68)	Fair	0.65 (0.51 to 0.79)	Good	0.92 (0.88 to 0.96)	Excellent
Mitchell and Medley, 1998[120]	2690			0.34 (0.26 to 0.43)	Poor		

[a] Technical codes used in the two studies assessing inter/intramachine variation: Doormewaard *et al.*: 'No technical code'/'Too few cells'/'Artefacts'; Mitchell and Medley: 'No technical code'/'Insufficient cells'/'Machine difficulties'/'Artefacts, bubbles'.

TABLE 44 *Reproducibility: whole-system variation in included studies*

Study	*n*	Whole-system variation			
		New technology		Conventional	
		κ_w (95% CI)	Agreement	κ_w (95% CI)	Agreement
Jenny *et al.*, 1997[118]	516	0.34 (0.23 to 0.45)	Poor		

any definite assertions as to why the two differ. The following reasons are all possible:

- differing prevalence of disease: 1.7% in one study and 31.6% in the other (it is of concern that agreement is worst where prevalence is more typical of that likely to be encountered in practice)
- one study may be measuring intermachine variation and the other intramachine variation
- different slide preparation protocols, with one aiding unambiguous classification
- changes in the physical characteristics of the slides between successive submissions.

The high level of agreement for decisions on slides in the study by Doornewaard and colleagues[117] is unsurprising. This was based on simultaneous viewing of two monitors, each having the images from the alternative scan. Decisions were based on the content of the images and the opportunity for reviewer bias was large, as the reviewer was aware of the contents of both arms at the same time.

Whole system variation

Jenny and colleagues reported the only study that measured the reproducibility of the PAPNET system as a whole.[118] With such an approach neither the machine nor the observers used remain constant between the two arms of the study. As with all these types of study, constant input (the same slide set) is required. The problem with such an approach, however, is that it does not reveal where the main source of variation lies, and therefore whether it is controllable.

However, given a cohort representative of the local screening population (in terms of spectrum of disease), whole-system variation could be more of an indication of the type of agreement that would be expected in the routine use of the PAPNET. Unfortunately, in the study by Jenny and colleagues the disease spectrum was quite different to that which would be expected in practice; 516 (43%) abnormal (all biopsy confirmed) and 684 normal slides were used.

The weighted kappa was calculated to be 0.34 (95% CI 0.23 to 0.45), showing that the net effect of the different variations was to give poor reproducibility of results. The conventional screening system was not used as a comparison.

Conclusion on reproducibility

All studies encountered evaluated the reproducibility of PAPNET. Cumulatively, they help shatter the myth that automated image

analysis devices, although computer based, do not share some of the same failings of systems based wholly on human interpretation. The results for observer variation ranged from poor to fair-to-good agreement. The strongest study of those reviewed on observer agreement, by Doornewaard and colleagues,[116] demonstrated the intraobserver and interobserver variation of the PAPNET to be equivalent to that of conventional screening. Kappa scores tend to be difficult to compare across studies, but within studies, particularly when compared with a conventional screening arm, they do provide a useful metric. Two studies measured machine variation, using either technical codes or percentage of air bubbles reported as the parameters to compare. While Doornewaard and colleagues[117] reported fair to good agreement, Mitchell and Medley's results[120] were less impressive. In addition to being undesirable features in their own right, imperfections in the observer and machine reproducibility inevitably contribute to the imperfect test performance demonstrated for PAPNET in the previous part of this chapter.

For this reason, it is of concern that no studies were encountered which assessed reproducibility, either for the current version of AutoPap or for its predecessor. The way in which AutoPap works, being a one-stage rather than a two-stage process, suggests that machine variation may be less. However, this needs to be confirmed. Further, rigorous information on whole-system reproducibility also needs to be provided, particularly in the conditions and circumstances in which the system is likely to be applied. Obtaining such information is an urgent priority for further research.

Part 3: Health outcomes

Introduction

As discussed in earlier chapters, it was thought sensible to check that there were no studies assessing the impact of introducing automation on health outcomes. A priori, the reviewers believed the study design most likely to have been carried out that may furnish information of this type would be historically controlled studies (pre–post), although recognising that the validity of such studies would need to be carefully considered. Searches were thus amplified to ensure that such studies were not overlooked. However, studies with between-subjects study designs from other components of the review, particularly the test performance component, were also considered, and as indicated in Part 1 of this chapter, an RCT identified unexpectedly by this means is included.

Method

The search was principally that applied in the other parts, with an additional broad search being applied to the 1990–1997 period to identify historically controlled trials that may not have been captured by the search strategies used in past systematic reviews on this subject. The searches referred to in previous parts of the review, particularly the search for foreign-language studies from 1990 to 1997, would have been a subset of the search designed to identify historically controlled trials.

Since, the design of particular interest was the historically controlled trial, the inclusion criteria were designed accordingly. Within-subject study designs were not considered for the purposes of providing information on health outcomes. Included studies needed to satisfy the same criteria for general relevance as used in stage 1 of the criteria used in the test performance review. In addition, they needed to be a between-subjects design, particularly a historically controlled trial quantifying the impact of introducing the new technology into a particular screening setting, on a health outcome (death, survival, cancer incidence). Levels of preinvasive disease were not considered to be a health outcome, but are considered in the context of impact on process in Part 4 of the review.

The criteria for inclusion are set out below.

- The study type was a primary study.
- The study evaluated cervical cytology as a screening test.
- The study evaluated an automated or a semi-automated cervical screening system.
- The screening system was used in a primary screening, rescreening only, or primary screening with rescreening mode.
- The study evaluated the impact on a health outcome of the implementation of the new technology using a between-subjects study design.

Study quality was assessed using the framework suggested by the Cochrane Collaboration.[49] Analysis relied on qualitative rather than quantitative summary of results of included studies.

Results: number of studies included

The results of applying the inclusion criteria are shown in *Table 45*. Two studies were included.[109,122]

TABLE 45 *Number of studies included and excluded in review of health outcomes*

	Historically controlled trials 1990–1997	Update search 1998–2000
Unduplicated articles	66	163
Inclusion/exclusion:		
Excluded	66	160
Remaining	0	2[a]
Included	0	2[a]

[a] Includes RCT originally identified in the course of part I of review.

TABLE 46 *Health outcomes: scope of included studies*

Study	Design	Intervention			Comparator				Screening modality		
		PAPNET	AutoPap	Auto Cyte SCREEN	Pap test	Pap +HPV	LBC	LBC +HPV	Primary	Rescreen	Both
Kok *et al.*, 2000[109]	RCT	✓			✓				✓		
Wertlake, 1999[122]	Pre–post		✓		✓					✓[a]	

[a] The QA step that AutoPap replaces is that current in the USA, rescreening of a 10% random sample of slides originally designated as 'within normal limits', in contrast to the UK system where there is rapid rescreening of all slides not originally designated as abnormal.

Results: details, quality and results of included studies

What was investigated?

As summarised in *Table 46*, one included study evaluated PAPNET[109] and the other the AutoPap 300 QC rescreening device[122] (one of the predecessors of the currently available device). The conventional Pap test was the technology used as the comparator in both cases; assessment of the impact of automation on systems incorporating LBC and HPV triage has not been done. The study by Kok and colleagues[109] on PAPNET investigated the impact of automation used in the primary screening step and the study by Wertlake the effect of AutoPap used in the quality assurance process. Thus, the health outcomes of AutoPap do not appear to have been assessed when used in the primary screening step, the way that the currently marketed version of AutoPap is designed to be used. The value of the assessment of AutoPap 300 QC is lessened in the context of the UK by the fact that the quality assurance step in the USA, where the study was conducted, is so different from that used in the NHSCSP.

Details of the intervention and comparator

These are presented in *Table 47* and considered subdivided by the screening step that the automation replaces or amplifies.

Primary screening mode

In the study by Kok and colleagues[109] patients' slides were randomly divided between PAPNET and the normal screening arm. The former involved central creation of data tapes, viewing the 128 tiles for each smear, and screeners making a decision on whether the slide was normal or abnormal based on this. Normal slides were archived and abnormal slides subject to review by a cytopathologist. In the normal screening arm, slides were manually screened by light microscopy and then if identified as normal archived, or if abnormal reported by a cytopathologist. Unlike the screening systems in the USA and the UK, the screening system in The Netherlands does not involve a quality assurance step.

Rescreening mode

As the pre–post study was conducted in the USA,[122] the manual control period used rescreening of a random sample of 10% of negatives as part of its quality control. As already highlighted, this is different practice to that in the UK. In the AutoPap period, slides to be rescreened manually in the quality assurance step were not considered on a 10% random basis, but were considered if they had not already been

TABLE 47 *Health outcomes: details of the intervention and comparator*

Study	Intervention description	Base comparator description
Kok *et al.*, 2000[109]	PAPNET scans some of slides, and images reviewed on monitor. Abnormalities identified from images, given targeted microscopic review.[a] All positive slides referred to a cytopathologist.[a] Negative slides archived[a]	Primary microscopic screening of all slides not screened by PAPNET. Abnormal slides referred to a cytopathologist,[a] with negative slides archived[a]
Wertlake, 1999[122]	All slides were manually screened. All negative slides were subject to scanning by the AutoPap 300 QC system, set at the 10% review rate (12.7% were selected) High-risk slides were excluded from AutoPap review. All review, QC review and process review received manual rescreening High-risk slides, including those with an abnormal smear or biopsy history, postmenopausal bleeding or history of sexually transmitted disease, were rescreened manually Abnormal slides were referred to a pathologist	All slides were manually screened. Of the negative slides a 10% (10.2%) random sample received manual rescreening High-risk slides were excluded from the 10% random sample review Manually rescreened slides that were found to be abnormal were referred to the pathologist for the final diagnosis

[a] Additional information obtained directly from the corresponding author, ME Boon.

identified as abnormal in the primary screening and had

- a high probability of abnormality ('QC review')
- failed to be processed by the device ('Process review')
- scanty material which made a decision difficult ('review').

The study excluded high-risk slides from the standard protocol.

It is noticeable that in the AutoPap study which set '10% review rates' this was not precisely achieved. This is explained by the fact that the AutoPap review rate is derived from a sample of the local population that may or may not be representative of the total population being screened by the device.

Combined primary screening and quality control
There were no between-subject studies assessing the impact of automation used to amplify both the primary and quality assurance steps.

Study population

Two issues need to be addressed. First, the population needs to be reasonably typical of that in which automation is likely to be used; and second, the characteristics of the population in the treatment and control arms or periods need to be similar so that any changes in outcome can be correctly attributed to the difference in screening programme.

Are populations typical?
As shown in *Table 48*, both studies consider population cohorts, and in this respect are likely to be typical of the screened population in the countries in question, The Netherlands and the USA. Unfortunately, because there is no population-based screening programme in the USA, the source population is likely to be less typical of that in the UK than the Dutch study. As a corollary, the percentage of negative smears (approximately 91%) is somewhat higher than that in the NHSCSP (approximately 82%).

Are populations equivalent?
The baseline characteristics are not clearly presented for either included study.

In the randomised trial by Kok and colleagues,[109] 109,104 slides were assigned at random to 'conventional microscopic screening and quality control' and 245,527 to PAPNET. One thus has to rely on adequacy of randomisation to ensure reasonable balance in characteristics at baseline, and no details on randomisation method are provided in the paper.

Wertlake[122] does provide information on age as a justification for the equivalence of the populations in the pre and post periods. Although the histograms look similar and the median for the pre and post periods is quoted as being 34.9 years and 33.8 years, respectively, it is debatable whether this is adequate proof of equivalence.

TABLE 48 *Health outcomes: study design and nature of population*

Study	Design and location	*n*	Comments
Kok *et al.*, 2000[109]	RCT, Leiden, The Netherlands	354,631 All smears from 1 January 1992 to 30 December 1997	System: routine 5-yearly smears offered to all women aged 30–60. 'Interval smears' can be requested by GP (patient initiated; based on assessment of greater risk, especially lifestyle or clinical findings) 194,258 routine smears; 160,373 interval smears
Wertlake, 1999[122]	Pre–post, Tarzana, California, USA	1,141,913 Unselected, consecutively accessioned cases from October 1995 to April 1998	System: not a population-based screening programme

TABLE 49 *Health outcomes: levels of disease as ascertained in the study by Wertlake*[122]

Location: Tarzana, California, USA **Population: Primary screening**	**Pre-auto**	**Post-auto**	**Difference (95% CI)**
Period	October 1995 to January 1997 (16 months)	February 1997 to March 1998 (14 months)	
Total	591,837	550,076	
Negative	91.41%	91.04%	–0.37 (–0.48 to –0.27)
Unsatisfactory	0.58%	0.71%	+0.12 (+0.10 to +0.15)
ASCUS	4.51%	5.58%	+1.07 (+0.99 to +1.15)
AGUS	1.06%	0.35%	–0.71 (–0.74 to –0.68)
LSIL	1.83%	1.87%	+0.042 (–0.008 to +0.09)
HSIL	0.594%	0.439%	–0.155 (–0.18 to –0.13)
Invasive	0.014%	0.009%	–0.005

* Significance difference between pre and post periods.

Restriction of the analysis to just one characteristic is a concern. Although not intended for the purpose of comparing baseline equivalence, there are data to compare the level of disease after primary screening, since this part of the process should have remained constant irrespective of the quality control policy. These are shown in *Table 49*. There are significant differences between virtually all the subclassifications of disease between the pre and post periods. This either challenges the validity of comparing outcomes in the pre and post periods, or suggests that changes apparently restricted to the quality assurance step actually affected primary screening too. The authors of the study present data claiming that increased feedback and training arising from the improved quality assurance provided by use of AutoPap were indeed responsible for an improvement in primary screening.

The size of the populations involved and the timescale suggest that change in the nature of the populations is unlikely to be the explanation for the change in spectrum of disease detected at the primary screening stage, but this does not preclude the possibility that the nature of women presenting for screening did not change. Even if it is felt reasonable to assume that the population was equivalent in the pre and post periods, it must be acknowledged that the effect being measured is probably not just that of automation with AutoPap, but the associated effect of audit, feedback and training.

Study outcomes

Neither of the included studies compared mortality. The difficulty in doing so is clearly indicated by data in the study by Kok and colleagues,[109] in which out of a combined screened

TABLE 50 *Health outcomes: number of cancers prevented in included studies*

	Manual	Automation	Decrease in cancers detected
Kok *et al.*, 2000[109]			
RCT comparing one stage manual system with automation using PAPNET			
No. of cancers detected	19	52	–33
No. of slides	109,104	245,527	
% of all smears	0.0174	0.0212	–0.0038
95% CI	0.011 to 0.0272	0.0162 to 0.0278	ns
Wertlake, 1999[122]			
Pre–post study comparing manual system using 10% random rescreening for QA, with manual system + AutoPap replacing QA step			
No. of cancers detected	83	49	34
No. of slides	591,837	550,076	
% of all smears	0.014	0.009	0.005
95% CI	0.011 to 0.017	0.007 to 0.012	ns

population (manual or automation) probably considerably in excess of 250,000 women, there were seven deaths from invasive squamous cancer of the uterine cervix, with an approximate median period of follow-up of 3 years.

Both included studies instead assessed the impact of automation on the number of invasive cancers detected. In the study by Kok and colleagues the focus was just on squamous cancers, which were ascertained via computerised laboratory files independently from the analysis of the cervical smears. Additional follow-up information on microinvasive [International Federation of Gynaecology and Obstetrics (FIGO) stage Ia] and fully invasive cancers (FIGO stage IB or higher) was obtained from the Dutch Automated Archive of Pathology Laboratories (PALGA), amplified with telephone follow-up. Such follow-up information was available in 69 out of 71 cases (97%).

In the study by Wertlake[122] cancers detected do not appear to have been restricted to squamous cancers, but also included adenocarcinomas. Unfortunately, the cancers are not divided by stage. Concerning ascertainment and verification of cancer cases, it is unclear how independent this was of obtaining the original cytological result. It appears that some additional histological verification was obtained on some high-grade slides; it is unclear, however, how incident cancer where cytology was negative would have been identified in this study.

Study quality

The foregoing alone suggests major concerns about the internal validity of both included studies. Selection bias presents an important threat to validity in both studies. In the study by Kok and colleagues[109] this arises because of limited information about the process of randomisation; in the study by Wertlake[122] the pre–post design intrinsically makes the results more likely to be susceptible. Little reassurance is provided that the characteristics of the population in the pre (manual) period do not have important differences from those in the post (automation) period.

Beyond selection bias, the study by Kok and colleagues appears to be relatively free of biases that may affect internal validity. In contrast, the study by Wertlake appears to be highly likely to be susceptible to performance and detection biases, and there is no information to assess whether differential loss to follow-up in the pre and post periods may also have been a source of bias.

Results of included studies

Only information on the impact of automation on the number of invasive cancers is available. The results are summarised in *Table 50*.

It should again be noted that the generalisability of the included study results to the current system in the UK is limited. Further, there are important threats to internal validity. With these important provisos, one included study[109] showed automation to be associated with an increase in numbers of cancers, the other a decrease.[122] Unsurprisingly, neither change was statistically significant, suggesting that the pattern across the two studies is most likely to be due to chance. Even if there were truly an effect of automation on the number of invasive cancers, it is highly

unlikely that even the two studies in combination would have sufficient power to detect a difference. Further, it is unclear whether the studies have considered the outcome over a long enough period for it to be plausible that any observed changes are attributable to automation. If such an effect is present it is likely that improved outcomes would be achieved by small improvements in the accurate detection of all levels of preinvasive and invasive disease, including lower grades where the avoidance of cancer as the health outcome would only occur many years in the future. Thus, the maximum benefit would only be likely to occur several years after the introduction of automation. Whether the approximate median 3-year follow-up in the study by Kok and colleagues[109] is sufficient is debatable; the 1-year period for the Wertlake study[122] is highly unlikely to be sufficient. Predicting when the maximum change in health outcome may occur after the introduction of a change in policy is something with which the DES model being developed as part of this project (see Chapter 9) may be able to help.

An interesting perspective concerning interpretation of number of cancers is also afforded by additional data presented in the study by Kok and colleagues.[109] Information on stage is provided for 69 of the 71 cancer cases on which there is follow-up (although not broken down by intervention). Twenty-eight were stage IA, 30 stage IB, three stage IIA, four stage IIB, one stage III and three stage IV. Three deaths occurred in the IB group, one in the IIB group and three in stage IV. The universal survival of stage IA (microinvasive cancer) suggests that with current treatments, an unchanged number of cancers, but a shift towards microinvasive cancer may be considered as much a 'success' as an overall reduction in cancer. The corollary of this is that in assessing the impact of automation on the health outcome of cancer incidence, breakdown by stage of cancer is important. Neither included study provided that breakdown for the manual arm or period compared with the automation arm or period.

Conclusions on health outcomes

On the grounds of generalisability alone, it is debatable whether the included studies add much to the current assessment of the impact likely to occur were automation to be introduced. What is highlighted, however, is that between-subjects designs, RCTs or pre–post studies, if appropriately conducted, have the potential to improve greatly the future assessment of impact. These designs should be included among further research on this topic.

Part 4: Process

Introduction

An important issue with which this health technology assessment project has attempted to deal is that the introduction of automation may be justifiable if test performance were equivalent to existing automated systems, but the time and effort required to process slides to screen a given population were reduced. For this reason a specific attempt was made to review systematically the published evidence on the impact on process of introducing automation to existing cervical screening programmes. Such information is often available from the manufacturers. This lacks independence and authenticity in many people's eyes. The purpose of this part of the review was thus to concentrate on independent assessments, most likely to reflect accurately the impact on actual practice, to see whether manufacturers' claims about their products could be substantiated.

Method

All studies passing step 1 of any search conducted for each of the preceding three parts of the review were further screened for information on the impact of an automated device on the screening process. Specifically, information on the following was sought:

- technical reliability of the technology, gauged in terms of breakdowns, need for maintenance and the numbers of slides rejected by the machines (and so needing to be processed manually)
- the average turnaround time or process time associated with using a device
- changes in the system design required for successful implementation of the new technology; these may be in terms of changes in the number of staff employed, changes to the type of work or the implementation of a different set of internal controls.

There were no further inclusion/exclusion criteria other than that the study provided information on the outcomes stated above.

The impact on the yield of slides at particular cytological gradings (i.e. ASCUS, LSIL, HSIL or borderline, mild, moderate and severe dyskariosis) has been addressed frequently in the studies encountered, using a variety of designs. For instance, the historically controlled study by Wertlake discussed in the previous section[122] and two other pre–post studies considered for inclusion in this section[123,124] superficially show that the use of the AutoPap 300 QA, to replace a

quality assurance step based on manual rescreening of 10% of slides selected at random, generally improves the yield of abnormalities at all levels (ASCUS, LSIL and HSIL), although only the first of these in a statistically significant manner. Unfortunately, viewed as a health outcome there must be debate about whether increased detection levels are good or bad, particularly in the absence of information about whether these are true or false positives, and uncertainty about the untreated fate of women with abnormalities at any of these grades, particularly the lower ones. Thus, studies purely giving information on changes in the yield of cytological diagnoses at particular levels associated with different methods of slide analysis were rejected as studies providing useful information on the impact on health outcome.

The impact on yield of slides at particular cytological grades may have been of value concerning the impact on process. This is because changes in cytological grade not only have possible health outcome implications (i.e. they are possibly more likely to identify preinvasive cancer at a stage where it can be successfully treated compared with the adverse effects of treatment for disease that would not have progressed to invasive disease untreated), but also have consequences for the screening programme. Thus, changes in the number of borderline or low-grade lesions increase the need for repeat smears, and hence may increase the average number of slides needing to be processed per woman enrolled in a screening programme. Similarly, changes in the number of high-grade lesions, irrespective of whether they are true or false positives, will increase the number of colposcopies required. These represent costs in addition to the new automation system, which need to be justified relative to any promised benefit. It is also worth noting that these costs are not just financial. In a situation where many cervical screening programmes are struggling to cope with processing existing slide volumes, introducing a system that increases the numbers of borderline and low-grade smears identified could have particularly important consequences. Unfortunately, what is required to obtain reliable information on such a consequence requires studies that refer to individuals rather than slides. Unfortunately, no such studies were encountered, and so studies providing just information on changes in the yield of slides associated with the introduction of automation have been effectively excluded for the purposes of considering the possible impact on both health outcomes and process.

Results for technical reliability: details and results of included studies

Table 51 shows all the studies that contributed information on the technical reliability in the day-to-day running of the automated cervical screening system reported. The technical reliability here was confined to mean the number of slides rejected by the system, as this type of slide has an impact on the efficiency of workflow of the new system.

The results from the studies have been pooled to give an average expected rejection rate that would be encountered in practice. The problem with this approach is that there is no consistent definition used for a rejected slide by the authors of the studies analysed. This was particularly the case with the PAPNET studies which, unlike trials on the AutoPap, did not have a clear category of 'process review', that is, those slides that the device is unable to process. Note that broken slides are not considered in the 'process review' category by the AutoPap, and for the analysis of both devices have been excluded from any pooled estimates. In contrast, the PAPNET studies tended to be less explicit about the 'unprocessed slides', with technical error often being quoted.

Slides with the 'cornflake' technical error would often be restained, have the coverslip reapplied and then be rescanned. Although this would involve extra processing, for the purpose of consistent analysis, slides that were successfully scanned after an initial failure were not considered rejects.

The pooled data on the PAPNET amounted to 64,277 slides being submitted for scanning with 1409 rejects, giving a rejection rate of 2.2% (95% CI 1.4 to 3.0%). Two studies contributed over 40,000 of the slides. In the PRISMATIC trial,[112] the only UK trial analysed (n = 21,700), the rejection rate was 2.0%, and in the Mitchell and Medley study[120] (n = 20,000), the rejection rate was 2.9%.

For the AutoPap, 562,793 slides were submitted for processing and 43,560 slides were rejected, giving a rejection rate of 7.7% (95% CI 7.5 to 8.0%). This average raises a number of points. The first is that the AutoPap has been treated generically and not subclassified. For the purpose of rejection rates, it was assumed that the sort of slides rejected by the rescreening model would be no different to those rejected by the primary screening model, that is, rejection is not in any way related to the presence or absence of

TABLE 51 *Rejection rates of the automated cervical screening systems*

PAPNET				AutoPap			
Study	**All slides**	**Rejected slides (%)**		**Study**	**All slides**	**Rejected slides (%)**	
PRISMATIC, 1999[112]	21,700	428	1.97	Bibbo and Hawthorne, 1999[125]	5,120	394	7.70
Chhieng *et al.*, 2000[126]	108	3	2.78	Fetterman *et al.*, 1999[123]	35,143	627	1.78
Doornewaard *et al.*, 1999[106]	6,343	161	2.54	Huang *et al.*, 1999[127]	400	31	7.75
Duggan, 2000[128]	2,200	36	1.64	Lee *et al.*, 1998[129]	683	151	22.11
Losell and Dejmek, 1999[130]	1,000	5	0.50	Marshall *et al.*, 1999[124]	31,240	2,777	8.89
Mitchell and Medley, 1998[120]	20,000	581	2.91	Wertlake, 1999[122]	463,836	38,617	8.33
O'Leary *et al.*, 1998[72]	5,478	128	2.34	Wilbur *et al.*, 1998[114]	26,171	963	3.68
Sherman *et al.*, 1998[107]	7,323	61	0.83				
Sturgis *et al.*, 1998[131]	75	6	8.00				
Totals	64,227	1,409	2.19	Totals	562,593	43,560	7.74

abnormality. As in previous parts of the effectiveness review, it should be noted that none of these studies relates to the current commercially available AutoPap device, designed for use in the primary screening step. It is of interest to note that the rejection rate obtained is compatible with the figure provided in a recent product insert for the AutoPap GS, of 7.9%.[35]

Finally, it should be noted that all studies, irrespective of whether they have been on AutoPap or PAPNET, have involved the processing of conventional slides. Since the advent of automated systems there has been a belief that the conventional Pap smear preparation was a significant part of the problem and that monolayer preparations (i.e. LBC) would greatly improve the chances of success, and so reduce rejection rates and slides classed as 'inadequate'. No studies from the search of the published literature were located that gave information on rejection rates of automated image analysis systems using a liquid-based preparation. It is thus impossible to corroborate the estimates from the AutoPap product insert claiming that the rejection rates for AutoCyte PREP slides (LBC system) fed into the AutoPap primary screening system were 418/6860 (6.1%) and 141/1665 (8.5%).[35] This may not be important, given that the figures are only slightly improved relative to the rejection rates for conventional smears.

Results for processing times: details and results of included studies

Information on the impact on processing times attributable to using automated systems was generally very poor. It appears that the major impetus of research into automated systems has been to evaluate their test performance metrics compared with the different manual systems around the world. This does not appear to have been accompanied by accurate assessment of processing times, although an important argument in favour of introducing automation requires not just demonstration of equivalent test performance, but also a clearly demonstrated reduction in average processing times, ideally taking into account the possibility that increases in the number of indeterminate and low-grade abnormal results may mean that a greater number of women would require repeat smears.

Table 52 provides brief details and results of the best information on the impact on processing times encountered. Few if any studies appeared to set out deliberately to assess such impact accurately, most results being reported as snippets in discussion sections, and frequently in passing. There were no published data on any version of AutoPap.

The main problem with nearly all the studies was the lack of clear definition of the part of the process whose duration was being determined, particularly the start and end-points. Without such clarity comparison was difficult.

The one study that did address these issues was the PRISMATIC trial,[112] and this consequently gives the best data, albeit restricted to the use of PAPNET. The screening process was broken down into stages that are presented in *Table 53*. The results appear to have been calculated from data collected for all 20,008 slides included in the study. It can be seen that the key stage for saving time was the primary screening step. Negative or inadequate slides in the conventional arm

TABLE 52 *Processing times of automated systems*

Study	Device	Mean time (minutes)			Comments
		Auto	Manual	Ratio	
PRISMATIC, 1999[112]	PAPNET	3.9	10.4	2.7	Total mean time for process. For breakdown of individual stage times see *Table 53*
Mitchell and Medley, 1998[120]	PAPNET	2.4			Average time to review a case (image with or without microscopy of slides)
O'Leary *et al.*, 1998[72]	PAPNET			3	Manual screen = 3 × PAPNET
					Saving = 2 minutes per slide (approx. two-thirds of cases not requiring manual rescreen)
Troni *et al.*, 2000[76]	PAPNET	3.8	4.5	1.2	Two experienced cytologists, trained to read PAPNET by the manufacturer, examined 1000 routine slides seeded with 81 false negatives. They assessed slides first in the usual manner, then using PAPNET. The two periods of examination were separated by 20 days
		4	4.8	1.2	
Veneti et al., 1999[113]	PAPNET	1 (approx)	5 (approx)	5 (approx)	Claims that PAPNET is five times faster than manual: 5 minutes vs 1 minute

TABLE 53 *Processing times for automated systems: details for the results of PAPNET in the PRISMATIC trial[112]*

	Conventional times (minutes per slide)		PAPNET (minutes per slide)	
	Mean	SD	Mean	SD
Primary screen (n = 20,008)	7.4	1.2	1.4	1.0
Rapid review of negatives/inadequates (n = 17,635)	3.0	1.9		
Abbreviated screen of negatives/inadequates (n = 18,053)			2.0	0.9
Check and report each abnormal smear found on primary or rapid review/abbreviated screen (Manual n = 2373 + 310; PAPNET n = 1955 + unknown[a])	5.4	1.1	5.1	1.9
Total time for screening and reporting	10.4	2.4	3.9	1.8
Difference (95% CI)	6.5 (6.54 to 6.46)			

[a] Number of slides initially identified by PAPNET as negative/abnormal, subsequently identified as abnormal on abbreviated screen, not clearly stated.

prompted rapid review, which was executed differently to the abbreviated screen in the PAPNET arm, which was seemingly quicker than the normal NHSCSP rapid review from the data presented. The difference in checking and reporting abnormal slides was marginal.

In summary, from this study it may be inferred that a screening process starting with the primary screening of the slides and ending with their reporting would have been between two and a half and three times faster with a PAPNET system than the existing conventional screening system in the UK. The study does not appear to consider the additional administration time needed to obtain the data tapes before the start of the screening process. The possibility of bias also needs to be considered, resulting from an inability to blind whether the slide was being assessed as part of the control or intervention (PAPNET) arm. The possibility that a greater number of repeat slides might have to be taken when automation was used does not seem to be an issue as far as the PRISMATIC study is concerned, as fewer slides were classified as borderline or mild in the automated relative to the manual arm (770+690 versus 1021+836). Despite these provisos, the data from PRISMATIC, supported by data from other studies, suggest that PAPNET used in a primary screening setting equivalent to that in the UK can reduce the average time taken to process a slide. There is uncertainty about the magnitude of

the improvement, but it seems clear from the available data that improvement is clinically important and statistically significant.

The key issue arising is thus whether there are any data to suggest that a similar level of benefit may be obtained from the only currently commercially available automated image analysis device, the AutoPap GS. Claims from the manufacturer on costs (see Chapter 8) suggest a four-fold increase in slide processing capacity, an improvement somewhat greater than that implied by the PRISMATIC trial for PAPNET. In the absence of any published data, the present authors cannot verify or refute this.

Of all the areas requiring urgent additional research, independent confirmation of the ability of the currently commercially available version of AutoPap to produce substantial reductions in mean total slide processing time is probably the greatest priority. Ideally, this should be combined with confirmation that sensitivity and specificity are equivalent to manual screening and that the need for repeat slides arising from increased cytological diagnoses of borderline and mild is unlikely to offset the benefit in terms of workload arising from reduced mean processing times. Replication of the PRISMATIC trial would achieve this. A well-conducted RCT or pre–post study would also be of value, although the ability of this design to produce valid information on test performance would be limited. Such between-subjects designs would, however, provide the basis for a longer term assessment of health impact.

Irrespective of the study design, the key issue is that the times taken for all stages of the process are recorded, including extra preparation times, cleaning and maintenance. Only when times for all processes are included in the analysis will it be clear whether over a given period a laboratory will be able to process more slides with the new system.

System changes associated with introducing automation

From the included 42 papers that passed step 1, and speaking to authors and investigators of automated screening systems, the following points were raised as potentially needing to be considered in assessing the overall effectiveness of introducing automation. The majority of the points relate to the AutoPap screening system, as these studies tended to give more information on the systems and process. Quantification of the potential effects highlighted was rarely attempted, and so the issues are merely listed.

- **Staining**: for the UK the staining protocol will probably need changing, as the AutoPap requires heavier staining of the nuclei with haematoxylin. A Laboratory Process Compatibility Assessment is carried out where staining scores are assigned by the machine to the slides. Trials of different staining protocols are performed on batches of slides until the batch staining score lies within the range of the system. Future slides are then stained according to the successful protocol. Unfortunately, changes in staining imply a change in appearance of the slides and laboratory staff will need to adapt.
- **System tolerance**: automated screening systems to date have not shown the same flexibility as their human counterparts. As indicated in the last point, the AutoPap can only tolerate a level of staining that lies within a predefined range, and other stages of preparation also have low margins for error. In terms of preparation a slide may be rejected owing to aberrant coverslips, barcoding or cornflaking, among others. These slides may need to go through the preparation process again before being accepted. The effect is that more time could well be spent on vigilant preparation.
- **Air conditioning**: those who have used the AutoPap have described instances where it has failed owing to overheating and that even during cold weather windows had to be left open to cool it down. In the AutoPap product insert it lists that a dedicated 20-A supply is required, at 100–200 V. This equates to power consumption between 2 and 4 kW. Some of this power will generate heat and this could be a large portion of the electricity consumption: 2 kW of heat would not be unreasonable. Thus, unless there is adequate ventilation when operating, the device could rapidly reach temperatures that cause it to break down. The product insert reports the operating temperature range to be 10–30°C.
- **Change of duties**: additional staff may be required for preparing, loading and maintaining the new system. Fetterman and colleagues[123] describe the need for an additional 3.5 laboratory assistants to operate four AutoPap 300 QC machines. Their "time was spent cleaning slides before and sorting after processing". They did believe, however, that the newer devices (study reported January 1999) were more tolerant of dust and dirt, and would require only one or two assistants for four machines. However, whatever the number of people required, the nature of the job is repetitive and tedious. Thus, as for the job of screening manually, recruitment and retention of staff may be a problem.

Conclusions on impact on process

Reviewed research evidence suggests that the number of conventional slides that automated slides image analysis systems cannot process is small, but not negligible. The rejection rates for PAPNET appear to have been lower, at 2.5%, and the rejection rates for AutoPap were approximately 7.5%. This means that for every 100,000 slides processed, 7500 will need to be processed manually, as in the current system. Debatably, the newer versions of AutoPap, using LBC slides rather than conventional Pap smears, may bring about an improvement in this rejection rate, but the manufacturer's own literature does not appear to be making major claims in this respect.

Rigorous assessments of the impact of introducing automation on mean processing times were sparse. Only one study on PAPNET was convincing, this suggesting that PAPNET used in the primary screening step in circumstances applicable to the NHSCSP was two-and-a-half to three times faster than the existing manual programme. It is debatable whether the size of benefit is quite as large as measured in this study, but what is clear is that the effect is substantial. No studies confirmed whether a similar impact might be expected with AutoPap, particularly the currently commercially available version. Obtaining such independent confirmation is hence probably the single greatest priority for further research.

The implementation of a new technology will impose changes on the organisation and system used. The AutoPap requires modification to the staining protocol, which will also affect the manual screening, as well as having low tolerance to aberrant slides. The increased preparation and the sorting of processed slides may require the employment of additional staff. Certainly a dedicated area with good ventilation is required for its efficient working.

Overall conclusion from all parts of systematic review of effectiveness

Taking all parts of the systematic review together, the overwhelming conclusion is that there continues to be inadequate research evidence on which to base a decision on whether automation should be implemented. Although this review has identified new research information on test performance not previously included in past systematic reviews and health technology assessments, and research information on reproducibility, impact on health outcomes and impact on process not previously searched for, the fundamental problem is that there is insufficient research to support implementation of an intervention as complex as automated slide analysis of cervical smears. It should also be noted that no information of any sort on patient acceptability or impact on patient quality of life of automated image analysis devices (two issues specifically mentioned in the commissioning brief for this project) was encountered in the course of the searches for effectiveness or other reviews contributing to this report.

Concerning past research, the only currently commercially available automated device (AutoPap GS System) appears to be substantially different from the predecessors and alternative types of automation evaluated. Further, past research uses as comparators manual systems not incorporating innovations such as LBC which are already under active consideration. Thus, past research is barely generalisable to the current situation and indicates strongly that there is an urgent need for further research on test performance, reproducibility, impact on health outcomes and process. Arguably, the importance of systematically reviewing the past research on this topic has been, first, to confirm that the trade-off between potential benefits and disbenefits has not been made sufficiently for it to be safe to assume that a device offering technical advantages over its predecessors will inevitably deliver net benefit. Second, if the need for further research is accepted the systematic review of past research provides clear lessons on the scope of research needing to be pursued and highlights shortcomings that need to be avoided in any future research.

Probably the single most urgent priority for further effectiveness research is independent confirmation that the currently commercially available image analysis device (AutoPap GS System) can achieve reductions in mean slide processing time similar to those achieved by PAPNET in the PRISMATIC trial.[112]

Chapter 7

Likelihood of publication bias in reviews of research on the effects of automation

Summary of key points

The main objective was to determine the risk of publication bias in the field of automated cervical screening.

A substantive methodology was developed and applied to the UK. Two nationwide questionnaire surveys of 215 research ethics committees and 212 cytology laboratories were carried out. In addition authors, British Society of Cervical Cytology (BSCC) members, the NHSCSP, manufacturers and regional Cervical Screening Quality Assurance Reference Centres were contacted. Online grey literature databases, including the NRR and Medical Editors Trial Amnesty (META), were also searched.

The response rates to the two surveys were 57% for the laboratory contacts and 59% for the ethics committees.

There were 13 verified unpublished studies. Although some of these were in-house pieces of research, most appeared to be major pieces of work in which substantial investigator time and effort had been invested. Most of the verified unpublished studies would almost certainly have been excluded in the systematic review of test performance. The authors of this report did not assess, and probably could not have assessed, whether they could have contributed to other components of the systematic review of clinical effectiveness.

There was uncertainty about whether two of the verified unpublished studies would definitely have been excluded. One of these is particularly important as it claims to assess the test performance of the currently available version of AutoPap in the context of the NHSCSP. It is imperative that this study, the 'Three Centre Trial', is fully published.

Both the volume and the nature of the unpublished studies identified in the UK suggest that it is plausible that major pieces of relevant research remain unpublished and unrepresented in this and past systematic reviews and health technology assessments. Unpublished literature in other countries should ideally be searched for, to the same level of scrutiny as has been undertaken in this report.

The experience of attempting to identify unpublished UK studies on automation suggests the need for new methodological research revisiting the concept of publication bias in the context of rapidly evolving options for disseminating the results of research. Guidance on when and how to conduct detailed searches for unpublished literature as part of attempts to address the threat of publication bias is also required.

Introduction

A priori, a concern was identified that unpublished research literature on the effects and effectiveness of automated cervical screening technologies might be particularly prevalent, and that reviews on effectiveness might be susceptible to publication bias as a result. Thus, the objective was to answer the question, 'With respect to reviews of evidence on the effects of automation, what risk to the validity of the overall conclusions is posed by unpublished research?'

A method was developed to locate unpublished studies that may have been omitted in the search for published literature. The method would ideally have been applied worldwide, but owing to time constraints was confined to the UK. It was developed in the context of a wider consideration, outlined in Appendix 9, of what one might want to achieve, with respect to ensuring the representativeness of included studies in a systematic review.

The chapter is divided into three parts. The first and main part deals with the description, application and results of the methods used to locate unpublished studies of automated cervical screening systems in the UK. The second part considers whether and how the method could be

generalised and extended to other countries, as well as other fields. The third part reflects the experience of attempting to identify unpublished literature on the established concept of publication bias and how it is dealt with.

Definition of published and unpublished studies in this health technology assessment

What constitutes a published and an unpublished study is dependent on the definition used. The definitions used in this project are reiterated below.

Definition of the set of published articles

The set of published studies is defined as consisting of all articles that may be retrieved from the following sources:

- MEDLINE, EMBASE, the Cochrane Library, HEALTHSTAR, CINAHL, CancerLit and other major bibliographic databases
- any recognised specialist journal not indexed in the above databases; a recognised specialist journal may be defined as a journal that, by consensus agreement of specialists in a field, would be a journal that they would consult for further information in that field
- publications from any health technology assessment body or equivalent on a worldwide basis.

Those articles that may only be found in the recognised grey literature databases, such as SIGLE, are excluded.

Definition of a published study

A published article or document is defined as one that may be found in the set of published studies.

Definition of the set of unpublished studies

The set of unpublished studies on automated cervical screening comprises all articles in the subject not found in the set of published studies.

Definition of an unpublished study

An unpublished article or document is defined as one that may be found in the set of unpublished studies.

Locating unpublished literature on automated cervical screening

Methods

The method developed was confined to the locating studies of three automated cervical screening systems over the period from January 1990 to May 2001. The three target devices of interest were:

- AutoPap (Neopath, now TriPath)
- PAPNET (NSI)
- AutoCyte SCREEN (AutoCyte, now TriPath).

Over the past decade these devices have been the most widely available technologies in automated cervical screening. The first two received FDA approval. Other technologies may have existed, but did not share the same success in dissemination across the global market. As a result these three devices are the ones that would have received the greatest scrutiny and testing, thereby being open to a greater potential for publication bias.

Baseline model of data sources

To find unpublished studies a model of the possible data sources was developed with the aim of systematically trawling through each to minimise the chances of missing any valuable studies. It is immediately apparent from *Figure 8* that these sources necessarily overlap as no individual source may be confidently considered to cover the whole field.

Description of the data sources used in the model

From *Figure 8* it can be seen that the different data sources of automated cervical screening in the UK were divided conveniently into those located centrally (i.e. no obvious division across the UK) and regionally (where a partition existed in which, for a particular data source, the UK could be divided into autonomous regions).

The different sources and how they were interrogated are expanded upon below.

Central data sources

- **Online sources**: the NRR at the Department of Health website and META (on the Cochrane Library) were searched for evidence of studies on the above devices.
- **Specialist bodies**: members of the BSCC were contacted for knowledge of conference proceedings or trials that had taken place in the UK. The NHSCSP was contacted and provided a list of all of the cervical cytology laboratories in England and Wales. Laboratories in Scotland were identified by contacting the Director of Cervical Screening Quality Assurance in Scotland.
- **Manufacturers**: there was only one commercially operational company covering the above three devices, TriPath. Offices in the USA

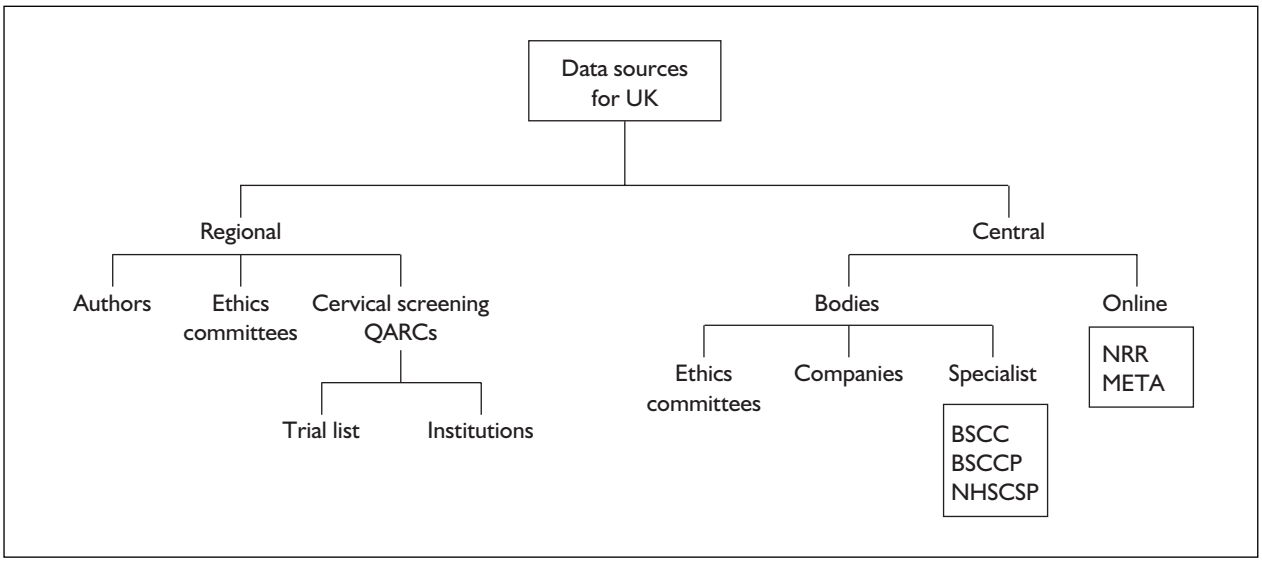


FIGURE 8 *Framework of data sources for ascertaining unpublished literature on automated cervical screening in the UK. NRR, National Research Register; META, Medical Editors Trials Amnesty; BSCC, British Society of Cervical Cytology; BSCCP, British Society of Colposcopy and Cervical Pathology; QARCs, Quality Assurance Reference Centre.*

and in Brussels (site of the European base) were contacted by telephone and e-mail, and representatives of the company were also interviewed. An independent UK distributor of the AutoPap, CellPath, was also contacted and the sales director was interviewed.

- **Research ethics committees (RECs)**: 212 RECs [both local (LRECs) and multicentre (MRECs)] across the UK were contacted by questionnaire. The questionnaire was piloted in the West Midlands, UK, before being circulated nationwide.

Regional data sources

- **Cervical Screening Quality Assurance Reference Centres (QARCs)**: in the NHSCSP, key staff at the QARCs have an informed overview of the programme local to their region. Each QARC was contacted and asked to provide a list of the laboratories and hospitals, and the corresponding lead pathologist and senior biomedical scientist (BMS) in their region. This served as a useful cross-check to the list supplied by NHSCSP. They were further asked to give details of any trials of the three devices that had taken place in their region.

  In total, 212 laboratories were identified across the UK. Either the lead pathologist or senior BMS of each laboratory was contacted by questionnaire. Initially, a questionnaire was circulated in 17 laboratories in the West Midlands as part of a pilot exercise to identify any ambiguities or flaws in the questionnaire that had not been revealed at this point. From the pilot there was a 53% response rate (9/17), with one positive reply. No problems with the questionnaire were noted. As a result, the questionnaires were circulated nationally. Non-responders were sent a second questionnaire.
- **Local research ethics committees (LRECs)**: these were treated as for the QARCs.
- **Authors**: the authors of studies in the UK were contacted to ascertain knowledge of any other studies that may have taken place.

Although for illustrative purposes the data sources are divided into central and regional, the approach was to treat the two surveys (cytology laboratories and RECs) as the primary sources of information, with all other (secondary) sources serving to augment the data retrieved from these. Specifically, details of studies located in the secondary sources were only recorded if they had not been located in one of the surveys.

Results: crude and verified numbers of studies

The results of the two surveys are shown in *Tables 54* and *55*. A reply was considered positive if details of the location and the automated device studied were given on the returned questionnaire. As there was a large number of duplicates the number of potential studies identified represents the number of potentially unique studies that may have taken place on a given device. Study verification was made after following up the details given on the returned questionnaire.

TABLE 54 *Results of laboratory contact survey*

	Raw data totals	Potential studies identified	Studies verified
No. of questionnaires circulated	212		
No. of replies	120		
No. of questionnaires with one or more positive reply	42		
Total no. of positive replies	75		
No. of positive replies relating to:			
PAPNET	42	12	7
AutoPap	25	6	4
AutoCyte	2	2	1
Unknown	6	5	
Totals	75	25	12

TABLE 55 *Results of RECs survey*

	Raw data totals	Potential studies identified	Studies verified
No. of questionnaires circulated	215		
No. of replies	127		
Could not retrieve	5		
Refused owing to confidentiality	4		
No. of questionnaires with one or more positive reply	7		
Total no. of positive replies	7		
No. of positive replies relating to:			
PAPNET	3	3	2
AutoPap	0	0	1
AutoCyte	0	0	1
Unknown	4	4	
Totals	7	7	4

None of the four verified studies in the REC survey was in addition to the laboratory contact survey.

The response rates of the two surveys were very similar and marginally better than the pilot study; 57% response (120/212) to the laboratory contact survey, compared with 59% (127/215) to the REC survey. However, there was a marked difference in the number of questionnaires with at least one positive reply; 20% (42/212, 95% CI 15 to 26%) for the laboratory contact survey and 3% (7/215, 95% CI 2 to 7%) for the RECs.

From all the leads and snippets of information provided by trawling the sources detailed in *Figure 8*, 13 studies were verified in total. Comparing the two surveys, the laboratory contact survey identified 92% (12/13, 95% CI 67 to 99%), compared with 31% (4/13, 95% CI 13 to 58%). Not only were the laboratory contacts a significantly better source of information in terms of the number of studies identified, but the quality of the information was also higher. Whereas in the former full information relating to the study was consistently given, in the latter the name of the institution would be the only piece of information returned on the questionnaire, details such as the name of the device, year or researcher's contact details being missing.

There are several possible reasons for the RECs being a less productive source of information in this survey. It may be that the ethics committees simply did not know of such studies. Certainly, a number of the studies were in-house evaluations, and ethics committee approval may not have been required or sought. This is borne out by a letter received from one of the ethics committees that claimed that the testing of an automated cervical screening device in their hospital did not require REC approval because it was done in parallel with manual screening and so did not involve patients or personal information. The truth of this

TABLE 56 *Results from other data sources*

Data source	Details	Additional studies detected
Central		
Specialist bodies	BSCC annual conference proceedings 1995–2000	1
	NHSCSP	0
	BSCCP	0
Online	NRR	0
	META	0
Manufacturers	TriPath	0
	CellPath (distributor)	0
Regional		
Cervical screening QARCs	All regional coordinators contacted	0
Authors		0

statement has not been tested, but it could account for the relative paucity of responses from the RECs.

A second possibility is that the RECs have approved such studies, but are unwilling to reveal this owing to a concern over breach of confidentiality. Four RECs refused even to look through their records on such grounds. This appears to be a problem that researchers are facing more and more.[132] Access to information that may benefit research is being restricted as RECs experience increasing anxiety about their legal obligations.[132–134] For the benefit of future research it is an issue that needs clarifying.

It must also be recognised that the size of the task may encourage apathy, particularly when the systems are not in place for easy data retrieval. In such cases the survey is reliant on motivated individuals to search through minutes that may lack indexing or organisation.

The results from the other sources are shown in *Table 56*. Note that only studies located in addition to the survey results are considered.

Verification of potential studies was made from follow-up investigations. Where possible, documentary evidence has been sought for the studies listed. Otherwise the information presented is that obtained from personal communications in the form of e-mail or conversation.

Attempts were made to assess whether the verified unpublished studies would have met the inclusion criteria for the systematic review of evaluations of test performance (see Chapter 6), and these are recorded in the tables. Inevitably, these decisions were based on the information available at the time of completing this component of the health technology assessment (January 2002). In some cases there was considerable certainty about whether a study would have been included or excluded; in others there was not and these cases are marked with an asterisk in the tables.

Results: details of verified unpublished studies

The 13 verified studies comprised eight on PAPNET, four on AutoPap and one on the AutoCyte SCREEN. Their details are presented in *Tables 57–59*. All of the studies identified satisfied the definition of an unpublished study used in this report. The PRISMATIC trial[112] was also located as part of the searches here, but has already been appraised earlier in the report (see Chapter 6).

TABLE 57 *Verified unpublished AutoCyte SCREEN study*

Source	Study description	Reference standard	Included/excluded[a]
Hospital A	Early version of the AutoCyte SCREEN, evaluated 1993–1994. AutoCyte SCREEN + PREP vs Manual screen of PREP. Manual arm screened blind of results from AutoCyte SCREEN arm	Probably manual arm	Excluded: no clear reference standard [2]

[a] The number in brackets indicates at which of the step 2 criteria the authors believe the study would have failed, where the study would have passed step 1 criteria.

TABLE 58 *Verified unpublished PAPNET studies*

Source	Study description	Reference standard	Included/excluded[a]
Hospital B	In-house assessment of PAPNET, no further information		Excluded at step 1
Hospital C	In-house assessment of PAPNET, no further information		Excluded at step 1
Hospital D	In-house assessment of PAPNET, no further information		Excluded at step 1
Hospital E	In-house assessment of PAPNET, no further information		Excluded at step 1
BSCC Proceedings 1995, A10, University College London[135]	PAPNET screening vs original diagnosis after rapid rescreening Blind review of images used as PAPNET diagnosis, by two cytotechnologists and one cytopathologist 350 slides with case-mix of abnormal, negative and false-negative slides (shown after rapid rescreening) October 1993 to April 1995	Not described, probably diagnosis obtained from normal manual protocol	Excluded: not a two-armed design [1]
BSCC Proceedings 1995, A11, St Mary's Hospital, London[136]	PAPNET rescreening vs rapid review (2–3 minutes) Blind review of images by two cytotechnologists independently 1017 slides with case-mix of 27 abnormal (15 borderline, 12 dyskaryotic), 26 unsatisfactory and 964 negative	Not described, but diagnosis obtained from normal manual protocol	Excluded: unsuitable reference standard [3]
BSCC Proceedings 1998, A34, Christie Hospital, Manchester[137]	PAPNET screening vs original diagnosis Multicentre European study Retrospective analysis of 12,422 archived slides with case-mix of approx. 300 slides in each abnormal category of atypia/borderline, mild, moderate/severe and carcinoma	Discordant slides between manual and PAPNET arm reviewed by at least one external expert	Excluded: not a two-armed design [1]
Hospital F*	PAPNET vs conventional, in primary screening. Prospective blinded comparison of 13,000 smears Order in which each arm screened a slide randomly assigned PAPNET diagnosis confined to the microscopic review of the coordinates identified on the image tiles (i.e. not full review of the slide)	Discordant slides between the two arms would have been investigated and adjudicated	? Excluded: not clear whether independent panel would adjudicate slides [3]

* Uncertainty about whether the study would have been included or excluded in the systematic review of test performance.
[a] The numbers in brackets indicate at which of the step 2 criteria the authors believe the study would have failed, where the study would have passed step 1 criteria.

Although the authors believe that all evidence presented here is factually correct, some of the evidence by its nature is not available for general perusal in the public domain. Such evidence may in certain circumstances be interpreted as being sensitive, and for this reason the identities of the sources have been anonymised as far as possible.

The single study on the AutoCyte SCREEN evaluated an early version of this device (1993),

TABLE 59 *Verified unpublished AutoPap studies*

Source	Study description	Reference standard	Included/excluded[a]
Leeds General Infirmary Pathological Society abstract[138]	AutoPap vs Conventional +rapid review 25,499 smears screened as cohort Data analysis was of 5981 smears designated by AutoPap as NFR. 18 were false negatives (eight mild, ten moderate/severe dyskaryotic)	Original diagnosis: pathology report	Excluded: not a two-armed design [1]
Three Centre Trial, UK TriPath product insert [35]*	AutoPap GS System vs ?original diagnosis in primary screening 6070 slides in prospective study Study in late 1998	Discordant slides subject to external review for truth diagnosis	? Excluded: not a two-armed design [1]
Hospital G	AutoPap vs original diagnosis in primary screening. In-house assessment over 6 months 6000 slides	Original diagnosis: pathology report	Excluded: not a two-armed design [1]
Hospital H	AutoPap 300 QC rescreening of a few hundred slides. Slides were sent to Seattle, USA, for analysis	Unknown	Excluded: not a two-armed design [1]

* Uncertainty about whether the study would have been included or excluded in the systematic review of test performance.
[a] The numbers in brackets indicate at which of the step 2 criteria the authors believe the study would have failed, where the study would have passed step 1 criteria.

which had received a number of updates before it was removed from the market following the formation of TriPath. There was no clear reference standard, so the study would almost certainly have been excluded from the systematic review of test performance.

Four out of eight PAPNET studies (marked hospitals B–E in *Table 58*) were in-house evaluations and it was difficult to verify whether any formal study design protocol had been used. On average, 2–3 months' data had been collected and in none of the cases had any attempt been made to publish the results. As these were relatively informal evaluations, and raw data were not made available in any of the four studies for independent appraisal or verification, the reports should be considered as anecdotal and would certainly not have been included in the systematic review of test performance.

The main reason for only limited data collection in each of these studies and in most of the other PAPNET studies listed below was the premature winding up of NSI. In some of the cases, despite the best attempts of the investigators, the study data were no longer available.

From the remaining four studies on the PAPNET, three were documented as conference proceedings from the annual scientific meeting of the BSCC. The results of these studies would have been presented at the meeting, but unfortunately were not recorded in an accessible format. One of the abstracts presented, from investigators at the Christie Hospital in Manchester,[137] reported the preliminary findings of a large European multicentre study, which on completion would have been submitted for publication. The design of this study did not meet the criterion of two-armed design, however, as the original diagnosis was used to indicate the result for the manual arm.

Only two of the verified unpublished studies on PAPNET would have passed the two-armed review inclusion criterion. The first, from St Mary's Hospital,[136] is only one of two studies in the unpublished and published literature reviewed in this project that compares the rescreening of smears by an automated system with the rapid rescreening (the standard method of rescreening in the UK) of around 1000 smears. It is documented as an abstract at the BSCC annual scientific meeting of 1995. The results were considered comparable and useful information on

time to screen by both methods was measured. Although the reference standard is not described, it was taken to be the report generated from a primary screen and rapid review, that is, the original diagnosis (Morse A, St Mary's Hospital, London: personal communication, 2001). On this basis, the results of this study would definitely have been excluded from the systematic review of test performance.

The second study that satisfactorily complied with a two-armed design was a much larger study. According to the investigators it was comparable to the PRISMATIC trial.[112] Over 13,000 smears were each screened conventionally and by the PAPNET at hospital F. The slides were allocated such that the order in which each arm screened a particular slide was random. The same screeners were involved, but were blind of results from the previous arm. The PAPNET was, however, used in a manner that departed from the manufacturer's protocol. When atypical cells were located on an image tile, instead of this being followed by a full microscopic review of the slide, the review was confined to the area defined by the coordinates of the cell on the image tile.

Unfortunately, it was not clear at the writing of this report whether the discordant slides would have been subject to adjudication by an independent panel (of at least two members), and for that reason the study would probably have been excluded from the systematic review of test performance. However, even if the study had potentially been includable, the information being provided by personal communication, without a written exposition of the original data, would have been a further problem in incorporating the data from the unpublished study in the review. It appears that the study was close to completion before the PAPNET was withdrawn. However, although the present researchers were informed that the study could probably have been successfully completed before PAPNET's withdrawal, the commercial failure of its manufacturer NSI dampened demand for published research on its evaluation.

There were four verified unpublished studies on the AutoPap, detailed in *Table 59*.

The first, an in-house study at hospital G, compared the AutoPap as a primary screening device with conventional screening (original diagnosis) processing around 6000 slides. The device was on loan and was suddenly withdrawn. The second in-house study was documented in a research thesis and evaluated the AutoPap 300 QC as a rescreening device on a few hundred slides at hospital H. In both cases the raw data were not available for appraisal and information is somewhat scant.

Two studies on the AutoPap, which are accessible, appear from different sources. The first appeared as a poster at the 2001 Pathological Society meeting and was located at Leeds General Infirmary.[138] The two arms were the AutoPap LGS screening and the normal conventional primary screening (including rapid review) of 25,499 smears. The data presented in the abstract, however, were on the 5981 slides that the AutoPap had classified as no further review (NFR). From the pathology reports, 18 abnormalities were detected in this set, of which ten were high grade. Two problems with this study would have meant that it would not have been included in the systematic review of test performance. The first is that the final pathology report, which is generated from the normal conventional protocol, was considered to be the reference standard. Thus, the manual arm is acting as the reference standard and is therefore not independent of the arms being tested. Second, the NFR category is really the negative category and this should be compared with the smears classified as normal in the manual arm, all 23,922 slides. What is not known from this study is how many of these slides contained abnormalities, and so the comparison is an unfair one.

The final study on the AutoPap in the UK involved three centres. The TriPath AutoPap Primary Screening System product insert[35] refers to a UK-based trial involving three centres, but gives no indication as to the timing or the location. The information presented in the product insert is summarised in *Table 60*. From this, there was uncertainty as to whether this was truly a two-armed design and whether the reference standard was appropriate. Although this study has been assessed as being likely to have been excluded from the systematic review of test performance, this designation can only be provisional. The reviewers are aware that there is a full report giving details of the study, a more detailed analysis of the clinical effectiveness, a full cost analysis, and a time and motion study (detailing the AutoPap workflow times compared with conventional screening). However, they were unable to obtain a copy the details of which could be transcribed. Given the direct relevance of this research to the UK and its potential to influence decisions on introduction, it is important that all

TABLE 60 *Details of the 'Three Centre Trial' of AutoPap GS System in the public domain, and areas of uncertainty*

Study/source	Description of study	Reference standard	Analysis
Three Centre Trial, UK TriPath product insert[35] Clinical investigators? Precise locations? When?	AutoPap GS primary screener vs conventional Up to 25% for NFR. Review slides were screened using PapMap overlays. Full review of the slide was made if FOV in PapMap showed an atypical cell 6070 slides enrolled 5531 slides analysed by both arms Excluded: 482 (7.9%) process 40 (0.7%) barcode 17 (0.3%) other Prospective two-armed design Was there blinding between the arms and were both arms subject to test conditions? Were the same screeners used for both arms?	Abnormal discordant slides subject to 'external discrepancy review' for final truth diagnosis Discrepancies in adequacy of the slides between the two arms subject to internal adequacy review How many did the reviewing and what was their expertise? Were they blind to the results of the arms?	Equivalence of performance tested in detecting abnormal (borderline+, mild dyskariosis+ and moderate dyskariosis+) and inadequate slides (McNemar's exact test) Insufficient data to estimate sensitivity and specificity

NFR, 'No further review' designation by AutoPap, i.e. slides ranked as least likely to contain abnormalities. FOV, field of view, location of any area where AutoPap GS detects potentially abnormal cells. PapMap, printout indicating the location of the FOVs on the smear assisting manual review of the slide; the workstation and motorised microscope stage can also move automatically to the *x*, *y*-coordinates of each FOV.

the results of the 'Three Centre Trial' should be fully published in a peer-reviewed journal.

Locating unpublished literature: conclusions

The context of the search for unpublished studies should be put into perspective. This was an extensive trawl across a country that has published one major primary study on automated cervical screening in the past decade.[112] Yet there were 13 'unpublished' studies, and although a significant number would have remained in-house evaluations, there were at least two major studies. The Three Centre Trial is particularly important because it deals with the only currently commercially available version, the AutoPap GS System.

The reporting of the Three Centre Trial is perplexing. If the results of the unavailable extensive report confirm those described in the product insert,[35] then there appears to be nothing to hide on clinical effectiveness. However, this report also deals with cost and process time and it is possible that the contentious issues relate to these. Whatever the reasons for it remaining inaccessible, it is essential that the results of the Three Centre Trial are fully published and peer reviewed.

Without full reports of the methods and results of the two studies on which there was uncertainty about inclusion and exclusion from the review of test performance, it cannot be stated with confidence whether the results concerning clinical effectiveness reported in Chapter 6 are supported or challenged. However, although there are no definite additional unpublished UK studies on test performance that directly contradict the findings on clinical effectiveness, there is a clear indication that major pieces of relevant research may remain unpublished and unrepresented in this and past systematic reviews and health technology assessments.

This suggests that unpublished literature in other countries should ideally be searched to the same level of scrutiny as has been undertaken in this assessment. That the amount of rigorous evidence on the benefits associated with automated image

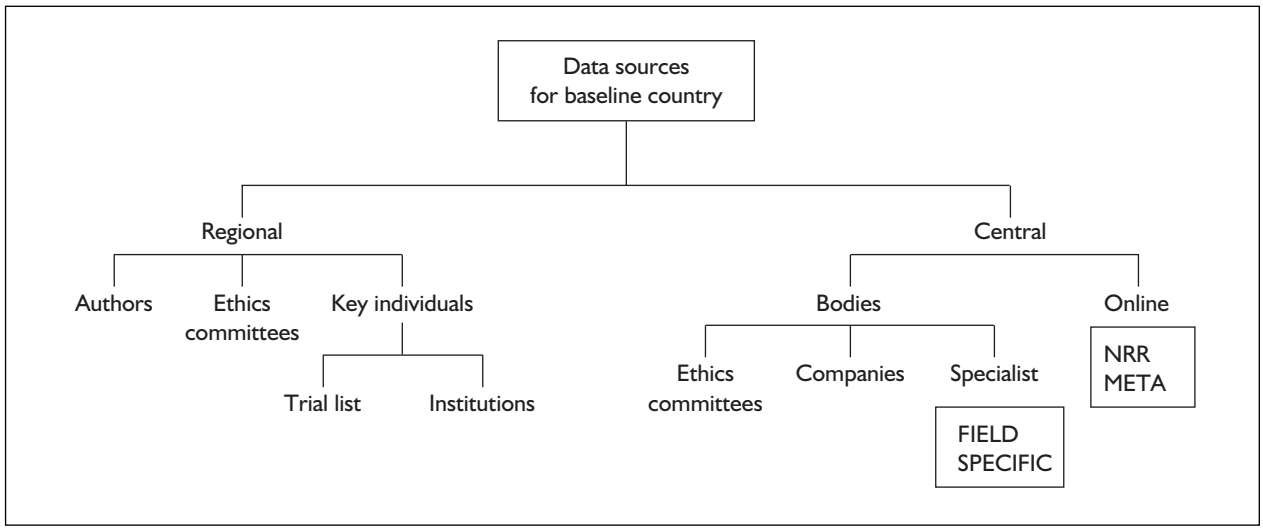


FIGURE 9 *General framework of data sources for ascertaining unpublished literature*

analysis is so limited and finely poised affirms that this is more than an academic exercise. The nature of publication bias, where studies showing interventions in a less favourable light are less likely to be published, further reinforces the need to ensure that assessment of the effectiveness of automated image analysis is based on all relevant literature reaching stated and appropriately high standards of methodological rigour.

The following paragraphs explore how the methods adopted for the search for unpublished material in the UK could be adapted to other countries.

Extending and generalising the search for unpublished literature

To extend the model used to other countries, and indeed other fields, requires a few modifications. A possible general model is outlined in *Figure 9*. As noted in the previous section, there is considerable overlap between the respective arms of the tree. This is expected, as no one arm is complete.

The search model developed in Part 1 was developed specifically for the field of automated cervical screening in the UK. However, there are two essential ingredients to this model, which would need to be present for its successful application to other situations:

- the ability to partition a country's health system into a number of smaller but more manageable subunits, so that a systematic trawl may be made; in the UK the partition is often determined according to geographical boundaries, and certainly in the case of cervical screening, the QARCs and the ethics committee are organised thus
- the use of a number of overlapping data sources that sum together to give, if not complete coverage, very close to complete coverage.

Certain data sources, including authors, companies and specialist bodies, may be expected to be a consistent whatever the field or country.

Theoretically, research that involves some risk to humans on a particular topic should have at least had guidance and where appropriate approval by an independent ethics committee, as laid down by the Declaration of Helsinki. As this guidance is intended to be worldwide this provides potentially a useful resource, since there should be at least some record corresponding to the advice or approval given by an ethics committee on all such studies.

In the UK, ethics committees are organised on a regional basis, with LRECs often being allocated to each hospital within a region. When the study stretches over a number of localities then application must be made to MRECs.

Another change necessary when attempting to generalise the UK specific model is identifying the equivalent of the cervical screening QARCs. The general equivalent used in *Figure 9* has been labelled as key individuals. The identification of the key staff in the QARCs greatly contributed to making the search for unpublished literature in automated cervical screening in the UK as

comprehensive as possible. Their importance therefore should not be underestimated. The concept that emerges from the UK-specific study is that for any speciality of medicine, in a given region, people probably exist who occupy a position that allows them a bird's eye view of developments in research that has been, or is currently being conducted in their speciality. If such individuals have counterparts in other regions throughout the country, then this provides a means of formulating a systematic search across the entire country. The key individuals should be able to provide two main types of information. First, from their own knowledge, they should be able to give details on past studies in their region. Second, they should be able to provide a list of contacts within institutions in their region who would potentially carry out such research. Those institutions and individuals identified would then be contacted. In developed countries, where infrastructures can be expected to be in place, this type of approach is more likely to have success than in less developed countries.

How the different data sources are interrogated is dependent on the sources being considered. The ethics committees and laboratories were considered to be amenable to questionnaire. This was piloted locally to iron out any design faults and to allow a formal evaluation.[139] In contrast, contact with authors, companies and specialist bodies was on a more individual and informal basis.

The success of this approach in other countries depends to a degree on gaining the cooperation of someone who would be able to coordinate and organise a similar strategy in their country. This could provide feedback on the pro forma method, to ascertain whether with appropriate changes it could be applied to that country. If cooperation could not be gained, then coordinating a search from overseas would be much more difficult and in some cases impractical.

One possible source of overseas contacts may be found in the major health technology assessment bodies around the globe. Many of these organisations are members of the International Network of Agencies for Health Technology Assessment (INAHTA). This network provides an important resource possessing a database of all reports produced by each member group, enabling the identification of those health technology assessment bodies and thereby authors who have contributed to the field. Such contributors may have an interest in coordinating a search for unpublished literature in their country. Otherwise a starting point would be the contacts supplied by INAHTA for each organisation or the recruitment of interested parties at international conferences in the field.

Once the method has been circulated and feedback returned, the question of coverage may be addressed. This will ultimately depend on the number of countries willing to take part.

For optimum coverage those countries that produce the highest number of publications should be chosen. If it proves too difficult to coordinate a search for the unpublished literature in that country, this will increase the uncertainty surrounding the impact of the unpublished literature on the results from published studies. Nonetheless, those countries that do participate will add to the understanding of the extent of the unpublished literature and its effect on the published results.

Reflecting on the nature of publication bias

Notwithstanding that the search for unpublished literature was restricted to the UK, it is unusual for systematic reviews and health technology assessments to go to such lengths to identify unpublished material. The primary reason for pursuing this component of the assessment was that there was such a strong prior suspicion that unpublished material might be important in the overall assessment of the clinical effectiveness of automated image analysis devices. However, given that this is an unusual undertaking, it also seems important to reflect on whether the way in which the search was done or its yield has any general implications. The following paragraphs describe the most important issues that arose during the course of this component of the health technology assessment, presented as a series of questions. It is beyond the scope of this project to answer them, but it is hoped that they may feed into further methodological research, extending that already undertaken on this topic by the NHS HTA programme.[100]

What is meant by unpublished studies?

Publication bias asserts that the results of unpublished studies are likely to be systematically different from those of published studies. According to Dickersin's original definition, it is "the tendency on the parts of investigators, reviewers and editors to submit or accept manuscripts for publication based on the direction

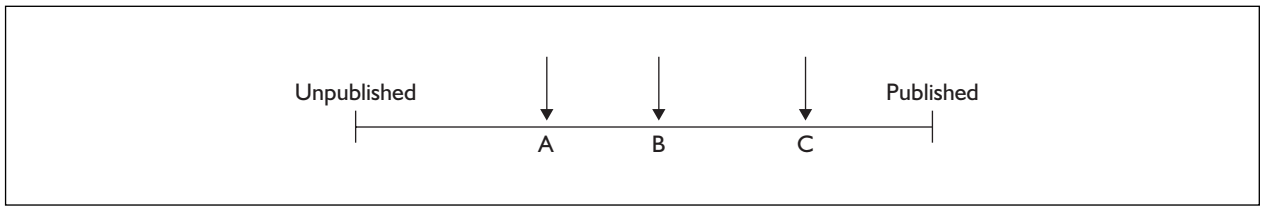


FIGURE 10 *Publication as a continuum*

or strength of the study findings".[101] The implications for reviewers, discussed in greater detail in Appendix 9, are either that as much unpublished literature as possible should be ascertained and included in the review, or that the likelihood of publication bias should be assessed theoretically. The two approaches are not mutually exclusive. Unfortunately, theoretical bases for assessing and adjusting for publication bias rest on a sufficiently large number of included published studies being available, a criterion which, as in this review, is not often met. In these circumstances, the only way of responding to the threat of publication bias is to attempt to identify any unpublished literature that would meet the inclusion criteria for the review.

However, to do this successfully requires a clear understanding of what is meant by published and unpublished studies. Without this it will be unclear whether the chance of including published and unpublished studies is sufficiently close for publication bias to have been minimised. Further, without a clear understanding, much time and effort may be wasted identifying unpublished research that is highly unlikely to meet most review inclusion criteria.

Unfortunately, the observation in this project was that there is not clarity about the concepts of published and unpublished research. In particular:

- The definitions of published and unpublished research in the methodological research defining publication bias are often absent or vague.
- The nature of publication has changed over recent decades, with electronic publication, particularly via the Internet, being a key innovation. In this respect the purpose and nature of what is published in traditional, peer-reviewed, paper-based journals may have changed.
- The facilities available to searchers and their skills have also improved, particularly in the context of undertaking systematic reviews, so that the practical barriers to identifying what in the past would be 'unpublished' research are less.

Pursuing the second point above, publication has been described as a continuum by one author.[140] *Figure 10* shows a schematic representation of this continuum hypothesis to publication. Unpublished and published studies are at the extreme (but different) ends of the scale. What lies between these (points A, B and C) could be described as 'partially published'. There is, however, a practical problem in applying this definition to the concept of publication bias. As there is an infinite number of categories along a continuum, for every point inside the limits there is another point more published than that point. Thus, C is more published than B, which is in turn more published than A.

From this, if an article in the *New England Journal of Medicine* is considered to represent published, then perhaps, conference proceedings could be represented by C, abstracts by B and in-house college circulars by A. Equally, conference proceedings could be considered completely published. The effect of this is to have the threshold for publication set at C. Everything above this threshold is considered published, and everything below as either partially published or unpublished. Unfortunately, it is difficult to imagine a universally agreed threshold for publication: should it be A, B or C? The arguments for any of these are equally valid. In effect, the definition used for publication in this project was created by artificially imposing a threshold between published and unpublished literature, without giving recognition to a partially published study. There is also the problem of deciding where on the continuum the different studies appear. For example, are conference proceedings more published than abstracts or vice versa, or are they equally published? If a single threshold is to be drawn, the studies must be ranked, with a unique position.

A corollary of the above, within the project, was that there were several studies where it was unclear whether they were effectively published or unpublished. The Three Centre Trial is a good example. This is unpublished in the sense that it

does not appear as an article in a traditional scientific journal, and it certainly could not have been identified via searches of bibliographic databases alone. However, a reasonably full account of the methods and results exists in the product insert for AutoPap Primary Screening System.[35] This account is widely available and the information has probably been reasonably well disseminated as a result. So, in this sense it is published. Indeed, it is probably better disseminated than if the information had only been made available as a journal article. Some key details about the method and results were missing, which hampered the assessment of whether the study would have been included in or excluded from the review. However, the details in question would often be absent from publications on this topic in peer-reviewed journals.

The same example illustrates why ambiguity about publication status is important in the context of a review. The key question from the reviewers' viewpoint is whether the Three Centres Trial represents the least optimistic end of the predicted spectrum of results likely to be obtained on the test performance of any intervention by chance alone from several studies. If its 'unpublished' status is a correct marker of its truly being at the less optimistic range of results likely to be obtained, then we ought to be reassured by the fact that the claimed performance of the AutoPap GS System in an 'unpublished' study is 'equivalent' to manual screening. If, however, the 'unpublished' status is coincidental, and has little to do with the results being less optimistic than the manufacturers, investigators, peer reviewers or journal editors had hoped, then no such reassurance can be drawn, even though the study was identified as part of a search for unpublished literature. In this sense, the question, 'What do we mean by unpublished studies' could be restated as, 'How good a marker for studies at the least optimistic end of the range of results expected by chance alone, is a study being designated "unpublished"?' Stated in this way, it also invites the question, 'Can the definition of "unpublished" be refined so that it is a better marker of studies at the least optimistic range of the results expected by chance alone?'

Are some unpublished studies less important than others?

A concern raised by the search for unpublished literature in this health technology assessment is that the vast majority of the unpublished work identified was not of a nature or quality to make it eligible for any review under any circumstances. That it was considered arises from encouragement that all unpublished studies are equally likely to be included in a review. However, this is not the case. There are some studies that are unlikely to be included in any review, and the characteristics of such studies can be used to help to limit the scope of enquiries about unpublished research. For instance, just as an ongoing or uncompleted published study would not be included in a systematic review (other than to indicate that results from this study would need to be considered for inclusion in updates of the review), uncompleted or ongoing unpublished studies should not be considered. Not only does this seem reasonable, because a study being completed is implicitly an inclusion criterion for published studies, but also a study being incomplete or ongoing is probably the most common reason for its remaining unpublished (or not being written up for consideration for publication). Excluding studies that remain unpublished because they are incomplete or ongoing is likely to improve the performance of 'unpublished' status as a marker for less optimistic evaluations of impact referred to at the end of the previous section.

There is legitimate debate about the desirability of abortive research and concern about the frequency with which it occurs. Researchers should clearly be dissuaded from undertaking research that they are unlikely to be able to complete. However, ascertaining such research as part of a search for unpublished literature in a systematic review is likely neither to reduce the frequency of such abortive research nor to improve the review.

Are some important unpublished studies accessible at all?

The final generalisable problem identified from the search for unpublished literature on the test performance of automated image analysis is how to deal with incomplete information on the method, analysis or results. By definition, this is likely to be a greater issue for an unpublished than a published study. Thus, to return to the example of the Three Centre Trial, there was uncertainty about some aspects of the method and level of detail of the results available. Because of this the reviewers did not feel confident about including or excluding the study, and effectively it was excluded. The conundrum is thus: although an unpublished study may be identified and thought potentially eligible, by definition it is more likely that details on method or analysis will be lacking, resulting in an inability to consider fully its results in the review. An extreme and worrying situation would be where sufficient information on the

method of an unpublished study was made available to allow a decision that it should be included in a review, but the detail of the results was suppressed because they were perceived to be poor. Supposition may indicate that this was the reason that the results were not available in full, but it is unclear whether it would be appropriate for a review to indicate in general terms that the results were unfavourable without recourse to the results in full.

Further methodological research on publication bias

The authors' experience in identifying and attempting to incorporate unpublished literature on this topic has identified that there are some important general practical problems associated with the exercise, on which there appears to be little to guide reviewers. Better guidance needs to be developed, particularly with respect to:

- what is the purpose of making specific detailed enquiries for unpublished research
- when it is most appropriate
- what type of unpublished research is most valuable and what are its characteristics
- how unpublished research of this type might be most effectively and efficiently identified
- how the results of unpublished research might best be incorporated into systematic reviews, particularly when details of methods and results are incomplete.

Of central concern to many of the above points is whether the positive association between publication in a peer-reviewed journal (versus not) and a large or statistically significant effect holds for other definitions of published and unpublished that are more relevant to the range of publication and dissemination options currently available. Without reaffirmation of this sort, the case for investing considerable effort in identifying unpublished research will be much undermined. Indeed, given the greatly increased range of options for publication, including online alternatives such as Biomed CENTRAL, it is worth confirming that the type of unpublished research central to the original demonstration of publication bias (a completed study, written up as a full manuscript, submitted and rejected by one or more journals) still exists. This description certainly does not describe the unpublished research identified in this health technology assessment.

The issues raised concentrate on the general problems highlighted. However, there may also be problems specific to the present attempt to deal with unpublished material in a review of a screening/diagnostic test. Publication bias in such situations is the subject of an ongoing HTA methods project; when this is complete, it may be appropriate to revisit the experience of conducting this search for unpublished studies on automated image analysis. One issue that should be emphasised in advance of this, however, is whether publication bias can operate when there is relatively little consensus on what constitutes an optimistic or a pessimistic result, in terms of either size of effect or statistical significance.

Chapter 8

Costs of automated cervical screening devices

Summary of key points

The objective was systematically to review information on the additional cost associated with introducing automated cervical screening devices, to inform the estimates of cost used in the simulation model.

Estimates of cost varied widely, even where the same devices used in the same mode were being considered. Details of how costs had been estimated were generally very limited, which precluded detailed examination of the causes of variation.

The 2001 UK device cost for one AutoPap GS machine, including VAT and maintenance charges over 5 years, was around £650,000, which is considerably greater than the cost quoted in the past.

In the absence of robust information to the contrary, a conservative interpretation of the manufacturer's cost estimates was suggested for the additional cost per slide processed figure to be used in the economic analysis, of £2.66 (range for sensitivity analysis £1.33 to £5.24).

These figures for the AutoPap GS do not include the need for additional facilities, especially laboratory space, nor do they incorporate the potential impact on staff costs. The average staff time required to process slides is claimed to be reduced, but there is no published research on the size of this effect.

The priority for future research should be to obtain accurate, transparent and independent cost impact estimates for AutoPap GS, and the parameters driving this, such as average slide processing time.

The likelihood of cointerventions such as LBC should be anticipated, as should any knock-on effects on other components of cervical screening programmes, such as colposcopy.

Introduction

The main purpose of this component of the report was systematically to collect source data for the simulation model-based economic analysis. However, not only is cost one of the key parameters determining cost-effectiveness, alongside effectiveness, but establishing the true cost of automated cervical screening devices is important because it has probably been a major factor inhibiting widespread uptake of this technology. To this end, based on primary and secondary research by other investigators and enquiry of manufacturers, the objective was to answer the question, 'What are the costs likely to be associated with the introduction of automated cervical screening devices to programmes similar to those operating, or likely to be operating in the UK?'

Method

Cost data were sought directly from the only manufacturer currently producing commercially available automated cervical screening devices. This was supplemented by locally derived cost information collected by the West Midlands Cervical Screening QARC in 1999. This included an assessment of the cost of introducing PAPNET, which is not currently commercially available.

To corroborate this, and to supplement it, particularly in respect of the wider costs of the technology, a systematic review of the literature was undertaken. A single search was done to identify studies providing both information on cost and prior assessments of cost-effectiveness (as described in Chapter 4). The sources interrogated for this combined search were:

- Cochrane Library 2001, Issue 2 (includes NHS EED)
- Medline, 1996 to March 2001
- EMBASE, 1998 to March 2001
- CINAHL, 1998 to May 2001
- CANCERLIT, 1998 to March 2001

- HealthSTAR, 1998 to December 2000
- EconLit, 1998 to May 2001.

The search was deliberately focused on recent literature to maximise the applicability of any results to the current day. The search strategies used combined series of terms capturing the condition of interest (e.g. cervix neoplasms/), the intervention of interest (e.g. image processing computer assisted/) and the type of literature desired (e.g. economics medical/). Full details of the search strategy for MEDLINE are provided in Appendix 5.

Searches of the bibliographic databases were supplemented by examining the reference lists of reviews and included studies identified, looking particularly for citations that indicated that an article had both dealt with automated screening and considered costs or efficiency, cost-effectiveness or cost–utility. The outputs of the searches for reviews and effectiveness were similarly scrutinised.

The inclusion criterion for this section on costs was simply that the study should provide information on the cost of an automated cervical screening device. The only exclusion criterion was where the study was a review or an editorial, where the main aim was to report others' assessments of costs without any significant additional analysis or modelling. Many studies whose main aim was to assess cost-effectiveness were included in the costs review, because they had gone to some lengths to assess cost as part of the process of assessing cost-effectiveness.

No formal appraisal of the included studies was done, although important strengths and weaknesses were noted with reference to the cost items in the Drummond checklist for health economic assessments.[70] Particular attention was paid to the degree to which the costings captured expenditure arising as a secondary consequence of introducing automation, particularly due to:

- additional materials aside from the automated device (i.e. disposable materials)
- need for maintenance/repair of automated device
- need for additional facilities, especially laboratory space
- additional/reduced staff time to prepare slides
- additional/reduced non-medical/medical staff time to read slides
- additional/reduced burden for other components of any cervical screening programme (e.g. colposcopy).

The data abstracted from included studies were primarily costs and resource use. Where costs were in US dollars, rough equivalents in UK pounds are provided, assuming an exchange rate of £1 = $1.4. Analysis was qualitative, in which conclusions were based on the pattern of results revealed in the tabulated results of included studies. Results were separated according to whether they had made direct estimates of cost impact or whether estimates were based on simulation models. Results were also presented by the type of automated device (e.g. AutoPap or PAPNET) and the mode in which it was used (e.g. primary screening or quality assurance).

Results

Yield of search for systematic review

In total, 223 citations were examined, 140 of which were immediately excluded, mostly for reasons of duplication (n=133). The full text of 82 of the 83 provisionally included studies was examined; one study in the *Journal of Clinical Ligand Assay* could not be obtained.[71] Of the 82 examined, 66 were excluded after detailed assessment, 59 because they did not deal with cost/cost-effectiveness and/or did not deal with automation, and seven because they did not provide any new primary data or undertake new analysis or modelling.

Thus 16 studies were included in either this chapter or Chapter 4. The 13 studies contributing information on costs are described here.[12,18,52,60,61,63,72–74,76,141–143]

AutoPap: primary screening mode

Information from manufacturer/local intelligence

In a letter dated 14 August 2001 from AutoPap's UK distributors (CellPath), the quoted basic price for the AutoPap® GS system [CT (cytotechnologist version), one slide wizard without microscope] was £500,000 plus VAT (at 17.5%, total cost £587,500). Based on a higher specification [AutoPap GS system CTM/PTM (combined cytotechnologist and pathologist version), one slide wizard without imaging and one with imaging], the cost of the automated device could rise to £534,000 plus VAT (at 17.5%, total cost £627,450). To these prices needs to be added a maintenance contract at 7% of the purchase price to extend cover beyond the 12-month warranty for parts and labour covered in the purchase price. On this basis the total purchase cost over 5 years for the lowest specification is £622,500, and £664,830 for the highest specification. The approximate cost of a

leasing arrangement over the same period would be approximately £13,000 per month plus VAT, or £916,500 in total.

The quoted capacity of the machines, operating for 7 days a week, 24 hours a day, is 50,000 slides per annum if conventional slides are being processed and 75,000 slides per annum if LBC slides are being processed. Thus, over 5 years, assuming that the automated device operates at maximum capacity throughout that period and that the slides are conventional, the cost to process each slide is £2.49–2.66, depending on specification. If the slides were LBC the cost to process each slide falls to £1.66–1.77, again depending on specification.

The letter also provides a spreadsheet allowing potential customers to calculate individual cost per slides, incorporating variation in annual numbers of slides to be processed, among other parameters. Making some assumptions about increases in the number of slides that AutoPap GS can process (the letter suggests a four-fold increase) and that laboratory staff time and labour cost will be saved in direct proportion, the spreadsheet also offers figures on the overall savings in the cost of slide processing/reading that may accrue from introducing AutoPap GS. The letter suggests considerable savings. However, this is dependent on the validity of the assumptions. Clearly, a major assumption is the marked reduction in processing time. No published study to support reduced processing times with the AutoPap GS System was identified in the review of effectiveness (see Chapter 6). The only study identified that considered the AutoPap GS System (as opposed to its predecessors) was the Three Centre Trial, identified in the search for unpublished literature and reported in Chapter 7. This study was mainly concerned with the test performance of AutoPap GS (which operates principally as a replacement for the primary screening stage). This study apparently included an associated time and motion analysis assessing the workflow and labour costs, and it may be on this that assumptions about increases in numbers of slides are based. However, neither the methods nor the details of the analysis of this time and motion study appear to have been published in any way.

The local evaluation of cost impact carried out by the West Midlands Cervical Screening QARC in March 1999 suggests a cost of £80,000 per annum per 55,000 slides. Ignoring the slight difference in maximum expected output, this would be equivalent to £400,000 over 5 years, considerably less than the figures of £622,500–664,830 derived directly for the distributor's quote above. The main source of the difference appears to be the assumed purchase price of the AutoPap system, which was only £300,000 in the West Midlands estimate in March 1999. On this basis, the cost of the technology in the UK at least seems to be increasing. This would be consistent with the development in the technology that appears to have taken place over the past 5 years.

Information from systematic review

No published studies were identified where the cost impact of introducing AutoPap in primary screening mode was directly observed. Two studies provided other information,[61,141] as detailed in *Table 61*.

Unfortunately, both studies have significant limitations with regard to the information provided on cost. Not least of these are that there are important potential conflicts of interest in both. Further, in the study by Grohs,[141] the figure of 1996 US$3.14 (≈UK£2.25) additional cost per slide of AutoPap over manual slide processing/reading is likely to be an underestimate. In the study by Smith and colleagues,[61] the limitation is a complete lack of information on how the suggested marginal cost of AutoPap over manual screening, 1997 US$4.50 (≈UK£3.20), was derived.

Summary of information on cost of AutoPap used in primary screening mode

It is debatable whether, in this instance, published estimates of cost increase our confidence about what the true costs of introducing AutoPap might be. What the analysis does show is that calculation of cost is not as simple as it might at first appear. How the costs were calculated, the year in which they were measured, effects of currency fluctuations and the scope of the costs included, in addition to whether the costs quoted are correct, all need to be taken into account. The corollary of this is that although there appears to be some consistency in the additional cost per slide data, extreme caution needs to be exercised in interpreting this as indicating that the true additional cost per slide of introducing AutoPap in primary screening mode will be in the region of £2.25–3.20 where conventional slides are being processed.

Particularly given that in the UK the change of greatest interest is probably the cost of adding automation to a screening programme where most slides are thin-layer preparations, the need for

TABLE 61 *Studies assessing the cost impact of introducing AutoPap in primary screening mode*

Study	Nature of assessment	Results	Comments
Grohs, 1998[141]	Detailed modelling of cost and cost per slide for an average laboratory processing 100,000 slides per year (US$, no cost year given, but probably 1996). Model only considers costs of production (no healthcare costs or consequences of screening incorporated). Laboratory costs for space, billing and collection, courier and/or transport and overheads are not included	Current practice (manual + 10% rescreen) $1.03 million, $10.31 per slide; AutoPap Screen (AutoPap looks at all first; 30% most normal not examined further; remaining 70% screened manually with 10% random manual rescreen) $1.35 million, $13.45 per slide (≈ £9.60). Additional cost of AutoPap over manual +$3.14 (≈£2.25) per slide	Author noted to be vice president of AccuMed International, manufacturer of AcCell/TracCell (the cost of which is also considered in the paper). Letter to Editor suggests that costs may be underestimated; acknowledged by author, emphasising that purpose of article was to demonstrate value of model, not to provide exact costings. Reviewer notes that cost of AutoPap is based on $5.00 per slide; this suggests a machine cost of c. $250,000 (≈£180,000) assuming an annual capacity of 50,000. Such a figure is far lower than current figures
Smith *et al.*, 1999[61]	Cost data (US$, 1997) used to populate cost-effectiveness model; source NeoPath (manufacturer)	Cost of AutoPap Screen = +$4.50 (≈£3.20) over manual screening. Cost of manual screening (taken from paper by Eddy, 1987[79]) = $20 ($12.00–36.00), thus cost of AutoPap Screen = $24.50 ($16.50–40.50) (≈£17.50)	Article reported in study was supported by NeoPath. No information is provided on how the additional cost of AutoPap provided by NeoPath was calculated

accurate, transparent and independent assessments of the likely cost impact of introducing the single currently commercially available automated system is paramount.

In the interim, the crude assessment of additional cost of AutoPap based on quoted machine/maintenance costs is probably the best approximation in the UK (i.e. £2.49–2.66 per conventional slide processed, or £1.66–1.77 per LBC slide). Manufacturers will undoubtedly argue the such an estimate is far too pessimistic, as it does not take into account the likelihood that, because 25% of slides are only examined by AutoPap without any manual examination, the machine/maintenance costs will be substantially offset by savings in laboratory staff time. Unfortunately, the impact of introducing AutoPap on the average time per slide (considering all slides received by a laboratory) is not clear. The suggestion received in the letter from the UK distributors of AutoPap that it may improve screening output four-fold, that is, change the average laboratory staff time spent to process a slide from, say, 8 minutes to 2 minutes, seems very optimistic. Further, as well as there being arguments that the crude estimate of additional cost of AutoPap based on machine/maintenance costs will be an overestimate, there are equally plausible reasons that it may be an underestimate. In particular, it assumes that it will be possible to organise a system where all machines operate to their promised annual capacity, something that is highly unlikely to be achieved; they do not take into account additional time required by laboratory workers to load the AutoPap machine; they do not include the cost of additional facilities; and they assume that no checking of the slides AutoPap archives will be required. The crude estimate of costs also assumes that the number of false-negative and false-positive results remains substantially unchanged with the introduction of AutoPap used in primary screening mode. This seems reasonable, but is based on few rigorous data and so still needs verification (see Chapter 6).

Finally, it should be emphasised that none of the above provides information on the possible wider cost impact of introducing AutoPap in primary screening mode, taking into account knock-on effects of increased or decreased requirements for colposcopy and its sequelae.

AutoPap: quality assurance mode

Information from manufacturer/local intelligence

There is no information from the manufacturer on the cost of AutoPap QC, as this machine is no longer available.

Information from systematic review

No published studies were identified that made direct observations of the cost impact of introducing AutoPap in quality control mode. Six studies contributed other cost information on the additional cost per slide of AutoPap used in quality control mode.[12,18,52,63,141,142] Further details are provided in *Table 62*.

The chief feature of the five included studies that provide easily derived approximate UK equivalent additional costs of the AutoPap used in quality control mode relative to manual screening is their variability [$2.00 (≈£1.43) per slide,[142] $3.29 (≈£2.35) per slide,[141] $4.58 per slide (≈£3.27),[18] $5.00 (≈£3.57) per slide[63] and $7.58 (≈£5.41) per slide].[12] The approximate two-fold variation is highly likely to be due to variation in the comparators, methods used to calculate cost and the time at which costs were calculated. Given that AutoPap used in quality control mode does not seem to be the current focus of attention, it is debatable whether identifying which of the estimates is likely to be most accurate is of value (if indeed this were possible, given the limited information available in many cases on how costs have been calculated). However, that there can be such variation in cost estimates indicates the need to have accurate, transparent, independent estimates for costs that are of key importance, that is, the additional costs of AutoPap used in primary screening mode. Further, the estimates of cost will almost certainly vary by country, and the fact that none of the identified information on cost addresses the situation in the UK is of concern.

Summary of information on cost of AutoPap used in quality assurance mode

The only information available on the cost impact of AutoPap used in quality control mode comes from published studies and health technology assessments. All were conducted in North America (five in the USA and one in Canada). The chief feature is the variability of the cost estimates. Analysis of the cause of this is probably of academic interest only. However, the fact that there is considerable variation is important and demonstrates the need to have accurate, transparent and independent assessments of cost impact derived in the country in which the automated screening device is to be applied.

PAPNET: primary screening mode

Information from manufacturer/local intelligence

No information is available from the manufacturer, as NSI has gone out of business and PAPNET is no longer commercially available.

The assessment of the West Midlands Cervical Screening QARC in March 1999 was that the costs of implementing automation across the West Midlands region using the PAPNET machine were likely to be similar to those for AutoPap (£724,000 per annum versus £888,000).

Information from systematic review

No studies were identified that directly observed the cost impact of introducing PAPNET in primary screening mode. One study did, however, directly observe and cost the potential time saved in this way, which is applicable to the use of PAPNET in primary screening mode.[76] Details of this study are given in *Table 63*.

The study by Troni and colleagues[76] provides some support for the possibility that PAPNET used as a primary screener would reduce cytologist time spent per slide and so offset the cost of the device and/or charges for its use. The magnitude of this is not unequivocally established by this single study, which is open to bias. It provides incomplete information about the full impact of introducing PAPNET, even on the costs of processing and reading slides, as it is debatable that the charge of $4.00 quoted is realistic (see next section) and covers all the cost implications of introducing PAPNET. Although not costed, other studies provide information on the impact of processing times and are discussed in Chapter 6. The most important of these with respect to the use of PAPNET in primary screening mode is the PRISMATIC study,[112] which estimated a reduction in average processing time from 10.4 minutes to 3.9 minutes.

Besides the study by Troni and co-workers, there were no studies considering cost impact. A number of published articles identified charges for PAPNET that would be likely to apply equally whether this automated device was used in primary screening mode or not. For convenience these are described in the next section of this chapter.

Summary of information on cost of PAPNET used in primary screening mode

Very little information was identified on the costs of using PAPNET in primary screening mode. The single partly relevant published article[76] indicates that PAPNET used in primary screening mode could reduce the time taken to process slides and so partially offset the charges associated with PAPNET. The size of reduction on processing time measured in this study, however, does not agree with other studies identified in the clinical

TABLE 62 *Studies assessing the cost impact of introducing AutoPap in quality assurance mode*

Study	Nature of assessment	Results	Comments
Brown and Garber, 1999[63]	Information (US$, 1996) based on survey of pathology laboratories in northern California and financial reports (no further details given)	AutoPap: additional cost per WNL slide for AutoPap-assisted rescreening = $5 (≈£3.57)($4.25 covers processing cost, $0.75 to rescreen manually 20% identified by AutoPap). Marginal costs added to cost per slide for manual Pap + 10% manual rescreening = $75.75, i.e. AutoPap $80.75 (≈£57.68) ($78–82 used in sensitivity analysis)	
ECRI, 1998[142]	Published charge/cost information	AutoPap charge $2 (≈£1.43) to $5 (≈£3.57) per slide cost to laboratory depending on contract, which stipulates minimum monthly volume. Device cost $350,000 (1998 ref.) (≈£250,000)	Does not provide any indication of wider cost impact
Grohs, 1998[141]	Detailed modelling of cost and cost per slide for an average laboratory processing 100,000 slides per year (US$, no cost year given, but probably 1996). Model only considers costs of production (no healthcare costs or consequences of screening incorporated). Laboratory costs for space, billing and collection, courier and/or transport and overheads are not included	Current practice (manual + 10% rescreen) $1.03 million; $10.31 per slide; AutoPap QC (manual then top 10% identified by AutoPap subject to rescreening) $1.36 million, $13.60 per slide; + $3.29 (≈£2.35) per slide over manual	Author noted to be vice president of AccuMed International, manufacturer of AcCell/TracCell (the cost of which is also considered in the paper). Letter to Editor suggests that costs may be underestimated; acknowledged by author, emphasising that purpose of article was to demonstrate value of model, not to provide exact costings
Hutchinson, 1996[18]	Cost (US$, year unknown) data used to populate simple arithmetic model estimating total costs and cost per additional case of various rescreening procedures, including 100% rapid rescreening, 100% manual rescreening and automation. Costs derived from cytological practice, professional press and literature, and transcripts of FDA proceedings, 7–8 August 1995	Selection cost per slide = $0 for manual rescreening systems, $5 for AutoPap. Costs of rescreening slides calculated for all at $5 per slide. Total cost for 50,000 slides: no rescreening $250,000, 100% rescreen $495,000 [$9.9 per slide (≈£7.07)]; 100% rapid (30 second) rescreen $291,000 [$5.82 per slide (≈£4.16)]; AutoPap (10% selected for triage) $520,000 [$10.40 per slide (≈£7.43)]. Additional cost for AutoPap over rapid rescreen $4.58 per slide (≈£3.27)	

continued

TABLE 62 *Studies assessing the cost impact of introducing AutoPap in quality assurance mode (cont'd)*

Study	Nature of assessment	Results	Comments
McCrory *et al.*, 1999[12]	Information (US$, 1997) used to populate cost-effectiveness model	Pap smear collection and processing (based on MEDSTAT data analysis for 20–64-year-olds) average $38.68 (range $25.32–43.57). Additional cost of AutoPap (manufacturer estimates) average of $7.58 (≈£5.41) per slide, ranging from $7.15 to 8.00. (Notes that when AutoPap used as screener, for every 100 slides processed, only 90 as opposed to 110 will need to be examined manually when using the AutoPap Screen system)	Assessment of cost impact of automation complicated by considering "new technology which improves rescreening sensitivity": this incorporates the potential of both AutoPap and PAPNET, considering them equally, with an incremental cost of $10.00 ($5–15 used in sensitivity analysis)
Noorani *et al.*, 1997[52]	Costs (Can$, ?1996) used to populate decision-analytic model used to estimate cost-effectiveness of introducing automation. Derived from questionnaire sent to manufacturer	Cost of AutoPap rescreening = Can$7 per slide. The modelled cost impact gives costs for 4 million slides processed as Can$34.4 million manual alone, Can$36.3 million manual+10% manual rescreen, AutoPap Can$59.5 million	Cost impact only considers costs associated with screening and processing slides

WNL, within normal limits.

TABLE 63 *Studies assessing the cost impact of introducing PAPNET in primary screening mode*

Study	Nature of assessment	Results	Comments
Troni *et al.*, 2000[76]	Two experienced cytologists, trained to read PAPNET by the manufacturer, examined 1000 routine slides seeded with 81 false negatives (42 true screening errors; 39 sampling errors). They first assessed slides in the usual manner, and then using PAPNET. The two periods of examination were separated by 20 days	The average time taken to process slides manually was 4 minutes 30 seconds (A), and 4 minutes 45 seconds (B). The average times when PAPNET was used were 3 minutes 50 seconds (A) and 4 minutes (B). The time saving associated with using PAPNET was thus –40 seconds per slide (A) and –45 seconds per slide (B). This was valued in Italian lira and converted into US dollars (1$ = 1900 Italian lira). On this basis the charge to process each slide by PAPNET ($4.00) was reduced by $1.40 (≈£1.00) attributable to the time saved	The order of assessment, manual then PAPNET or vice versa, is unclear; if manual was first and PAPNET second it is possible that greater familiarity with the slides might have accounted for part of the reduction in average time to process the slides. The basis upon which the time was valued was unclear. The stated charge for PAPNET is extremely low

TABLE 64 *Studies making direct observations of the cost impact of introducing PAPNET in quality assurance mode*

Study	Nature of assessment	Results
Brotzman *et al.*, 1999[74]	Prospective assessment of 1200 consecutive slides over 6 months (laboratory would normally process 6000 slides over this period), previously assessed as negative after manual screening + 10% manual rescreening, then submitted to PAPNET rescreening. Records kept of: (a) clerical time required to collect and dispatch slides to PAPNET for imaging, (b) cytotechnologist time spent reviewing data tapes and/or glass slides, and (c) pathologist time reviewing PAPNET cases triaged for PAPNET review. Time used to derive additional laboratory cost per PAPNET slide	For 1200 slides: (a) clerical time = 20.3 hours, (b) cytotechnologist review time = 37.8 hours, and (c) pathologist time = 0.6 hours (37 smears required pathologist triage). Using labour costs of (?1997) US$11.72 for laboratory aid and $22.18 for a laboratory cytologist (mid range of salary), an additional cost of labour of $1 per slide was derived. The charge to the laboratory from PAPNET was $18 per slide, giving a total of $19 (≈£13.57). The study estimated that the equivalent cost of manually rescreening each slide was $2.20 (assumes cytotechnologist dealing with ten slides per hour and an hourly rate of $22.18)
O'Leary *et al.*, 1998[72]	Assessment of 5478 slides identified retrospectively as 'within normal limits' or 'benign changes' between 1994 and 1995 after manual + 10% random rescreening. Time to deal with slides manually and to assess PAPNET data tapes apparently measured, although not explicitly stated and no detail on how this was achieved	3864 (71%) were triaged as negative without further review; 29% required microscopic review. In these slides the time taken to undertake a PAPNET-assisted rescreen was found to be one-third as long as the time taken to do a manual rescreen (how?). This was used to estimate cost per slide for manual and PAPNET-assisted screening: (?year) US$3 vs $9.38 (≈£6.70). The latter is derived by adding fees charged by NSI ($7.50) + cytotechnologist time (1/3*$3 – manual cost per manual slide) + extra time required to re-examine those slides highlighted by PAPNET as needing microscopic review (0.29*$3)

effectiveness review, especially the PRISMATIC study.[112] Several studies report charges for PAPNET, which probably apply equally whether PAPNET was being used in primary screening or quality assurance mode. These charges are reported in full in the next section.

PAPNET: quality assurance mode

Information from manufacturer/local intelligence

No information is available from the manufacturer, as NSI has gone out of business and PAPNET is no longer commercially available.

Information from systematic review: studies directly measuring cost impact

Two published studies were identified,[72,74] details of which are given in *Table 64*.

The two included studies, by Brotzman and colleagues[74] and O'Leary and colleagues[72] were reasonably well conducted, although details of how the units of costs were derived are scant in the latter. Both studies suggest that there is an important additional cost associated with the use of PAPNET in quality assurance mode. The magnitude of this was an average of $1.00 (≈£0.71) per slide and $1.25 (≈£0.89), respectively.

As important a finding, particularly considering the results of all the studies directly assessing the cost of PAPNET, is the variation in charges for the use of PAPNET: $4.00 (≈£2.86),[76] $7.50 (≈£5.36)[72] and $18.00 (≈£12.85).[74] Such variation is perhaps not surprising given that the studies were done in different locations, at different times, with different volumes of slides to be processed. However, it does point to an important issue that needs to be taken into account in assessing cost and cost-effectiveness, that a single charge (often provided by the manufacturer) may not adequately reflect true variation in cost, where processing by the automated device is undertaken by a third party, outside the laboratory.

Finally, the included studies do not provide information about costs beyond charges and changes in laboratory staff time. In particular, there is no information about costs arising from consumables and facilities. In the latter respect it is notable that the PAPNET machine has special requirements in that it should be sited in a separate room, which is vibration free with a concrete floor and air conditioning. As for previous sections, no information is available on what the knock-on costs might be to other

elements of the cervical screening programme (e.g. colposcopy services).

Information from systematic review: studies assessing cost impact, other than by direct observation

Eight studies provided other information on the costs of using PAPNET.[12,18,52,60,63,73,141,142] Details of these are provided in *Table 65*.

The range of charges identified in the health technology assessment by ECRI[142] confirms the point made in the immediately preceding subsection.

Beyond this, the chief feature of the six included studies that provide easily derived approximate UK equivalent additional costs of the PAPNET used in quality control mode relative to manual screening is their variability (cost per slide): $7.00 (≈£5.00),[73] $10.00 (≈£7.14),[63] $10.32 (≈£7.37),[18] $10.41 (≈£7.41),[60] $15.00 (≈£10.71)[12] and $18.66 (≈£13.33).[142] The approximate three-fold variation is highly likely to be due to variation in the comparators, methods used to calculate cost and the time at which costs were calculated. Of some concern is the fact that the two lowest estimates of additional cost come from authors with conflicts of interest likely to favour PAPNET, and the highest estimate comes from an author with a conflict of interest likely to disadvantage PAPNET. As in the equivalent section to this for AutoPap used in quality control mode, it is debatable whether identifying which of the estimates is likely to be most accurate is of value (if indeed this were possible, given the limited information available in many cases on how costs have been calculated). However, again the pronounced variation in cost estimates indicates the need for accurate, transparent, independent estimates for costs that are of key importance, and that these costs estimates need to be conducted in the country where the technology is to be applied.

One final important point is that the additional costs of automated devices over manual screening consistently suggest that PAPNET is a more costly option than AutoPap, when both are used in quality assurance mode.

Summary of information on cost of PAPNET used in quality assurance mode

Unfortunately, the analysis of cost information on PAPNET used in quality assurance mode has become academic, given that the device is no longer available. From the point of view of cost alone this appears to be justified, as PAPNET seems to be considerably more expensive than the main alternative, AutoPap.

The main conclusions to be drawn from this section arise from the observed pronounced variation in cost estimates, showing the need for accurate, transparent and independent assessments of cost impact derived in the country in which the automated screening device is to be applied.

Other devices

One other study providing information on costs of AutoCyte used as a primary screener was identified.[143] This study, by Bishop, suggests that from the point of view of cost, the combination of LBC and semi-automation can be relatively inexpensive compared with AutoPap or PAPNET. Some caution is required in taking these results at face value, particularly as the cost estimates do not appear to have been replicated and there is a potential conflict of interest.

Conclusions

The systematic review of cost data on the impact of introducing automated image analysis devices superficially revealed a wealth of data. Unfortunately, its value was limited by:

- irrelevance to the policy decision of most interest: the introduction of AutoPap GS (as opposed to PAPNET and predecessors of the AutoPap GS, which are no longer available)
- poor quality, particularly the level of detail supplied about which elements of cost were included and how they were measured.

Variation in the latter almost certainly contributed to the most marked feature of the cost data, their variability. This was true even where the costs were for the same device used in the same mode. The need for well-conducted future cost research is emphasised, as too is the need for such analyses to be carried out free from conflicts of interest, which were frequently observed in past studies.

Because the cost data on PAPNET and predecessors of the current version of AutoPap are largely of academic interest, the conclusions focus on the cost data relating to the AutoPap GS.

According to the UK distributor, the 2001 device costs including VAT and maintenance charges over 5 years would be approximately £650,000 per machine, with some variation depending on the

TABLE 65 *Studies making an assessment of the cost impact of introducing PAPNET in quality assurance mode, other than by direct observation*

Study	Nature of assessment	Results	Comments
Brown and Garber, 1999[63]	Information (US$, 1996) based on survey of pathology laboratories in northern California and financial reports (no further details given)	PAPNET: additional cost per WNL slide = $10 (≈£7.14) Processing $8.50; viewing PAPNET images + manual rescreening those identified as abnormal (assumed to be 20%) = $1.50. Marginal costs added to cost per slide for manual Pap + 10% manual rescreening = $75.75, i.e. PAPNET $85.75 (≈£61.25) ($81–90 used in sensitivity analysis)	
ECRI, 1998[142]	Published charge/cost information	PAPNET $8 (≈£5.71) to $18 (≈£12.86) per slide cost to laboratory depending on contract (1997 ref.). Handling cost needs to be added to charge; estimate of $5 (≈£3.57) per slide (1997 ref.). PAPNET charge to patient where insurer does not cover = $35–50 (1998 ref.)	Does not provide any indication of wider cost impact
Grohs, 1998[141]	Detailed modelling of cost and cost per slide for an average laboratory processing 100,000 slides per year (US$, no cost year given, but probably 1996). Model only considers costs of production (no healthcare costs or consequences of screening incorporated). Laboratory costs for space, billing and collection, courier and/or transport and overheads are not included	Current practice (manual + 10% rescreen) $1.03 million, $10.31 per slide; PAPNET (manual screen and 10% rescreen, then PAPNET examines all WNL, cytotechnician examines data tape; 25% require further manual rescreening) $2.90 million, $28.98 per slide; +$18.66 per slide (≈£13.33) over manual	Author noted to be vice president of AccuMed International, manufacturer of AcCell/TracCell (the cost of which is also considered in the paper). Letter to Editor suggests that costs may be underestimate; acknowledged by author, emphasising that purpose of article was to demonstrate value of model, not to provide exact costings
Hutchinson, 1996[18]	Cost (US$, year unknown) data used to populate simple arithmetic model estimating total costs and cost per additional case of various rescreening procedures, including 100% rapid rescreening, 100% manual rescreening and automation. Costs derived from cytological practice, professional press and literature, and transcripts of FDA proceedings, 7–8 August 1995	Selection cost per slide = $0 for manual rescreening systems, $10 for PAPNET. Costs of rescreening slides calculated for all at $5 per slide. Total cost for 50,000 slides: no rescreening $250,000; 100% rescreen $495,000 [$9.9 per slide (≈£7.07)]; 100% rapid (30 second) rescreen $291,000 [$5.82 per slide (≈£4.16)]; PAPNET (27% selected in triage for further examination) $807,000 [$16.14 per slide (≈£11.53)] Additional cost for PAPNET over rapid rescreen $10.32 per slide (≈£7.37)	
McCrory *et al.*, 1999[12]	Information (US$, 1997) used to populate cost-effectiveness model	Pap smear collection and processing (based on MEDSTAT data analysis for 20–64-year-olds) average $38.68 (range $25.32–43.57). Additional cost of PAPNET (manufacturer's estimate of costs) average of $15 (≈£10.71), ranging from $12 to 18 depending on volume. (Notes that charges by laboratories for PAPNET rescreening may be substantially higher, at $30 or $40)	Assessment of cost impact of automation complicated by considering "new technology which improves rescreening sensitivity": this incorporates the potential of both AutoPap and PAPNET, considering them equally, with an incremental cost of $10.00 ($5–15 used in sensitivity analysis)

continued

TABLE 65 *Studies making an assessment of the cost impact of introducing PAPNET in quality assurance mode, other than by direct observation (cont'd)*

Study	Nature of assessment	Results	Comments
Noorani *et al.*, 1997[52]	Costs (Can$, ?1996) used to populate decision-analytic model used to estimate cost-effectiveness of introducing automation. Derived from questionnaire sent to manufacturer	Cost of PAPNET rescreening = Can$14 per slide. The modelled cost impact gives costs for 4 million slides processed as Can$34.4 million manual alone, Can$36.3 million manual+10% manual rescreen, PAPNET Can$ 84.8 million	Cost impact only considers costs associated with screening and processing slides
Radensky and Mango, 1998[60]	Cost data (US$, 1997) used to populate model; source PAPNET manufacturer. Additional cost includes processing, reading of data tapes and manual review of triaged slides	Cost of manual screening = $35.60 per slide; PAPNET = $46.01, i.e. additional cost of PAPNET over manual of +$10.41 (≈£7.41) per slide	Evaluation of this study needs to take into account exchange of Letters to the Editor concerning a study by O'Leary *et al.*, pointing to problems with the CEA conducted by Radensky and Mango. The correspondence also reveals that in 1998 Dr Mango was vice president and chief medical officer of NSI (manufacturer of PAPNET) and that Dr Radensky was a paid consultant of the same company
Schechter, 1996[73]	Costs (US$, ?1994) used to populate Markov model used to estimate cost-effectiveness of introducing PAPNET	Cost of PAPNET rescreening (from manufacturers) = $30 per slide; manual screening = $23 per slide, i.e. additional cost of PAPNET over manual of +$7 (≈£5) per slide	Unable to identify modelled impact on total cost; results of modelling only expressed as marginal cost-effection ratios. Evaluation of this study needs to take into account exchange of Letters to Editor concerning a study by O'Leary *et al.*, pointing to inaccuracies and an inability to replicate the findings of the Schechter CEA. Also reveals that in 1998 Dr Schechter was a consultant to NSI (manufacturer of PAPNET)

specification. The costs over 5 years would be approximately £900,000 if the devices were leased. The device costs quoted appear to be significantly greater than those supplied in previous UK costing exercises and the AutoPap device costs quoted in past literature. This is in keeping with the technological development that has occurred.

Care is required in translating these figures into an additional cost per slide processed. This requires assumptions about the number of slides that can be processed, and can, as in the manufacturers' estimates of cost per slide processed, include assumptions about savings resulting from claimed reductions in processing time. In view of this, the present authors have adopted an explicit approach to the data on cost per slide processed to be used in the economic evaluation. In this, the annual slide capacity is assumed to be 50,000 slides (quoted capacity for normally prepared cervical smears), with any savings associated with reductions in staff costs being accounted for separately.

On this basis the additional cost per slide processed is £2.66, with a range of £1.33–5.32 being used for the purposes of a sensitivity analysis (increased or decreased by a factor of 2 decided upon arbitrarily, but with reference to observed values quoted in the literature). This figure does not include:

- any need for additional facilities, especially laboratory space
- potential impact on staff costs
- knock-on effects to other components of the cervical screening programme (e.g. altered need for colposcopy).

In any assessment of cost impact, changes resulting from these factors would need to be accounted for separately. The figure of £2.66 per slide is consistent with the numerical values of additional cost of an AutoPap Screen smear over manual obtained from two articles identified in the systematic review by Grohs[141] and Smith and colleagues.[61] The reassurance provided by this observation is, however, undermined by concerns raised by conflicts of interest, as a minimum.

As already indicated, possible reductions in average staff time required to process slides are likely to be particularly important influences on net cost, with a potential to offset substantially or even completely the additional device costs, if manufacturers' claims are to be believed. However, whether this occurs is critically dependent on the size of effect on average processing time, in which respect two points are important:

- So far, there has been no published research on the directly observed impact of using AutoPap GS, or indeed the use of any version of AutoPap in primary screening mode on processing times.
- When the impact on processing time has been measured for PAPNET, variability has been pronounced, presumably indicating that the parameter is sensitive to the method by which it is calculated, the setting and the individuals doing the testing.

An inevitable consequence of the above is that the main conclusion of this component of the health technology assessment must be that further research on the impact on processing times is a priority. Without it, assessment of cost impact of the only currently commercially available automated image analysis is incredibly speculative. With regard to collecting the necessary data, great care should be taken with regard to the method and independence, learning from the observed problems with data on PAPNET.

Ideally, if further research effort is going to be devoted to improving the estimates of cost impact, the effects of the need for increased facilities and knock-on effects on other components of the cervical screening programme should also be considered, as should the possibility of simulating how the cost of introducing automation may interact with other related technologies such as LBC or HPV.

Chapter 9

Modelling the health economic impact of automation

Summary of key points

The objective was to assess, the question, 'What is the cost-effectiveness of introducing automated cervical screening devices to programmes similar to those operating, or likely to be operating in the UK?'

Based on an analysis of the strengths and weaknesses of previous attempts to model the impact of automation, a new model using DES was developed. The screening and treatment procedures in the model were based on what occurs in the NHSCSP. The model parameters were partly taken from UK-specific data, and partly from data used in other models.

For some of the model parameters the authors were not confident about the appropriateness of the data they had to use. HPV incidence and progression, in particular, appear to have been derived from populations that may not be typical of the general screened population.

Deriving the probabilities of smear outcome (normal, inadequate, borderline, low grade or high grade) according to actual condition (well, HPV, LSIL, HSIL or invasive) was particularly problematic.

Overall, the model could not be sufficiently well validated for it to be used to assess the cost-effectiveness of introducing automation.

If it is assumed that test performance is equivalent, cost-minimisation analysis can be used to suggest that if a reduction in processing times similar to that achieved for PAPNET in the PRISMATIC trial could be achieved for AutoPap GS, automated image analysis would be efficient. However, such a reduction in processing times has not been independently demonstrated (see Chapters 6–8). To reiterate conclusions from these chapters, rigorous assessment of the impact on processing times of the AutoPap GS is an urgent research priority. The cost-minimisation analysis could help in the planning of such a study.

Only a fully developed DES model will help to deal with the more difficult decisions about the cost-effectiveness of automation in combination with other related technologies such as LBC and HPV, or the cost-effectiveness of automation relative to other non-technological approaches to enhancing cervical screening programmes, such as improving coverage.

This, and its value in helping to plan other research called for in this health technology assessment, suggest that a high priority for further research should be the development of the DES model begun in this assessment.

Introduction

The objective of this component of the health technology assessment was to assess the question, 'What is the cost-effectiveness of introducing automated cervical screening devices to programmes similar to those operating, or likely to be operating in the UK?' In this, an attempt was made to integrate the information on effects and cost identified in the preceding components of the project. However, recognising that uncertainty concerning key data may be a major factor limiting conclusions, ancillary objectives were to develop a model that could incorporate information from future research, and to highlight those parameters that seem to have the greatest influence on cost-effectiveness.

As was made clear in the review of past assessments of cost-effectiveness and methodological issues arising from them (see Chapters 4 and 5), none of the existing models contained a complete representation of the current policy for surveillance screening or any allowance for competition for resources between women of different ages. Therefore, a new model was developed.

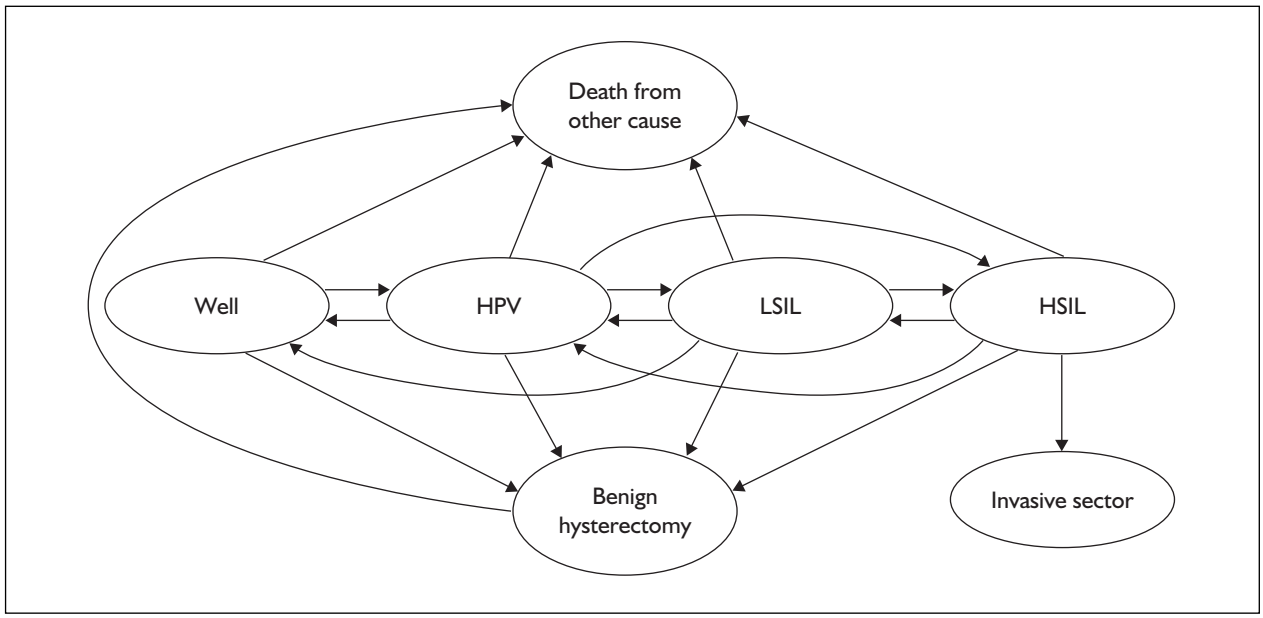


FIGURE 11 *The preinvasive sector of the natural history model*

Model for cervical cancer progression and screening

General description

The model is a DES model, written in Borland Delphi and using an event-based executive. Individual women pass through the model, their status being changed at various times according to the natural history of cervical cancer and the effects of a screening programme. The natural history sector of the model is based on that used by McCrory and colleagues (AHCPR),[12] while the screening procedure is based on a specially developed patient walk-through (see Appendix 10). The starting point for this was a translation of the NHSCSP guidance[14] into a step-by-step algorithm. Areas of uncertainty were resolved by discussion with members of the steering group or experts suggested by them.

Natural history sector

The AHCPR[12] reports a Markov model with a 1-year cycle which follows a cohort of women from 15 to 85 years of age. For the model reported here, this has been converted into a DES model. The key differences in model structure are:

- Times of events are recorded to full computer accuracy, rather than simply being regarded as taking place in a given year.
- DES allows realistic modelling of the time taken to report screening results and the possibility that this could be influenced by the number of screenings performed in a given period.

The AHCPR model contains 20 states, of which one can only be reached by screening. The remaining 19 states may conveniently be divided into preinvasive and invasive states.

The possible state transitions in the preinvasive sector are shown in *Figure 11*. Women may progress from the state 'Well' through acquisition of human papillomavirus ('HPV') to low-grade squamous intraepithelial lesion ('LSIL') and high-grade squamous intraepithelial lesion ('HSIL'). The assumption from the AHCPR model that HPV is a necessary precursor of SIL has been maintained. The transitions shown in *Figure 11* are those included in the AHCPR model. However, because of the different way of handling time in this model, as described above, it is possible to progress from Well to HSIL (or even to Invasive) within 1 year. To maintain a measure of compatibility with the AHCPR model, direct transitions between HPV and HSIL, and from LSIL to Well, have been included.

In addition, women may undergo hysterectomy for causes other than cervical cancer or SIL; such events are referred to as benign hysterectomies. Finally, all women are followed through to death. Death from other causes is possible in any state.

Figure 12 shows the transitions in the invasive sector. Entry to this sector is by progression from HSIL to undetected cervical cancer. Cervical cancer is taken to have four stages. As long as the cancer remains undetected, progression to further

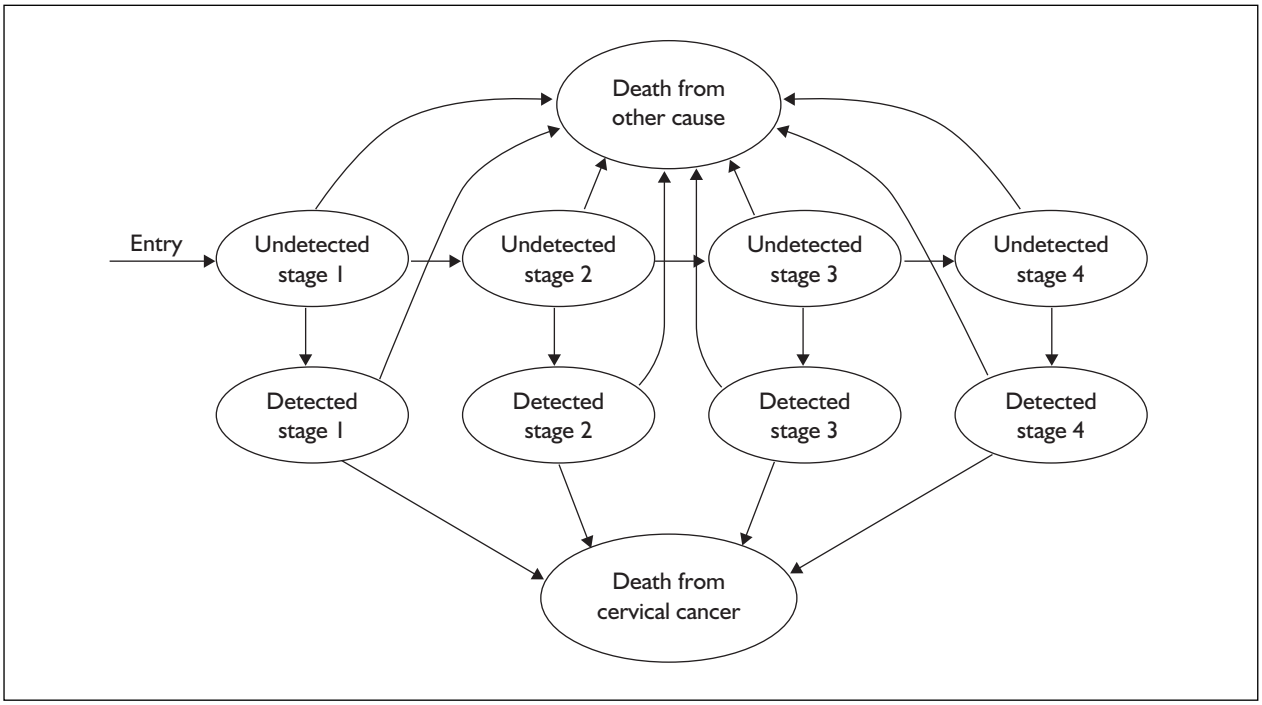


FIGURE 12 *The invasive sector of the natural history model*

stages is possible. The cancer may be detected through either screening or clinical signs. Once the cancer has been detected, it is assumed that treatment starts and that no further progression is possible. In this model, the only event following detection of cancer is death, which may be either from cervical cancer or from other causes. The AHCPR model considers all deaths more than 5 years after detection of cancer to be deaths from other causes, and includes 'cancer survivor' states for women more than 5 years after detection.

An important assumption in this part of the model is that all cancers become symptomatic before death, and are thus detected and treated. The structure in this model could allow for a difference in life expectancy according to time spent in the undetected stage, and according to whether detection was through screening or symptoms. In the absence of the necessary data, this has not been done in the version of the model presented here. It would also be possible to allow for death from cervical cancer to follow directly from an undetected cancer; again, this has not been done in this model.

Screening procedure: outline

Any individual woman will enter the screening programme at the age of 20 years with her first smear test. After any smear test, she will be waiting for the result. As long as the smears return a normal result, she will continue the routine cycle of a smear at regular intervals (in the base-case analysis of the model, every 5 years). If the smear is inadequate, or shows borderline or low-grade dyskaryosis, she will be under surveillance in the screening programme; she will then have a repeat smear after a shorter recall time. This surveillance continues until she has either a normal smear result, in which case she returns to routine screening, or sufficiently many non-normal results (details appear later in this section), in which case she will be referred for colposcopy, possibly leading to biopsy and treatment. A single smear showing high-grade dyskaryosis is sufficient for referral to colposcopy. After colposcopy and possible biopsy, a woman may return to the smear programme under surveillance, until a sufficient number of normal smears has occurred, when she returns to routine. The woman remains in the screening programme until one of the following occurs:

- death from other cause
- benign hysterectomy
- detection of cervical cancer
- reaching the age of at least 65 years in routine smearing.

Note that a woman reaching the age of 65 under surveillance remains in the screening programme as long as surveillance is thought necessary.

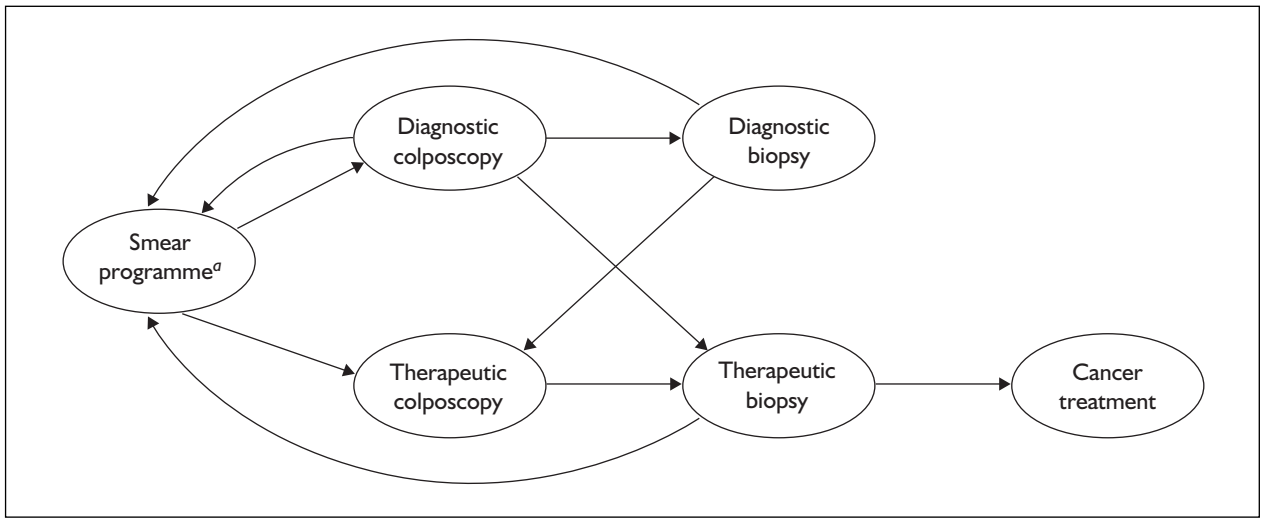


FIGURE 13 *Stages of the screening programme. [a]This includes both routine and surveillance recall.*

From a modelling point of view, the important distinction is whether the next screening event for the woman is a smear test or result, or colposcopy or biopsy. This is illustrated in *Figure 13*.

Details of the screening programme

Each time a smear is taken, the result is recorded on a scale of normal, inadequate, borderline, low grade or high grade. A high-grade smear leads immediately to referral for further investigation. Other results contribute to a cytological score; when this score reaches an appropriate level, referral for further investigation is made.

Sasieni and colleagues[144] use a simple score of no points for a normal smear, 1 point for borderline, 2 points for low grade and 3 points for high grade, with a referral threshold at 3 points. This system does not allow for referral after repeated inadequate smears, as required by NHSCSP-derived patient walkthrough. The points system in the present model allows for referral after three inadequate or borderline smears, two low-grade smears or one high-grade smear; it differs from Sasieni's model in that a borderline smear followed by a low-grade smear does not lead to referral for women with no previous diagnosis of CIN.[144] To take account of the above, the scores in *Table 66* are used.

Women returned to the smear programme after colposcopy or biopsy are under surveillance and will be referred back to colposcopy after one low-grade or two inadequate or borderline smears. This is modelled by setting the cytological score to 3 points before the first smear taken under surveillance.

TABLE 66 *Cytological scores used for deciding referral in the model*

Smear result	Cytological score
Normal	0
Inadequate	2
Borderline	2
Low-grade	3
High-grade	6

TABLE 67 *Screening times*

Parameter	Default setting
Earliest screening age	20 years
Latest screening age	65 years
Routine recall time between screenings	5 years
Possible recall times between screenings while under surveillance	1 year 6 months 3 months

The policy used by Sasieni and colleagues,[144] that two successive normal smears cancel out the existing score, was adopted. In practice, women may be referred, even after a normal smear, for reasons other than the smear result; such referrals have been excluded from the model.

The timings of the screenings are based on six parameters as shown in *Table 67*. For ease of reading, the default settings are used in the following description, although it is emphasised that they can all be changed.

Provision is made for a limited capacity at the laboratory where smear reading takes place. This is done by having a minimum time between issue of successive smear results. The minimum time is 1 year divided by the annual capacity. When a smear is taken, the time for the result is set to be either 2 weeks after the smear was taken or the minimum interval after the previous scheduled result time, whichever is the later. If a substantial queue builds up, it is possible for the woman to change state before the smear is read. The result depends on her state at the time the smear was taken, not at the time it is read.

The recall times between screenings are taken from the most recent smear or colposcopy; if the queue has become so long that the next smear is overdue at the time of issue of a smear result, the next smear is taken immediately. If a smear result leads to referral to colposcopy, this is assumed to take place 2 weeks after issue of the smear result; it is assumed that there are no delays due to shortage of capacity anywhere other than smear reading.

Individual characteristics

At any time, an individual woman in the screening programme has a cytological score and a recall programme. Initially, the cytological score is set at zero and the recall programme is set to every 5 years. The first smear takes place at the age of 20 years.

Smear testing

The result may be any of the following.

- Normal: set next smear according to the individual woman's recall programme, unless that would take the woman over the age of 65 years and in routine recall, in which case no further screening takes place. If this is the second successive normal smear, then the cytological score is set to zero.
- Inadequate: add 2 to the cytological score, and set next smear for 3 months' time, unless the cytological score is at least 6, in which case refer for diagnostic colposcopy.
- Borderline: add 2 to the cytological score, and set next smear for 6 months' time, unless the cytological score is at least 6, in which case refer for diagnostic colposcopy.
- Low grade: add 3 to the cytological score, and set next smear for 6 months' time, unless the cytological score is at least 6, or there is a previous diagnosis of CIN at biopsy, in which case refer for diagnostic colposcopy.
- High grade: refer immediately for therapeutic colposcopy.

Diagnostic colposcopy

Visual inspection may reveal the cervix to appear either normal or atypical.

- Normal: set cytological score to 3. Immediate smear is taken and referred for result as above. Two further recalls at 6-month intervals before return to routine.
- Atypical: immediate therapeutic biopsy if previous diagnosis of CIN at biopsy, otherwise referred for diagnostic biopsy.

Diagnostic biopsy

It is assumed that biopsy is 100% accurate diagnostically and that the act of taking a biopsy itself does not alter the condition of the woman, either immediately or in terms of future progression. Action taken is as follows.

- Well or HPV: set cytological score to 3, individual recall programme to one recall at 6 months, two recalls at 1 year, then return to routine. Next smear at 6 months according to this programme.
- LSIL: record diagnosis of CIN at biopsy, set cytological score to 3, individual recall programme to one recall at 6 months, two recalls at 1 year, then return to routine. Next smear at 6 months according to this programme.
- HSIL or invasive: therapeutic colposcopy.

Therapeutic colposcopy

This is the first part of a process leading to therapeutic biopsy: it is recorded as a separate stage so that diagnostic colposcopy can lead directly to therapeutic biopsy as described above.

Therapeutic biopsy

It is assumed that biopsy is 100% accurate diagnostically. If it shows cancer, then the woman is moved from undetected to detected stage *n* and treatment for cervical cancer follows.

Otherwise, the action taken is as follows, depending on the state of the woman before treatment.

- Well or HPV: set cytological score to 3, individual recall programme to one recall at 6 months, two recalls at 1 year, then return to routine. Next smear at 6 months according to this programme.
- LSIL: record diagnosis of CIN at biopsy, set cytological score to 3, individual recall programme to one recall at 6 months, two recalls at 1 year, then return to routine. Next smear at 6 months according to this

TABLE 68 *Age-related risk of death from other causes (deaths per 1000 women per year)*

Age (years)	Risk
15–20	0.3
20–25	0.3
25–30	0.4
30–35	0.5
35–40	0.8
40–45	1.3
45–50	2.1
50–55	3.3
55–60	5.4
60–65	8.7
65–70	14.9
70–75	25.9
75–80	41.9
80–85	73.2
>85	153.3

TABLE 69 *Age-related risk of benign hysterectomy (per woman per year)*

Age (years)	Risk
15–20	0.00035
20–25	0.00035
25–30	0.0026
30–35	0.0042
35–40	0.0067
40–45	0.0088
45–50	0.0069
50–55	0.0069
55–60	0.00028
60–65	0.00021
65–70	0.00021
70–75	0.00021
75–80	0.00021
80–85	0.0001
>85	0.00007

programme. Following the treatment, the woman may be in state Well, HPV or LSIL.

- HSIL: record diagnosis of CIN at biopsy, set cytological score to 3, individual recall programme to two recalls at 6 months, four recalls at 1 year, then return to routine. Next smear at 6 months according to this programme. Following the treatment, the woman may be in state Well, HPV or HSIL.

Parameters used in the model

In the DES model, ages are in fractions of years; thus, someone aged 19 years and 300 days belongs to the age group 15–20, while someone aged 20 years and 4 days belongs to the age group 20–25. Unless otherwise stated, these parameters are taken from the AHCPR model by McCrory and colleagues.[12]

- **Death from other causes** is age-related, as shown in *Table 68*. These are based on UK death rates from 1998.[145]
- **Benign hysterectomy** is age-related, as shown in *Table 69*. These are the data used in AHCPR model multiplied by 0.7. The multiplier was added to adjust for the higher levels of hysterectomy in the USA than in the UK. The multiplier was based on information from published UK estimates of hysterectomy incidence[146] and additional information from the corresponding author of this paper.
- **HPV acquisition** is age-related, is as shown in *Table 70* and based on the data used in the AHCPR model.
- **Progression from HPV** is at a fixed rate of 0.2 per HPV-positive woman per 36 months.

TABLE 70 *Age-related probability of HPV acquisition (per HPV-negative woman per year)*

Age (years)	Probability
15–17	0.1
17–18	0.12
18–19	0.15
19–20	0.17
20–21	0.15
21–22	0.12
22–24	0.1
24–30	0.05
30–50	0.01
>50	0.005

TABLE 71 *Age-related probability of regression from LSIL (per woman per 2 years)*

Age (years)	Probability of regression
15–20	0.17
20–25	0.20
25–30	0.23
>30	0.28

Progression is directly to HSIL in 10% of cases and to LSIL in the remaining 90%.

- **Regression from HPV** is at a fixed rate of 0.7 per HPV-positive woman per 12 months.[147]
- **Progression from LSIL** is set at 0.19 per woman per 2 years.[148]
- **Regression from LSIL** is as shown in *Table 71*, again using data derived from Furber and colleagues.[148] Regression from LSIL is to Well 90% of the time, and otherwise to HPV.

TABLE 72 *Stage-related probability of progression from undetected cancer (per woman per year)*

Stage	Detected	Next stage
1	0.15	0.9/4
2	0.225	0.9/3
3	0.6	0.9/1.25
4	0.9	NA

TABLE 73 *Parameters of post-cancer survival curves*

Stage	1	2	3	4
Shape	1.04	0.94	0.66	0.39
Scale	30.6	5.26	1.75	0.495

- **Progression and regression from HSIL** are not age dependent. The rate of progression is 0.4 per 12 years and the rate of regression is 0.382 per 6 years. Regression is to HPV or LSIL, each with a probability of 0.5.
- **Progression from undetected cancer** depends on stage. The cancer can become detected through clinical signs, or can progress to the next stage while remaining undetected. The probabilities are shown in *Table 72*.
- **Detected cancer** survival is stage dependent. A survival time is selected from the appropriate distribution. The one remaining event in the woman's life will be death from cervical cancer or from other cause, whichever happens first. This approach is structurally different to that used in the AHCPR model, where women surviving for more than 5 years are moved to a 'cancer survivor' state, in which state they remain until death, which is always taken to be from other causes. In the present model, death can be attributed to cervical cancer at any time after detection. However, no distinction is made between survival time after detection through screening or through the appearance of symptoms.

  The distribution used is the Weibull distribution with stage-dependent parameters as shown in *Table 73*. If the shape parameter is a and the scale parameter is b, then the probability of survival to at least time t is $\exp\left(-\left(\frac{t}{b}\right)^a\right)$. A shape parameter $a > 1$ means an increasing risk with time, while a shape parameter $a < 1$ means a decreasing risk with time. The parameters used were fitted to survival data obtained from the West Midlands Cancer Intelligence Unit.

The remaining parameters are concerned with the screening programme.

- **Smear results** require the probability of any outcome given the actual condition of the woman being tested. The table from Cuzick and colleagues[41] was used, incorporating a 10% inadequacy rate, as shown in *Table 74*. The appropriateness of the use of this table is discussed below, under the heading Validation.
- **Colposcopy** was assumed in the base-case analysis to be 100% sensitive and specific, to be varied in sensitivity analysis.
- **Outcomes from therapeutic biopsy** are as shown in *Table 75*, based on data from Hulman and colleagues[149] and Zaitoun and colleagues.[150]

TABLE 74 *Assumed outcomes for smear test*[41]

Actual condition	Probability of outcome				
	Normal	Inadequate	Borderline	Low-grade	High-grade
Well or HPV	0.882	0.1	0.009	0.0045	0.0045
LSIL	0.45	0.1	0.18	0.09	0.18
HSIL	0.36	0.1	0.135	0.18	0.225
Invasive	0.27	0.1	0.135	0.18	0.315

TABLE 75 *Outcomes from therapeutic biopsy*

State before treatment	Probability of being in given state after treatment			
	Well	HPV	LSIL	HSIL
LSIL	0.466	0.466	0.068	0
HSIL	0.31	0.31	0	0.38

TABLE 76 *Proportions of subgroups within stage 1 cancers*

Stage	Proportion
1a1	0.471
1a2	0.043
1b1	0.351
1b2	0.135

- **Cancer treatment**: stage 1 cancers are further classified into 1a1, 1a2, 1b1 and 1b2. Based on data from the West Midlands Cancer Intelligence Unit in June 2001 on cases of malignant neoplasm of cervix uteri (ICD10 C53) in 1998 and 1999, it was estimated that the proportions of the various subgroups in stage 1 were as shown in *Table 76*.

  Treatment for each type is as follows:
  – stage 1a1: LLETZ
  – stage 1a2 or 1b1: radical hysterectomy
  – stage 1b2, 2, 3 or 4: radiotherapy.
  (LLETZ is large loop-excision of the transformation zone of the cervix.)

Sample results

The model was run for a number of populations with an arrival rate of 1000 new women per year. The screening capacity was set to the smallest multiple of 1000 screens per year, which would allow the model to run without the queue becoming unacceptably long. In all the runs reported here, a capacity of 11,000 per year was used.

Results were obtained that could feed into a cost-effectiveness analysis. On the cost side, the number of each procedure carried out in each year of simulated time was collected. For effectiveness, in the absence of any data on quality of life the model worked simply in terms of life-years saved. In this model, the age at which each woman would have died from other causes was known. Each time a death from cervical cancer occurred, the years of life lost could be determined.

Because the model is a stochastic model, the results will be slightly different each time that different sets of random numbers are used. In addition, the model starts empty, so it is necessary to run for a warm-up period until it reaches a steady state. The figures reported in each table are averages per year over three runs each of 100 (simulated) years' running of the model after discarding a warm-up period of 100 years. Quasi-standard errors for the difference are shown. These could be reduced by increasing the number of runs of the model.

For a base-case run, the sensitivity of AutoPap was speculatively entered as being 10% higher than conventional testing, conditional on an adequate smear. It is acknowledged that the true value is likely to be no difference, but the value of entering no difference into the model to explore its performance was also a consideration. The results of the base case are shown in *Table 77*.

As an example of a sensitivity analysis, the sensitivity and specificity for diagnostic colposcopy

TABLE 77 *Sample base-case results from the model*

	No AutoPap	With AutoPap	Difference	QSE (diff.)
Screenings	10,009.8	10,063.0	53.2	8.3
Normal/inadequate	9,722.2	9,758.8	36.6	7.5
Borderline/low/high	287.6	304.2	16.7	1.5
Colposcopies	157.6	170.1	12.5	1.1
Diagnostic biopsy	11.4	13.3	1.9	0.3
Therapeutic biopsy	104.7	114.1	9.4	0.9
Cancer treatments:				
Stage 1a1	1.987	1.670	–0.317	0.106
Stages 1a2 and 1b1	1.620	1.413	–0.207	0.106
Stages 1b2, 2, 3 and 4	3.560	2.693	–0.867	0.150
Life-years lost[a]	215.2	168.4	–46.8	7.5

QSE, quasi-standard error.
[a] Each time a death from cervical cancer was recorded in the model, the number of life-years lost for that woman could be found by subtracting her age at death from the age at which she would have died from other causes. This approach has the advantage over measuring average age at death that variation in normal lifetime is eliminated from the variance of the estimate, thereby increasing the precision of the result from a small number of runs of the model.

TABLE 78 *Sample sensitivity analysis result*

	No AutoPap	With AutoPap	Difference	QSE (diff.)
Screenings	10,026.2	10,082.0	55.8	8.7
Normal/inadequate	9,736.7	9,773.2	36.4	7.9
Borderline/low/high	289.5	308.9	19.3	1.4
Colposcopies	159.4	172.3	12.9	1.1
Diagnostic biopsy	15.1	18.7	3.6	0.3
Therapeutic biopsy	105.0	114.0	9.0	0.9
Cancer treatments:				
Stage 1a1	2.007	1.650	–0.357	0.111
Stages 1a2 and 1b1	1.637	1.450	–0.187	0.105
Stages 1b2, 2, 3 and 4	3.297	2.530	–0.767	0.141
Life-years lost[a]	205.4	170.8	–34.6	7.5

[a] Each time a death from cervical cancer was recorded in the model, the number of life-years lost for that woman could be found by subtracting her age at death from the age at which she would have died from other causes. This approach has the advantage over measuring average age at death that variation in normal lifetime is eliminated from the variance of the estimate, thereby increasing the precision of the result from a small number of runs of the model.

were both reduced to 80%. The results are shown in *Table 78*. Compared with *Table 77*, the differences are within the range of random error. This is not surprising, as the number of diagnostic colposcopies is quite small.

Validation

The validity of a model may be tested by comparing the output from the model with data about the 'real' system being modelled. One comparison that can be made is the age-related incidence of cervical cancer. Yearly rates for incidence and detection of cancer are plotted in *Figure 14* along with English cancer registry data for 1998,[151] which are provided in 5-year age bands. As can be seen, detection has sharp peaks at the standard screening ages. A better comparison can be seen in *Figure 15*, where the detection rate in the model is adjusted to 5-year bands.

Comparing the model output with the English data, it is clear that cancers are occurring in the model at a somewhat earlier age than in reality. However, the comparison at least matches the general shape in that there is a peak age of incidence. It is possible that changes in sexual behaviour mean that the model is a better predictor of the future age of peak incidence than the current data. In comparison with other UK models, Payne and colleagues[38] had a steadily increasing incidence throughout the screening age range, and Cuzick and colleagues[67] did not give a comparison, stating: "Validation studies were not in the scope of this review project".

Another aspect of validation is to compare the output from the model with available data on screening results. *Table 79* shows the reported number of smear results for England in 1999–2000.[152]

Combining this with the best available data for predictive value of a smear gives the proportion of all smears that have any given combination of actual state and smear result, as shown in *Table 80*. The data sets used were as published by Jones and colleagues[153] for borderline and low-grade smears and KC61C returns to the Department of Health (obtained via Statistics Division 2B) for high-grade smears.

If accurate and complete values for *Table 80* were available, they would allow a replacement for *Table 74* above. This would be done by scaling the numbers in each row to add to 100%. Note that the borderline and low-grade results did not include any cases of invasive cancer. Clearly, a very large sample would be necessary for these numbers to be known accurately enough to be sensibly scaled.

The validation issue here is that the prevalence of actual states at smear should be in line with *Table 80*. *Table 81* shows the prevalence of states at smear in the model, compared with the minimum required by *Table 80*. The actual prevalences must be higher because of inadequate and negative smears. For example, more than 1.89% of smears taken in reality come from women who have HSIL, whereas only 1.01% of smears taken in the model

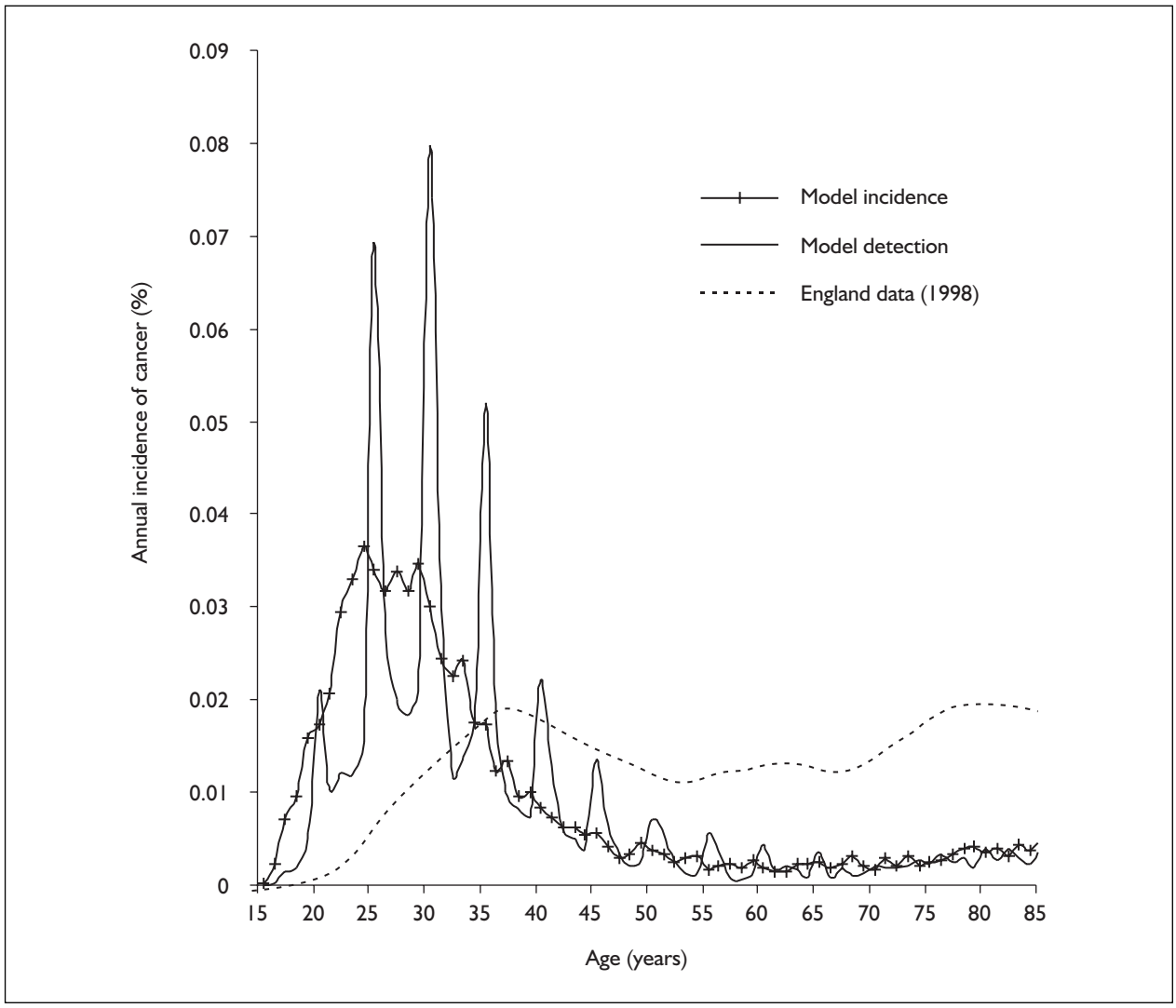


FIGURE 14 *Age-related incidence and detection of cancer*

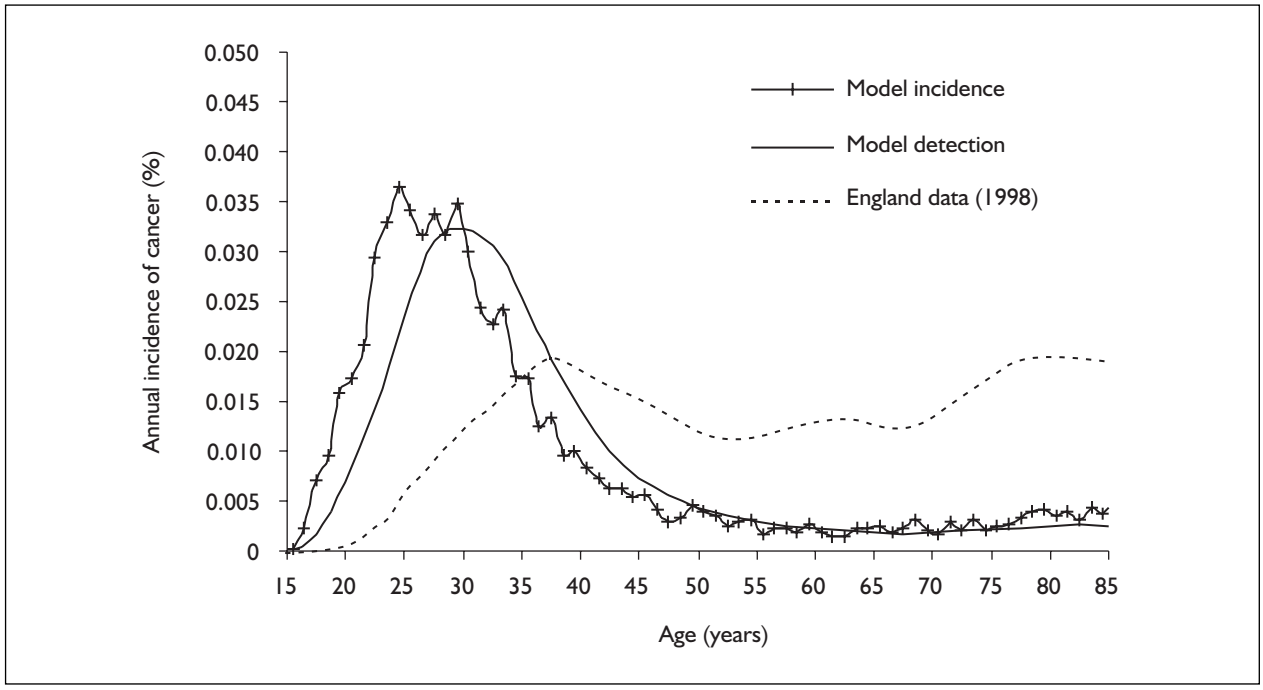


FIGURE 15 *Age-related incidence and detection of cancer (model rates by 5-year bands)*

TABLE 79 *Reported data of actual smear results for England, 1999–2000*[152]

Smear result	Number	%
Negative	3,491,822	81.98
Inadequate	414,210	9.72
Borderline	183,475	4.31
Low-grade[a]	103,510	2.43
High-grade[b]	66,437	1.56
Total	4,259,454	100.00

[a] Low-grade: 'mild dyskaryosis'
[b] High grade: 'moderate dyskaryosis' + 'severe dyskaryosis' + '? invasive carcinoma' + '? glandular neoplasia'.

come from such women. It is clear that the model is underestimating the prevalence of states from LSIL upwards.

It would be possible to adjust the risk of progression in the model to ensure that the model reported acceptable prevalences for SIL. However, this would involve making arbitrary assumptions about the distribution of increases that would be made, and in any case would be dependent on the assumptions for HPV that have already been discussed. Thus, there is no guarantee that such adjustment would have any useful predictive value.

In summary, there was sufficient concern about the validation of the model to indicate that proceeding to use it to estimate cost-effectiveness was unreasonable. The problems experienced may relate to the use of data for particular parameters that may not be entirely appropriate. For instance, data on HPV incidence and progression tend to be derived from higher risk populations that may not be typical of the screened general population whose cancer incidence and smear results the model is attempting to replicate. Further, the model is primarily trying to assess the cost-effectiveness of introducing automation in the general population.

Capability of the model

This model is an improvement on previous models in that it does not impose specific shapes on the survival times in any given state. It also allows for inadequate smears and takes realistic account of the timing of repeat smears. It is thus able to accommodate guidelines for repeat and surveillance screening.

As it stands, the model largely uses data collected for models that have more restrictive assumptions placed upon them. In principle, the model can

TABLE 80 *Percentages of smears by actual state and smear result*

Actual state	Smear result				
	Negative	Inadequate	Borderline	Low-grade	High-grade
Cancer			0.00	0.00	0.07
HSIL			0.40	0.42	1.07
LSIL		Not known	0.63	0.78	0.19
HPV			0.95	0.39	0.06
Well			2.33	0.83	0.17
Total	81.98	9.72	4.31	2.43	1.56

TABLE 81 *Percentage prevalence of states at smear: model versus reality*

State at smear	From model	Required[a]
Cancer	0.08	0.07
HSIL	1.01	1.89
LSIL	1.24	1.61

[a] Numbers in this column were obtained by summing the numbers from the relevant row of *Table 80*. They represent the proportion of total smears from women who are in the indicated state and give a result of borderline or higher. They are thus lower bounds for the proportion of women who are in the indicated state.

allow changes in a woman's condition to be dependent on previous history in any way that can be specified. In particular, annual rates of progression can depend on any combination of the woman's age, how long she has been in her current state and how long she was in any previous state.

One particular matter that would be usefully addressed is the issue of survival times from detection of cancer. As the model stands, a woman with undetected stage 4 cancer will have her life expectancy reduced by earlier detection of the cancer, since survival time is effectively measured from detection. This only has a minor effect on the model. Earlier stages of cancer are also affected by this problem, but it is offset by the possibility that detection through screening could prevent progression to later stages.

Other possibilities for future development of the model include assessment of the transient effects of a change in screening policy. The model can be set to reach steady state under one screening policy and then introduce a new policy from a given calendar date. The short-term effects and time taken to reach a steady state can be shown. The model could also allow for a growing population by increasing the number of women entering the model each year. This would require some method for systematically increasing the screening capacity with the population if waiting times for screening results are not to increase without limit.

Conclusions concerning DES model

To test the possibility that automation may change the test characteristics in any way requires a model that adequately reflects the realities of the screening programme. A model has been constructed that could be developed for this purpose, given adequate data. Such a model would also be able to address other questions with regard to the cervical screening programme, and therefore its development should be given a high priority for further research.

Concerning this project, this still leaves the question of cost-effectiveness unaddressed, as it was intended that the *de novo* DES model would be used to assess this. However, given that the evidence that automation makes any difference to the test characteristics of Pap testing is weak, simpler approaches to assessing the relationship between cost and benefit/disbenefit of introducing automation may be appropriate as an alternative. Such an approach is developed in the remaining paragraphs of this chapter.

Alternative approach to the assessment of efficiency

Main potential benefits and costs of introducing automated image analysis

Reference to Chapters 6 and 8 identifies the main potential benefits and costs to be:

- improved test performance: fewer false-negative and false-positive diagnoses, which should in turn translate into improved health outcomes
- reduced processing times, leading in turn to reduced costs or increased capacity
- costs associated with purchasing, running and maintaining the automated image analysis devices.

With three main components, and particularly with a desire to indicate what improvements in test performance might mean in terms of health outcomes, simulation modelling was required. However, as indicated in the section on effectiveness, the available data seem most compatible with equivalence test performance of automated image analysis devices with existing manual screening. Although the need to confirm that this is the case for the newer automated image analysis devices remains (and indeed the need to investigate properly whether there may be improvements), acknowledgement that equivalence is an increasingly reasonable assumption suggests that testing whether the costs can be justified in terms of reduced processing time alone would be useful. Simple cost-minimisation is all that is required to achieve this.

Cost minimisation analysis for automated screening

Assuming that there is equivalence in test performance between automated image analysis incorporated into a manual screening system and a manual screening system alone, the benefit of automation may be calculated by comparing the value of screeners' time saved to the capital and running costs of the machine.

Using data from earlier in the report (see *Table 53*), the average screeners' time for each type of smear is as shown in *Table 82*. The data are derived from the PRISMATIC study, and so it is assumed that reductions in average processing time demonstrated for PAPNET are also true of AutoPap GS.

The relevant unit labour costs (at 1999/2000 prices) are £27 per hour for technicians and £82 per hour for pathologists.[154] The equipment costs were based on the data presented in Chapter 8.

TABLE 82 *Average screeners' time per smear with and without automation*

		Without automation[a]		With automation[a]	
		Screeners' time (minutes)		Screeners' time (minutes)	
Smear result	**%[b]**	**Technician**	**Pathologist**	**Technician**	**Pathologist**
Negative/inadequate	91.70	10.4	0	3.4	0
Borderline or worse	8.30	7.4	5.4	1.4	5.1
Average		10.15	0.45	3.23	0.42

[a] Data taken from the PRISMATIC trial, where automation was PAPNET used in a primary screening mode.
[b] Percentage of all smears giving the indicated result (*Table 80*).

TABLE 83 *Details of cost-minimisation analysis*

	Minutes per slide	Cost per hour	Cost per slide
Without automation			
Technician	10.15	£27	£4.57
Pathologist	0.45	£82	£0.61
Total cost per slide			£5.18
With automation			
Technician	3.23	£27	£1.46
Pathologist	0.42	£82	£0.58
Machine cost per slide			£2.66
Total cost per slide			£4.69

This analysis assumed the higher cost configuration at £664,830 over 5 years.

Assuming that the machine operates at a full capacity of 50,000 slides per year, the net saving associated with the use of automation is estimated at approximately 49p per slide or £24,400 per year. The details of the calculations are shown in *Table 83*.

The above calculations assumed that the machine would be used to full capacity. It is unlikely that any region would have a demand exactly matching the capacity of one or more machines. The annual net saving associated with AutoPap varies with the achieved throughput, as shown in *Figure 16*.

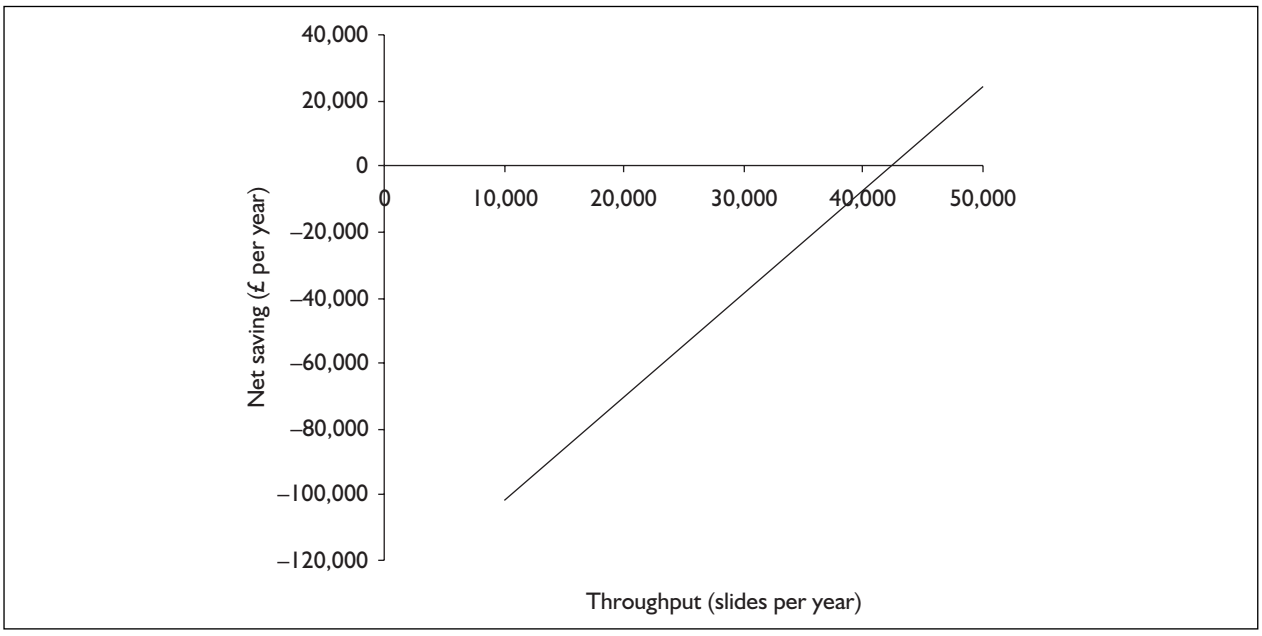


FIGURE 16 *Relationship between net savings associated with introducing automation and throughput*

Conclusions concerning efficiency of introducing automation

If alteration of test performance is felt to be a critical component of the cost–benefit relationship, cost-minimisation is inappropriate, and from the perspective of this report, assessment of cost-effectiveness should await the further validation of the DES model.

If, however, the test performance of automation is felt to be equivalent to that of manual screening alone, then the cost-minimisation analysis suggests that the additional cost of automation is justifiable, with the following key provisos.

- The equivalence of test performance for AutoPap GS has not yet been formally demonstrated in a peer-reviewed publication. This should be substantiated.
- The reduction in test processing time is assumed to be the same as that demonstrated for PAPNET; although greater improvements are claimed by the manufacturers for AutoPap GS, these claims need to be independently substantiated using rigorous methods.
- Other potential costs, such as locating the automated devices and the knock-on effects on other components of the cervical screening programme are assumed to be negligible; this ought to be tested.
- A single AutoPap GS machine must be able to process reliably at least 45,000 smears per annum. The processing failure rate of approximately 7% noted in Chapter 6 needs to be considered in this respect.
- The device costs remain similar to those quoted in August 2001.

Chapter 10

Discussion and overall conclusions

Main findings

Given the scope of this report, there are a great many findings at various levels of detail that could be mentioned in any summary. The following are those that have been particularly influential in shaping the conclusions.

Automated image analysis is not a homogeneous technology. From a theoretical perspective there were important differences between different competing alternatives, especially PAPNET and AutoPap. However, these differences are now of largely academic importance, as there is only one currently commercially available device, the AutoPap GS System. Independently of device type, whether automated image analysis is used to replace the quality assurance or the primary screening step is also likely to make a difference on impact; the AutoPap GS System mainly aims to replace the primary screening step. Finally, it is undoubted that there has been development in the systems since they were introduced and that systems marketed now bear little relationship to those introduced in the early 1990s.

The implementation costs of the only commercially available system would be considerable. The device alone costs in excess of £0.5 million (2001). With the manufacturer's estimated annual capacity of 50,000 slides, a very conservative estimate of the device purchase price alone to process the number of slides received by the NHSCSP (approximately 4 million) would be £40 million.

The overwhelming finding from the systematic reviews was the very limited amount of rigorous primary research available, particularly relative to the complexity of the decision attempting to be addressed. However, the generalisability of this research is also important, because the assessments of many dimensions of impact will be dependent on the exact nature of the 'normal' manual screening with which the system enhanced with automated image analysis is being compared. The 'normal' systems differ quite markedly from one country to the next. Taking this into account, the amount of rigorous primary research available to inform decisions appropriately in any particular national system, including the NHSCSP, is even more limited.

Concerning impact on test performance, two studies were initially included, which together considered just over 13,000 slides. Because the authors were not convinced that all the original inclusion criteria were acting as good indicators of true study quality, and to maximise the opportunity to capture all potentially valuable research information, the inclusion criteria were relaxed slightly. This allowed consideration of a further five studies in a sensitivity analysis. The seven studies considered in total in this analysis represented approximately 64,000 slides. The pronounced variability in what was being assessed in each of the included studies (clinical heterogeneity), the threats to internal validity identified in the quality assessment process, and the actual observed variation in the sensitivity and specificity results made the interpretation of the results difficult. Debatably, they are most compatible with automated image analysis having similar sensitivities and specificities to manual screening. However, potentially important improvements are not excluded; nor indeed is deterioration, although this is less likely than improvement.

Concerning reproducibility, six studies provided information on interobserver, intraobserver and machine variability. The most notable feature about the results was that kappa values less than 0.4 (the cut-off between 'fair' and 'poor' agreement) were sometimes obtained.

Two studies, together involving over 1.5 million slides, were identified which could be used to assess the impact of introducing automated image analysis on health outcomes, particularly numbers of invasive cancers. One was a probable RCT and the other a pre–post design. Threats to validity limited their contribution to the conclusions of this particular report, but the potential value of such an approach, if properly conducted, was highlighted.

Information was identified from a larger number of studies on various aspects of impact on process, including technical rejection rates. This was

approximately 2% for PAPNET and 7% for AutoPap machines. Of particular importance were five studies contributing information on the impact on processing times. All showed reductions associated with automation. The most robust evaluation on over 20,000 slides, directly generalisable to the NHSCSP, was provided by the PRISMATIC trial, where the average slide processing time was reduced from 10.4 minutes to 3.9 minutes using PAPNET. This difference was statistically significant.

The major proviso to all the findings reported above is that none refers to the AutoPap GS System. The vast majority of the results refer to PAPNET. With the exception of the data on technical rejection rates, only one study on AutoPap contributed data on test performance, no studies on reproducibility, one study on health outcomes and no studies on test processing times. All of the AutoPap studies dealt with predecessors of the currently available device (AutoPap Primary Screening System for test performance and AutoPap QC for health outcomes).

No studies were identified which considered the effects of automated image analysis in combination with either LBC or HPV screening.

The detailed search for unpublished research on UK-based assessments of test performance revealed 13 studies whose nature could be verified. Although several of these were clearly in-house evaluations, many were substantial, involving considerable amounts of investigator time and effort. Although by definition the nature of the information available makes it impossible to know with absolute certainty, there were two studies in particular (one on PAPNET, with around 13,000 slides; and one on AutoPap GS System, with around 6000) of a quality that might have meant that they could have been considered in the systematic review of test performance. This needs to be seen in the context of the fact that there were no UK-based studies in the main analysis, and just one such study in the sensitivity analysis of the systematic review of test performance. These findings indicate that there may be a substantial unpublished literature, and that particularly if considered on a worldwide basis there could be a number of additional unpublished studies for inclusion in the systematic reviews reported. If publication bias applies to research on automated image analysis and operates in the established manner, the finely poised nature of the published findings means that locating unpublished studies and incorporating them is very important.

Systematically identified data on cost showed a high degree of variability and point to the need to consider carefully the costing method adopted and the scope of costs that are being captured. Costs associated with housing automated image analysis devices and costs arising from knock-on effects, such as changes in the need for colposcopy, are rarely considered in costing exercises. Cost savings resulting from reduced labour costs are claimed by the manufacturers of AutoPap, yet there are no rigorous published data to support the size of reduction in average slide processing time beyond PAPNET.

A *de novo* DES model of cervical screening was developed. Its potential to deal with scenarios beyond the scope of other approaches based on population cohorts was demonstrated. Unfortunately, the authors were not satisfied with its validation and did not proceed to use it to estimate the cost-effectiveness of introducing automation. Although the data used for certain parameters are likely to have been the main problem, time constraints meant that this could not be investigated and solutions could not be developed.

Assessment of efficiency instead used a cost-minimisation analysis. This assumes that the effectiveness of an automated image analysis-enhanced screening system and an existing manual system is equivalent. This seemed a reasonable assumption from the systematic review of effectiveness, although noting that this equivalence needs to be verified for the AutoPap GS System. The cost-minimisation analysis showed that with a device cost of £644,000, a capacity of 50,000 slides per annum and improvement in average slide processing times similar to that obtained by PAPNET in the PRISMATIC trial, automated image analysis is efficient (net cost per slide is less for automated image analysis). Costs associated with housing the automated system and knock-on effects were not considered.

Findings in the context of past health technology assessments

Considerable efforts were made to develop this assessment to build on the methods and findings of previous reports. As summarised in the earlier chapters of this report, the findings of those reports adopting the most rigorous methods, without conflicts of interest,[12,52,53,62,63] have universally found there to be no convincing evidence on effectiveness, particularly the test

performance of automated image analysis systems. Inevitably, the uncertainty about effectiveness is reflected in considerable uncertainty about associated estimates of cost-effectiveness. However, the general sentiment has been that the technology is at best at the margins of what is normally considered cost-effective.

As in earlier health technology assessments, this report indicates that the main limitation on conclusions remains the lack of rigorous evidence. However, the amount of evidence identified and considered here is much greater than before. This is primarily because several important studies have been published in recent years. This is not the only explanation, however, and in particular one advance offered over previous health technology assessments is the broadening of the scope of effects considered appropriate in attempting to weigh benefits against costs. This has resulted in systematic reviews of information on reproducibility, health outcomes, technical rejection rates and processing times, aspects of the clinical effectiveness of automated image analysis that have not previously received much attention. Information on the last issue has been particularly important, and has reinforced that automated image analysis may be justifiable on the basis of improvements in process alone, provided that test performance is demonstrated to be equivalent. This is a perspective that is again not apparent in previous assessments. Other important differences in the conduct of the review that may account for any differences in the conclusions are, first, that this is the only health technology assessment to consider explicitly and attempt to address publication bias and, second, that the report was conducted at a time when one dimension of complexity had been removed. The nature of the intervention available has now resolved to one system, the AutoPap GS System, by a combination of commercial pressure and genuine development of the systems in question. Earlier reviews were undoubtedly made more difficult by different devices used in different modes being real policy alternatives at the time the reviews were conducted.

Conclusions

Although the present findings are ostensibly similar to past reviews in highlighting the limited available evidence, they differ in that the general case for automated image analysis may now just have been made. As indicated above, the best grounds for this are likely to be equivalence of test performance combined with the possibility of important reductions in average slide processing times. The latter may be desirable in its own right (because it is believed that it is almost impossible to recruit and retain sufficient screeners to process the growing numbers of slides) or because it may result in the cost of the system actually being lower. Even with this argument there are still many uncertainties, which should be the target for further research:

- Is the case for equivalence sufficiently proven?
- Is the validity of the estimate of reduction in slide processing time adequate?
- By how much are the true costs of introducing automated image analysis likely to be underestimated by ignoring the costs of housing the devices and not estimating knock-on costs?

However, the one area of uncertainty that suggests that the appropriate recommendation for further action is research rather than implementation is that most of the evidence considered in the review relates to PAPNET, not the current commercially available device, the AutoPap GS System. Such evidence is claimed to exist, but it has certainly not been fully published. Even if it had been, given the concern about conflict of interest, heightened by the findings on the volume and nature of unpublished literature, there should be great reluctance about basing an implementation decision with major financial consequences on findings that have not been corroborated by independently conducted research, meeting appropriately rigorous methodological standards. In conducting future research, greater involvement from public funding bodies and a system of prospective registration for any new studies on the AutoPap GS System should be sought to reduce the likelihood of publication bias and improve confidence in findings. The need to be able to generalise the research findings to a range of different existing systems also needs to be considered.

Implications for patients and practice

The main implications are for those making policy and those funding research. However, women in cervical screening programmes and healthcare professionals involved in delivering them should be encouraged to support the additional research advocated on this topic. Information on the acceptability of automated devices to patients has been raised as an important issue in the past, which certainly does not seem to have been

addressed thus far. Arguably, until it was clear exactly what automation was and what it might realistically achieve, assessing acceptability may have been premature. Given that greater clarity has emerged, and will hopefully continue to emerge, the time may be approaching where formal assessment of acceptability by women is appropriate.

Implications for further research on automation in cervical screening

There is a major mismatch between the complexity of the problem that this study attempted to address and the amount of rigorous research available to help answer it. On these grounds virtually any high-quality research on any aspect of automated cervical screening could be justified.

However, in the authors' view the areas of greatest priority are:

- Clinical effectiveness of the AutoPap GS System relative to existing cervical screening programmes: this should include reproducibility, impact on test performance, impact on processing times, impact on costs and ideally the impact on health outcomes of introduction.
- Further development of the DES model presented in this report, particularly its validation: the DES model could help with the planning of many of the components of the clinical effectiveness research programme, and be the main tool for the further assessment of cost-effectiveness.
- Further assessment of cost-effectiveness: the cost-effectiveness of other technological and non-technological approaches should be considered alongside the cost-effectiveness of introducing the AutoPap GS System.
- Further research on the effectiveness and costs of these other approaches: this report notes in particular a complete absence of research on the effectiveness and costs of automated image analysis combined with LBC and HPV screening.

As already noted from the point of view of avoiding publication bias, public research funding bodies may need to consider taking a greater lead in future research to ensure its independence and methodological rigour. In any event, automated image analysis has arguably now reached a stage in its development where the likelihood that it represents an effective and efficient intervention is sufficiently great that the responsibility for supporting further research should rest as much with the public as with the private sector.

Implications for methodological research

Several issues were identified in the course of this work where there did not seem to be clear advice in the methods literature on how to proceed. All of these areas could undoubtedly benefit from further methodological research; however, the authors would prioritise the following.

- Research on the advantages and disadvantages of different research designs assessing test performance of screening or diagnostic tests: two-armed designs, where both a test of interest and an existing test are compared against a reference standard, was a design that had not been encountered to a great extent previously. Although they have been criticised, their role does not seem to have been fully assessed, and if properly conducted they may offer advantages for the many tests where no clear gold standard is available.
- Research on the conduct of systematic reviews of dimensions of the impact of screening and diagnostic tests, other than test performance: reviewing assessments of reproducibility would be of particular interest. Such reviews are rarely conducted, yet interpretation of the results of test performance assumes some knowledge of reproducibility, and appraisal checklists frequently refer to such knowledge. An important part of the problem may be that there is no guidance on the topic, yet there are several real problems. Not least of these is that the interpretation of one of the most commonly used methods of expressing reproducibility, Cohen's kappa, is not straightforward. Reviewing assessments of impact on process is also an area where better guidance could be developed, particularly as there are likely to be other new tests whose introduction could be justified on the basis of equivalent accuracy, but reduced time or inputs to achieve the results.
- Further research on publication bias: in this project, in response to a strong perceived threat of publication bias, a detailed survey of unpublished literature was undertaken. A priori, it was felt that this was the only way to proceed, anticipating that the available literature would

be insufficient to use techniques such as funnel plots. Guidance on when such an approach may be of assistance, what it may achieve and how to conduct it efficiently appears to be absent. A closely related issue that also needs further investigation is how the original concepts of publication bias are affected by the increasing range of publication options in the early twenty-first century, that is, does publication bias exist if a definition of publication/non-publication more relevant to the current day is applied?

When to repeat this health technology assessment

This assessment was completed in April 2002. Further health technology assessments and systematic reviews should be undertaken when primary research of the type indicated above on automation in cervical screening has been completed, or in 2010, whichever is sooner. Future reviews should focus on the currently commercially available automated image analysis device, AutoPap GS System and its successors. The authors anticipate there would be little value in revisiting the literature on other devices, such as PAPNET and earlier versions of the AutoPap device.

Acknowledgements

The authors particularly acknowledge the following: Dr Gill Lawrence (West Midlands Cancer Intelligence Unit), Dr David Poller (Department of Pathology, Queen Alexandra Hospital, Portsmouth), Dr Christine Waddell (Cytology Laboratory, Birmingham Women's NHS Trust), Dr Angela Raffle (Department of Public Health, Avon Health Authority, Bristol) and Dr Richard Winder (NHSCSP, Department of Health). They acted as a steering group during the course of the project, providing advice on its conduct. They also helped with providing specific pieces of information and contacts as requested by the main project worker. All members of the steering group commented on drafts of the report.

The authors also gratefully acknowledge the information provided by the staff of the West Midlands Cancer Intelligence Unit and the Department of Health Statistical Division, the goodwill of reviewers and investigators of past work on this topic who responded to further enquiries, and all those who responded to the surveys for unpublished information.

Ann Massey and Anne Fry-Smith provided invaluable support with respect to administration of the project and the conduct of literature searches, respectively.

This report was commissioned by the NHS R&D HTA programme. The report and the ideas therein are the responsibility of the authors alone.

Contributions of authors

Brian H Willis (Senior House Officer, Medicine) was the main project worker, and the lead author for all chapters except for 4, 8, 9 and 10. Pelham Barton (Lecturer in Mathematical Modelling) was the main worker on the simulation model and lead author for Chapter 9. Philippa Pearmain (Deputy Director of Cervical Screening Quality Assurance) was the advisor on cervical screening, and had particular input on Chapters 1, 3 and 9. Stirling Bryan (Professor of Health Economics) was the supervising health economist and lead author for Chapter 4. Chris Hyde (Senior Lecturer in Public Health) was the overall supervisor and guarantor of the project and lead author for Chapters 8 and 10. All authors have read and contributed to changes to the final draft of the report. All authors agree with the overall conclusions and executive summary.

References

1. Secretary of State for Health. *Saving lives: our healthier nation*. London: HMSO; 1999.
2. Office for National Statistics. *Mortality statistics; cause, 1997*. Series DH2, No. 24. London: The Stationery Office; 1999.
3. Fidler HK, Boyes DA, Worth AJ. Cervical cancer detection in British Columbia. A progress report. *Journal of Obstetrics and Gynaecology in the British Commonwealth* 1968;**75**:392–404.
4. Van Oortmarssen GJ, Habbema JD. Epidemiological evidence for age-dependent regression of pre-invasive cervical cancer. *Br J Cancer* 1991;**63**:559–665.
5. Sasieni P, Adams J. Effect of screening on cervical cancer mortality in England and Wales: analysis of trends with an age period cohort model. *BMJ* 1999;**318**:1244–5.
6. Quinn M, Babb P, Jones J, Allan E. Effect of screening on incidence of and mortality from cancer of the cervix in England: evaluation based on routinely collected statistics. *BMJ* 1999; **318**:904–8.
7. Government Statistical Service. *Cervical screening programme, England: 2000–01*. London: Department of Health; September 2001.
8. Papanicolaou GN, Traut NF. *Diagnosis of uterine cancer by the vaginal smear*. New York: Commonwealth Fund; 1943.
9. Guidozzi F. Screening for cervical cancer. *Obstetrical and Gynaecological Survey* 1996; **51**:247–52.
10. Shingleton, HM, Orr JW Jr. *Cancer of the cervix*. Philadelphia, PA: JP Lippincott; 1995.
11. Evans DM, Hudson AE, Brown CL, Boddington MM, Hughes HE, MacKenzie EF, *et al*. Terminology in gynaecological cytopathology: report of the working party of the BSCC. *J Clin Pathol* 1986;**39**:933–44.
12. McCrory DC, Mather DB, Bastian L, Datta S, Hasselblad V, Hickey J, *et al*. *Evaluation of cervical cytology*. Evidence Report/Technology Assessment No. 5. AHCPR Publication No. 99-E010. Rockville, MD: Agency for Health Care Policy and Research; February 1999.
13. NHS Cervical Screening Programme. *Cervical screening. A pocket guide*. Sheffield: NHS Cervical Screening Programme; November 1996.
14. Report of a Working Party set up by RCPath, BSCC and NHSCSP. *Achievable standards, benchmarks for reporting and criteria for evaluating cervical cytopathology*. NHSCSP Publication No. 1. 2nd ed. Sheffield: NHS Cervical Screening Programme; May 2000.
15. Eddy DM. Screening for cervical cancer. *Ann Intern Med* 1990;**113**:214–26.
16. Duncan ID, (editor). *Guidelines for clinical practice and programme management*. NHSCSP Publication No. 8. 2nd ed. Sheffield: NHS Cervical Screening Programme; December 1997.
17. Report of a Working Party convened by the NHSCSP and chaired by Dr John Pritchard. *Quality assurance guidelines for the cervical screening programme*. NHSCSP Publication No. 3. Sheffield: NHS Cervical Screening Programme; January 1996.
18. Hutchinson ML. Assessing the costs and benefits of alternative rescreening strategies. *Acta Cytol* 1996;**40**:4–8.
19. Baker RW, O'Sullivan JP, Hanley J, Coleman DV. The characteristics of false negative cervical smears – implications for the UK cervical screening programme. *J Clin Pathol* 1999; **52**:358–62.
20. Petticrew MP, Sowden AJ, Lister-Sharp D, Wright K. False-negative results in screening programmes: systematic review of impact and implication. *Health Technol Assess* 2000;**4**(5).
21. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. *Am J Epidemiol* 1995;**141**:680–9.
22. Poller DN, Willis BH, Codling BW. Search tasks in human visual perception: relevance to diagnosis of cervical small cell and pale dyskaryosis and other cytology specimens. *Acta Cytol* 1996;**40**:851–3.
23. Treisman A, Gormican S. Feature analysis in early vision: evidence for search asymmetries. *Psychol Rev* 1988;**1**:15–48.
24. Koss LG, Lin E, Schreiber K, Elgert P, Mango L. Evaluation of the PAPNET cytologic screening system for quality control of cervical smears. *Am J Clin Pathol* 1994;**101**:220–9.
25. Anderson TL. Automated screening of conventional Papanicolaou smears: the AutoPap 300 and AutoPap 300 QC Systems. *Tutorials of Cytology* 1994;306–11.

26. Howell LP, Beln T, Agdigos R, Davis R, Lowe J. AutoCyte Interactive Screening System. Experience at a University Hospital Cytology Laboratory. *Acta Cytol* 1999;**43**:58–64.

27. Banda-Gamboa H, Ricketts I, Cairns A, Hussein K, Tucker JH, Husain N. Automation in cervical cytology: an overview. *Anal Cell Pathol* 1992;**4**:25–48.

28. Prewitt JM, Mendelsohn ML. The analysis of cell images. *Ann N Y Acad Sci* 1966;**128**:1035–43.

29. Taylor J, Bahr GF, Bartels PH, Bibbo M, Richards DL, Wied GL. Development and evaluation of automatic nucleus finding routes and thresholding of cervical cell images. *Acta Cytol* 1975;**19**:289–96.

30. Liedtke CE, Gahm T, Kappei F, Aeikens B. Segmentation of microscopic sciences. *Anal Quant Cytol Histol* 1987;**9**:197–211.

31. Cahn RL, Poulsen RS, Toussaint G. Segmentation of cervical cell images. *J Histochem Cytochem* 1977; **25**:681–7.

32. Holmquist J, Bengtsson E, Eriksson O, Nordin B, Stenkvist B. Computer analysis of cervical cells, automatic feature extraction and classification. *J Histochem Cytochem* 1978;**26**:1000–17.

33. http://www.prnewswire.com/comp/110214.html

34. http://www.tripathimaging.com

35. AutoPap® Primary Screening System. Product Insert. Burlington, NC: TriPath; 2001.

36. Zahniser D, Sullivan P. CYTYC Corporation. *Acta Cytol* 1996;**40**:37–44.

37. Knesel EA Jr. Roche Image Analysis Systems Inc. *Acta Cytol* 1996;**40**:60–6.

38. Payne N, Chilcott J, McGoogan E. Liquid-based cytology in cervical screening: a rapid and systematic review. *Health Technol Assess* 2000;**4**(18).

39. Herbert A, Johnson J. Personal view: is it reality or illusion that liquid-based cytology is better than conventional cytology? *Cytopathology* 2001; **12**:383–9.

40. van Ballegooijen M, van den Akker-van Marle ME, Warmerdam PG, Meijer CJ, Walboomers JM, Habbema JD. Present evidence on the value of HPV testing for cervical cancer screening: a model-based exploration of the (cost-) effectiveness. *Br J Cancer* 1997;**76**:651–7.

41. Cuzick J, Sasieni P, Davies P, Adams J, Normand C, Frater A, *et al*. A systematic review of the role of human papillomavirus testing within a cervical screening programme. *Health Technol Assess* 1999; **3**(14).

42. http:// www.cancerscreening.nhs.uk/cervical/ hpv.html. Accessed 10 January 2005.

43. Williams GH, Romanowski P, Morris L, Madine M, Mills AD, Stoeber K, *et al*. Improved cervical smear assessment using antibodies against proteins that regulate DNA replication. *Proc Natn Acad Sci USA* 1998;**95**:14932–7.

44. Ferris DG, Schiffman M, Litaker MS. Cervicography for triage of women with mildly abnormal cervical cytology results. *Am J Obstet Gynecol* 2001;**185**:939–43.

45. Cronje HS, van Rensburg E, Cooreman BF, Niemand I, Beyer E. Speculoscopy vs the acetic acid test for cervical neoplasia. *Int J Gynaecol Obstet* 2000:**69**:249–53.

46. Quek SC, Mould T, Canfell K, Singer A, Skladnev V, Coppleson M. The Polarprobe – emerging technology for cervical cancer screening. *Ann Acad Med Singapore* 1998;**27**:717–21.

47. National Institute of Clinical Excellence. *Guidance on the use of liquid based cytology for cervical screening*. Technology appraisal guidance No. 5. London: NICE; June 2000.

48. Khan KS, ter Reit G, Glanville J, Sowden AJ, Kleijnen J, editors. *Undertaking systematic reviews of research on effectiveness*. CRD Report 4. (2nd ed). York: NHS Centre for Reviews and Dissemination, University of York; March 2001.

49. Clarke M, Oxman AD, editors. Cochrane Reviewers' Handbook 4.0 [updated July 1999]. In *The Cochrane Library* [database on CD-ROM] (Issue 3). The Cochrane Collaboration. Oxford: Update Software; 2000.

50. Irwig L, Tosteson ANA, Gastonis C, Lau J, Colditz G, Chalmers TC, *et al*. Guidelines for meta-analyses evaluation diagnostic tests. *Ann Intern Med* 1994;**120**:667–76.

51. Briggs AH, Gray AM. Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technol Assess* 1999;**3**(2).

52. Noorani H, Arratoon C, Hall A. *Assessment of techniques for cervical cancer screening*. Ottawa, Ontario: Canadian Co-ordinating Office for Health Technology Assessment; May 1997.

53. Australian Health Technology Advisory Committee. *Review of automated and semi-automated cervical screening devices*. Canberra: Technology Section, Commonwealth Department of Health and Family Services; April 1998. pp. 1–86.

54. Egger M, Smith GD. Principles of and procedures for systematic reviews. In Egger M, Smith GD, Altman D, editors. *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ Publishing Group; 2001. Chapter 2.

55. Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. In Egger M, Smith GD, Altman D, editors. *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ Publishing Group; 2001. Chapter 7.

56. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup D. Improving the quality of reporting of meta-analysis of randomized controlled trials: the QUOROM statement. *Lancet* 1999;**354**:1896–900.

57. Abulafia O, Sherer DM. Automated cervical cytology: meta-analyses of the performance of the PAPNET system. *Obstet Gynaecol Surv* 1999; **54**:253–64.

58. Abulafia O, Sherer DM. Automated cervical cytology: meta-analyses of the performance of the AutoPap 300 QC System. *Obstet Gynaecol Surv* 1999;**54**:469–76.

59. Mango LJ, Radensky PW. Interactive neural network assisted screening: a clinical assessment. *Acta Cytol* 1998:**42**:233–45.

60. Radensky PW, Mango LJ. Interactive neural network-assisted screening: an economic assessment. *Acta Cytol* 1998;**42**:246–52.

61. Smith BL, Lee M, Leader S, Wertlake P. Economic impact of automated primary screening for cervical cancer. *J Reprod Med* 1999;**44**:518–28.

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

63. Brown AD, Garber AM. Cost-effectiveness of 3 methods to enhance the sensitivity of Papanicolaou testing. *JAMA* 1999;**281**:347–53.

64. Myers ER, McCrory DC, Subramanian S, McCall N, Nanda K, Datta S, *et al*. Setting the target for a better cervical screening test: characteristics of a cost-effective test for cervical neoplasia screening. *Obstet Gynecol* 2000;**96**:645–52.

65. Nanda K, McCrory DC, Myers ER, Bastian LA, Hasselblad V, Hickey JD, *et al*. Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. *Ann Intern Med* 2000;**132**:810–19.

66. Austin RM, Ramzy I. Increased detection of epithelial cell abnormalities by liquid-based gynecologic cytology preparations. A review of accumulated data. *Acta Cytol* 1998;**42**:178–84.

67. Cuzick J, Sasieni P, Davies P, Adams J, Normand C, Frater A, *et al*. A systematic review of the role of human papillomavirus testing within a cervical screening programme: summary and conclusions. *Br J Cancer* 2000;**83**:561–5.

68. Egger M, Zellweger Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet* 1997;**350**:326–9.

69. Szczepura A, Kankaanpaa T, editors. *Assessment of health care technologies. Key concepts and strategic issues*. Chichester: John Wiley; 1996.

70. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. *BMJ* 1996;**313**:275–83.

71. Mango LJ. Neural network-assisted cervical cancer screening. *Journal of Clinical Ligand Assay* 1998; **21**:203–7.

72. O'Leary TJ, Tellado M, Buckner SB, Ali IS, Stevens A, Ollayos CW. PAPNET-assisted rescreening of cervical smears: cost and accuracy compared with a 100% manual rescreening strategy. *JAMA* 1998;**279**:235–7.

73. Schechter CB. Cost-effectiveness of rescreening conventionally prepared cervical smears by PAPNET testing. *Acta Cytol* 1996;**40**:1272–82.

74. Brotzman GL, Kretzchmar S, Ferguson D, Gottlieb M, Stowe C. Costs and outcomes of PAPNET secondary screening technology for cervical cytologic evaluation. A community hospital's experience. *Archives of Family Medicine* 1999;**8**:52–5.

75. Raab SS, Zaleski MS, Silverman JF. The cost-effectiveness of the cytology laboratory and new cytology technologies in cervical cancer prevention. *Am J Clin Pathol* 1999;**111**:259–66.

76. Troni GM, Cipparrone I, Cariaggi MP, Ciatto S, Miccinesi G, Zappa M, *et al*. Detection of false-negative pap smears using the PAPNET system. *Tumori* 2000;**86**:455–7.

77. van Ballegooijen M, van den Akker-van Marle, Patnick J, Lynge E, Arbyn M, Anttila A, *et al*. Overview of important cervical screening process values in European Union (EU) countries, and tentative predictions of the corresponding effectiveness and cost-effectiveness. *Eur J Cancer* 2000;**36**:2177–88.

78. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996.

79. Eddy D. The frequency of cervical cancer screening: comparison of a mathematical model with empirical data. *Cancer* 1987;**60**:1117–22.

80. Sherlaw-Johnson C, Gallivan S, Jenkins D, Jones MH. Cytological screening and management of abnormalities in prevention of cervical cancer: an overview with stochastic modelling. *J Clin Pathol* 1994;**47**:430–5.

81. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine. How to practice and teach EBM*. 2nd ed. Edinburgh: Churchill Livingstone; 2000.

82. Barratt A, Irwig L, Glasziou P, Cumming RG, Raffle A, Hicks N, *et al*., for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. XVII. How to use guidelines and recommendations about screening. *JAMA* 1999;**281**:2029–34.

83. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;**274**:645–51.

84. Intersociety Working Group for Cytology Technologies. Proposed guidelines for primary screening instruments for gynecologic cytology. *Acta Cytol* 1997;**41**:924–9.

85. Intersociety Working Group for Cytology Technologies. Proposed guidelines for secondary screening (rescreening) instruments for gynecologic cytology. *Acta Cytol* 1998;**42**:273–4.

86. Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley; 1981.

87. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA* 1994;**271**:389–91.

88. Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA* 1994;**271**:703–7.

89. How good is that test (II)? *Bandolier* 1996;(27):2. URL:http://www.jr2.ox.ac.uk/bandolier/band27/b27-2.html.

90. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–30.

91. Lachs MS Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med* 1992;**117**:135–40.

92. Nanda K, Myers ER. Selecting a cervical cytology screening test – what are the issues? *Journal of Clinical Outcomes Management* 2001;**8**:29–32.

93. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol* 1992;**45**:581–6.

94. Melnikow J, Nuovo J, Willan AR, Chan BK, Howell LP. Natural history of cervical squamous intraepithelial lesions: a meta-analysis. *Obstet Gynecol* 1998;**92**:727–35.

95. McIndoe WA, McLean MR, Jones RW, Mullins PR. The invasive potential of carcinoma *in situ* of the cervix. *Obstet Gynecol* 1984;**64**:451–8.

96. Chock C, Irwig L, Berry G, Glasziou P. Comparing dichotomous screening tests when individuals negative on both tests are not verified. *J Clin Epidemiol* 1997;**50**:1211–17.

97. Miller WC. Bias in discrepant analysis: when two wrongs don't make a right. *J Clin Epidemiol* 1998; **51**:219–31.

98. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H. Publication bias and clinical trials. *Control Clin Trials* 1987;**8**:343–53.

99. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;**337**:867–72.

100. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technol Assess* 2000;**4**(10).

101. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990; **263**:1385–9.

102. Sohn D. Publications bias and the evaluation of psychotherapy efficacy in reviews of the research literature. *Clin Psychol Rev* 1996;**16**:147–56.

103. Hetherington J, Dickersin K, Chalmers I, Meinert CL. Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. *Pediatrics* 1989:**84**:374–80.

104. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods of diagnostic test accuracy. *J Clin Epidemiol* 1995;**48**:119–30.

105. Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with confidence*. 2nd ed. London: BMJ Books; 2000.

106. Doornewaard H, van der Schouw YT, van der Graaf Y, Bos AB, Habbema JDK, van den Tweel JG. The diagnostic value of computer-assisted primary cervical smear screening: a longitudinal cohort study. *Mod Pathol* 1999;**12**:995–1000.

107. Sherman ME, Schiffman M, Herrero R, Kelly D, Bratti C, Mango LJ, *et al*. Performance of a semiautomated Papanicolaou smear screening system. *Cancer (Cancer Cytopathol)* 1998;**84**:273–80.

108. Hølund B, Ejersbo D, Hjortebjerg A. Evaluering af PAPNET – et semiautomatisk screeningssystems anvendelse i screening mod livmoderhalskraeft. [Evaluation of PAPNET – a semiautomated system used in the screening against cervical cancer] [in Danish]. *Ugeskr Laeger* 1998;**160**:5802–6.

109. Kok MR, Boon ME, Schreiner-Kok CT, Koss LG. Cytological recognition of invasive cancer of the uterine cervix: comparison of conventional light-microscopical screening and neural network-based screening. *Hum Pathol* 2000;**31**:23–8.

110. Duggan MA, Brasher P. Paired comparison of manual and automated Pap test screening using the PAPNET system. *Diagn Cytopathol* 1997; **17**:248–54.

111. Lerma E, Colomo L, Carreras A, Esteva E, Quilez M, Prat J. Resecreening of atypical cervicovaginal smears using PAPNET. *Cancer (Cancer Cytopathol)* 1998;**84**:361–5.

112. PRISMATIC Project Management Team. Assessment of automated primary screening on PAPNET of cervical smears in the PRISMATIC trial. *Lancet* 1999;**353**:1381–5.

113. Veneti S, Papaefthimiou M, Symiakaki H, Ioannidou-Mouzaka L. PAPNET for cervical cytology screening. Experience in Greece. *Acta Cytol* 1999;**43**:30–3.

114. Wilbur DC, Prey MU, Miller WM, Pawlick GF, Colgan TJ. The AutoPap system for primary screening in cervical cytology. *Acta Cytol* 1998; **42**:214–20.

115. Wilbur DC, Prey MU, Miller WM, Pawlick GF, Colgan TJ, Taylor DD. Detection of high grade squamous intraepithelial lesions and tumors using the AutoPap system. Results of a primary screening clinical trial. *Cancer (Cancer Cytopathol)* 1999;**87**:354–8.

116. Doornewaard H, van der Schouw YT, van der Graaf Y. Reproducibility in double scanning of cervical smears with the PAPNET system. *Acta Cytol* 2000:**44**:604–10.

117. Doornewaard H, van der Schouw YT, van der Graaf Y, Bos AB, van den Tweel JG. Observer variation in cytologic grading for cervical dysplasia of Papanicolaou smears with the PAPNET testing system. *Cancer (Cancer Cytopathol)* 1999;**87**:178–83.

118. Jenny J, Isenegger I, Boon ME, Husain OA. Consistency of a double PAPNET scan of cervical smears. *Acta Cytol* 1997:**41**:82–7.

119. Mitchell H, Medley G. Differences between false-negative and true-positive Papanicoloaou smears on Papnet-assisted review. *Diagn Cytopathol* 1998; **19**:138–40.

120. Mitchell H, Medley G. Detection of unsuspected abnormalities by PAPNET-assisted review. *Acta Cytol* 1998:**42**:260–4.

121. Mitchell H, Medley G. Detection of laboratory false negative smears by the PAPNET cytologic screening system. *Acta Cytol* 1998:**42**:265–70.

122. Wertlake P. Results of AutoPap system assisted and manual cytologic screening. A comparison. *J Reprod Med* 1999;**44**:11–17.

123. Fetterman B, Pawlick G, Koo H, Hartinger J, Gilbert C, Connell S. Determining the utility and effectiveness of the NeoPath AutoPap 300 QC system used routinely. *Acta Cytol* 1999:**43**:13–22.

124. Marshall CJ, Rowe L, Bentz JS. Improved quality-control detection of false-negative Pap smears using the Autopap 300 QC system. *Diagn Cytopathol* 1999;**20**:170–4.

125. Bibbo M, Hawthorne C. Performance of the AutoPap Primary Screening System at Jefferson University Hospital. *Acta Cytol* 1999:**43**:27–9.

126. Chhieng DC, Elgert PA, Xiong Y, Cangiarella JF, Cohen J-M. Use of computer-assisted rescreening as an ancillary tool to subclassify AGUS cervical smears. *Diagn Cytopathol* 2000;**23**: 165–70.

127. Huang TW, Lin TS, Lee JS. Sensitivity studies of AutoPap® System Location-Guided Screening of cervical-vaginal cytologic smears. *Acta Cytol* 1999: **43**:363–8.

128. Duggan MA. Papnet-assisted, primary screening of cervico-vaginal smears. *Eur J Gynaecol Oncol* 2000; **21**:35–42.

129. Lee JS, Kuan L, Oh S, Patten FW, Wilbur DC. A feasibility study of the AutoPap System Location-Guided Screening. *Acta Cytol* 1998:**42**:221–6.

130. Losell K, Dejmek A. Comparison of Papnet-assisted and manual screening of cervical smears. *Diagn Cytopathol* 1999;**21**:296–9.

131. Sturgis CD, Isoe C, McNeal NE, Yu GH, DeFrias DV. PAPNET computer-aided rescreening for detection of benign and malignant glandular elements in cervicovaginal smears: a review of 61 cases. *Diagn Cytopathol* 1998;**18**:307–11.

132. Kmietowicz Z. Registries will have to apply for right to collect patients' data without consent. *BMJ* 2001;**322**:1199.

133. Mayor S. UK researchers told to maintain confidentiality. *BMJ* 2000;**321**:917.

134. Brown P. Cancer registries fear imminent collapse. *BMJ* 2000;**321**:849.

135. Casey A, McGloin J, Kocjan G. Evaluation of rapid re-screening compared with PAPNET assisted cervical screening as a method of detecting false negative cervical smears. *BSCC Annual Scientific Meeting* 1995; A10.

136. Morse A, Douglas G, Coleman D. Evaluation of the PAPNET system for quality control in cervical screening. *BSCC Annual Scientific Meeting* 1995; A11.

137. Desai M, for the European PAPNET Study Group. A multicentric European study comparing PAPNET assisted screening of archival cervical smears with conventional manual screening. *BSCC Annual Scientific Meeting* 1998; A34.

138. Sampson H, Brooke D, Sutton J, Quirke P. Performance of the AutoPap Primary Screening System. *J Pathol* 2001;**195**(Suppl):39a.

139. Stone DH. Design a questionnaire. *BMJ* 1993; **307**:1264–6.

140. Smith R. What is publication? A continuum. *BMJ* 1999;**318**:142.

141. Grohs DH. Impact of automated technology on the cervical cytologic smear. A comparison of cost. *Acta Cytol* 1998;**42**:165–70.

142. *Automated Pap smear rescreening technologies*. Windows on medical technology; No. 4. Plymouth Meeting, PA: ECRI; March 1998.

143. Bishop JW. The cost of production in cervical cytology: comparison of conventional and automated primary screening system. *Am J Clin Pathol* 1997;**107**:445–50.

144. Sasieni PD, Cuzick J, Lynch-Farmery E, National Co-ordinating Network for Cervical Screening Working Group. Estimating the efficacy of screening by auditing smear histories of women with and without cervical cancer. *Br J Cancer* 1996;**73**:1001–5.

145. http://www.statistics.gov.uk. Accessed 7 February 2002.

146. Redburn JC, Murphy MFG. Hysterectomy prevalence and adjusted cervical and uterine cancer rates in England and Wales. *Br J Obstet Gynaecol* 2001;**108**:388–95.

147. Ho GYK, Bierman R, Beardsley L, Chang CJ, Burk RD. Natural history of cervicovaginal papillomavirus infection in young women. *N Engl J Med* 1998;**338**:423–8.

148. Furber SE, Weisberg E, Simpson JM. Progression and regression of low-grade epithelial abnormalities of the cervix. *Aust N Z J Obstet Gynaecol* 1997;**37**:107–12.

149. Hulman G, Pickles CJ, Gie CA, Dowling FM, Stocks PJ, Dixon R. Frequency of cervical intraepithelial neoplasia following large loop excision of the transformation zone. *J Clin Pathol* 1998;**51**:375–77.

150. Zaitoun AM, McKee G, Coppen MJ, Thomas SM, Wilson PO. Completeness of excision and follow-up cytology in patients treated with loop excision biopsy. *J Clin Pathol* 2000;**53**:191–6.

151. Office for National Statistics. *Cancer statistics; registrations. Registrations of cancer diagnosed in 1998, England*. Series MB1, No. 29. London: Office for National Statistics; 2000.

152. Government Statistical Service. *Cervical screening programme, England: 1999–2000*. London: Department of Health; November 2000.

153. Jones MH, Jenkins D, Singer A. Regular audit of colposcopic biopsies from women with a mildly dyskaryotic or borderline cervical smear results in fewer cases of CIN III. *Cytopathology* 1996;**7**:17–24.

154. Netten A, Curtis L. *Unit costs of health and social care*. Canterbury: Personal Social Services Research Unit, University of Kent; 2000.

155. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. *JAMA* 1994;**272**(17):1367–71.

156. Sterne JAC, Bartlett C, Jüni P, Egger M. Do we need comprehensive literature searches? A study of publication and language bias in meta-analyses of controlled trials. In *3rd Symposium on Systematic Reviews: beyond the basics*, 3–5 July 2000, Oxford.

157. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, *et al*. Should unpublished data be included in meta-analyses? Current convictions and controversies. *JAMA* 1993;**269**:2749–53.

158. Angell M. Negative studies. *N Engl J Med* 1989; **321**:464–6.

Appendix 1

Secondary literature on clinical effectiveness of automation and related technologies: MEDLINE search 1997–2000

1. vaginal smears/
2. cervix neoplasms/
3. cervical intraepithelial neoplasms/
4. cervix dysplasia/
5. or/1-4
6. mass screening/
7. exp mass screening/
8. exp diagnosis computer assisted/
9. exp image processing computer assisted/
10. automat$.tw
11. or/6-10
12. Review and Meta analysis filter
13. 11 and 5 and 12
14. papnet$.tw
15. (pap adj net$).tw
16. autopap$.tw
17. (auto adj pap$).tw
18. autocyte$.tw
19. (auto adj cyte$).tw
20. thinprep$.tw
21. (thin adj prep$).tw
22. or/14-21
23. 12 and 22
24. 13 or 23

Appendix 2

Secondary literature on clinical effectiveness of automation and related technologies: excluded studies

Reasons for exclusion are given in brackets.

MEDLINE

1. Mitchell MF, Cantor SB, Brookner C, Utzinger U, Schottenfield D, Richards-Kortum R. Screening for squamous intraepithelial lesions with fluorescence spectroscopy. *Obstet Gynecol* 1999;**94**:889–96. [Not a new technology.]
2. Gotay CC, Wilson ME. Social support and cancer screening in African American, Hispanic, and Native American women. *Cancer Practice* 1998;**6**:31–7. [Not a new technology.]
3. Fylan F. Screening for cervical cancer: a review of women's attitudes, knowledge, and behaviour. *Br J Gen Pract* 1998;**48**:1509–14. [Not a new technology.]
4. Kim YB, Ghosh K, Ainbinder S, Berek JS. Diagnostic and therapeutic advances in gynaecologic oncology: screening for gynaecologic cancer. *Cancer Treat Res* 1998; **95**:253–76. [Not a systematic review of clinical/cost effectiveness or a *de novo* model.]
5. Heatley MK. What is the value of proliferation markers in the normal and neoplastic cervix. *Histol Histopathol* 1998;**13**:249–54. [Not a new technology.]
6. Nuovo J, Melnikow J, Hutchinson B, Paliescheskey M. Is cervicography a useful diagnostic test? A systematic overview of the literature. *J Am Board Fam Pract* 1997; **10**:390–7. [Not a new technology.]
7. Cava M, Greenberg M, Fitch M, Spaner D, Taylor K. Towards an inclusive cervical cancer screening strategy: approaches for reaching socioeconomically disadvantaged women. *Canadian Oncology Nursing Journal* 1997; **7**:14–18. [Not a new technology.]
8. DeMay RM. Common problems in Papanicolaou smear interpretation. *Arch Pathol Lab Med* 1997;**121**:229–38. [Not a new technology.]
9. Cenci M, Giovagnoli MR, Olla SV, Drusco A, Vecchione A. Automation of cytological analysis of cervical smears (L'automazione applicata all'analisi citologica dei preparati eso-endocervicali). *Minerva Ginecol* 1999;**51**:291–8. [Not a systematic review of clinical/cost effectiveness or a *de novo* model.]
10. Arbyn M, Schenck U. Detection of false negative pap smears by rapid reviewing. A meta-analysis. *Acta Cytol* 2000;**44**:949–57. [Not a new technology.]
11. Jepson R, Clegg A, Forbes C, Lewis R, Sowden A, Kleijnen J. The determinants of screening uptake and interventions for increasing uptake: a systematic review. *Health Technol Assess* 2000;**4**:1–133. [Not a new technology.]
12. Yabroff KR, Kerner JF, Mandelblatt JS. Effectiveness of interventions to improve follow-up after abnormal cervical cancer screening. *Prev Med* 2000;**31**:429–39. [Not a new technology.]
13. Rock CL, Michael CW, Reynolds RK, Ruffin MT. Prevention of cervix cancer. *Crit Rev Oncol Hematol* 2000;**33**:169–85. [Not a new technology.]
14. Linder J, Zahniser D. The ThinPrep Pap test. A review of clinical studies. *Acta Cytol* 1997;**41**:30–8. [Not a systematic review of clinical/cost effectiveness or a *de novo* model.]

Cochrane Library

Not given are controlled trials, or reviews already excluded in the MEDLINE search.

1. O Leary TJ, Tellado M, Buckner SB, Ali IS, Stevens A, Ollayos CW. PAPNET-assisted rescreening of cervical smears: cost and accuracy compared with a 100% manual rescreening strategy. *JAMA* 1998;**279**:235–7. [Not a systematic review of clinical/cost effectiveness or a *de novo* model.]

2. Cuzick J. Screening for cancer: future potential. *Eur J Cancer* 1999;**35**: 685–92. [Not a systematic review of clinical/cost effectiveness or a *de novo* model.]
3. Suh-Bergman EJ, Goodman A. Surveillance for endometrial cancer in women receiving tamoxifen. *Ann Intern Med* 1999;**131**:127–35. [Not a new technology.]
4. Foreman J, Hader J. Cervical cancer screening. Summary Report No. 8. Health Services Utilisation and Research Commission (HSURC); 1999. [Not a new technology.]
5. Grant CM. Cervical screening interval: costing the options in one health authority. *J Public Health Med* 1999;**21**:140–4. [Not a new technology.]
6. Raab SS, Bishop NS, Zaleski MS. Cost effectiveness of rescreening cervicovaginal smears. *Am J Clin Pathol* 1999;**111**:601–9. [Not a new technology.]
7. Raab SS. The cost effectiveness of cervical-vaginal re-screening. *Am J Clin Pathol* 1997; **108**:525–36. [Not a new technology.]
8. Healthcare Insurance Board (CVZ), Amstelveen, The Netherlands. Evaluation of the PAPNET system in cervix screening – primary research. [Not a systematic review of clinical/cost effectiveness or a *de novo* model.]
9. Jouveshomme S, Baffert S, Charpentier E, Souag A. Computer assisted analysis system for cervico-vaginal smears (systematic review, primary research, expert panel) [in French]. Paris: Comite d'Evaluation et de Diffusion des Innovations Technologiques (CEDIT); 1998. [Not a systematic review of clinical/cost effectiveness or a *de novo* model.]

Appendix 3

Secondary literature on clinical effectiveness of automation and related technologies: appraising the included studies

A number of guidelines has been developed to appraise and determine the level of quality of a systematic review. In particular, the QUOROM (Quality of Reporting of Meta-analyses) statement is an exhaustive checklist produced at a conference of participating clinical epidemiologists, clinicians, statisticians and researchers who conduct meta-analysis (*Systematic Reviews in Health Care*, 2001). Unfortunately, the checklist is aimed at systematic reviews of RCTs, and not reviews of other types of study. One of the consistent features of reviews of diagnostic tests is that they tend not to contain RCTs.

A more general approach has been produced by the guidelines developed by Oxman and colleagues,[155] whose checklist may be used on reviews of aetiology, diagnosis, prognosis, therapy and prevention. The problems with universal checklists in general are that they can end up being extensive if all the possible combinations of review are to be covered. Furthermore, there will be items on the list that are wholly inappropriate for the review being appraised. However, if one attempts to use briefer and thus more workable shorter checklists, there is the risk of valuable elements of quality assessment being omitted.

Notwithstanding this, the principle of using a checklist is not contended here, but rather that they should be made more specific for the problem being considered. To be able to design a checklist that assesses the quality of a particular review requires determining the core principles that all good checklists are attempting to assess.

This led the authors to consider the appraisal of reviews in terms of satisfying a number of objectives. In this sense the person appraising the review had a degree of control over the tests that should be applied to satisfy these objectives. The other consideration was to produce a list of objectives that not only comprehensively covered the prerequisites of a systematic review, but was easy to implement.

The list of aims that a high-quality review were considered to address consist of objective, completeness, accuracy, inference and reproducibility. These may be conveniently rearranged to form the mnemonic CAIRO.

The explanations and order in which each of these objectives were applied are described below.

Objective

This aim uses the word objective to encapsulate two different concepts. In essence,

- Does the review have an objective (and how have the researchers ensured that it is met)?
- Is the review objective, as in has everything possible been done to minimise bias?

To satisfy the first part, two questions were considered: (a) Was there a clearly focused question that the review was trying to answer? (b) Were the reported search strategy and inclusion criteria likely to find the right studies to answer this question?

The answer to question (a) in general would be found in the abstract. Question (b) involved considering the search terms and their combination to identify studies. It also considered the inclusion criteria and whether they have been appropriately designed to include high-quality articles. Here, the authors specifically looked at the search terms used to locate new technologies in cervical screening and considered the criteria for including a good study on cervical screening.

It should be noted that although this review is concerned with cervical screening there is a large degree of flexibility in this approach to accommodate the different types of review required for diagnostic, prognostic or therapeutic appraisals. The guidelines on whether the inclusion criteria are then capturing the relevant

studies to answer the objective may be found in articles on evidence-based medicine.

To satisfy the second part, ideally, would involve considering all types of bias at the different stages of the process. However, it was considered that adequate control of selection bias would ensure that the conclusions were not materially different from the truth.

The anticipated sources of selection bias were in the search strategy, inclusion criteria and data abstraction process. As each of these should have been explicitly designed before it was implemented, this was checked. Evidence-based guidelines on the prerequisites of a good study in other areas, such as diagnostics, will highlight whether selection bias is present in the criteria.

Completeness

The idea behind the objective of completeness is to achieve maximum coverage of the field being reviewed, so that the evidence pertinent to the conclusion has not been inadvertently missed. To do this the search, the sources of information and the inclusion criteria were considered in turn. A possible combination of questions to address this objective was:

- Was the combination of search terms used sufficient to cover all the relevant studies?
- Were sufficient databases interrogated to offer maximum coverage of the subject area? This would include the way in which the review dealt with unpublished studies.
- Were the inclusion criteria not too restrictive to exclude relevant articles?

Accuracy

A review should accurately report on the studies included in the appraisal. This may be assessed on two levels. The easiest but less valid way is to be satisfied with identifying whether any internal controls were used to ensure accurate transfer of information from the primary studies. The second involves doing this and testing the information.

Possible internal controls carried out by the investigators are:

- independent duplication of data abstraction
- independently checking that the information on the abstraction sheets corresponds to that in the source documents.

External validation involves the appraiser taking the data in the review from the included studies and checking them against the included primary studies. A sample of the primary studies was verified for accuracy of reporting in those reviews that were instrumental to this project.

Inference

Based on the information presented in the review, were the conclusions inferred by the reviewer reasonable and accurate? The only proviso with this was that all the conclusions had to be based on the included studies and no others, and followed coherently from the information reported.

Reproducibility

Was the reporting of information sufficiently complete that the review could be reproduced independently and the same conclusions derived? The robustness of any piece of scientific work is that the conclusions may be reproduced given the same circumstances. This adds to the strength of the results.

This required full disclosure of the search algorithms used, the databases searched and the period covered by the search. Inclusion criteria were preferably developed as an algorithm so that selection bias was minimised and inter- intra-rater agreement was maximised.

Appendix 4

Secondary literature on clinical effectiveness of automation and related technologies: detailed evidence tables

TABLE 84 *Automation: detail of clinical effectiveness reviews (Set A)*

Study/device	Objective	Search	Inclusion criteria
Abulafia and Sherer, 1999[57] AutoPap	To estimate the overall FNR of the AutoPap 300 QC system when applied as a (1) primary screening system, (2) quality control (rescreening) system	From ? to October 1998 Databases searched: • MEDLINE only • Cross-referencing Keywords not given No algorithm given *n* = 14	Not explicitly stated, but it appears that studies were included if: • English articles only • articles on the AutoPap 300 QC • provided complete data for analysis Objective 1 *n* = 4 Objective 2 *n* = 5
Abulafia and Sherer, 1999[58] PAPNET	Compared with conventional screening, to determine whether the PAPNET as a primary screener (1) identifies a larger number of abnormalities, (2) has a lower FNR; and whether the PAPNET as a rescreener reduces the FNR	From ? to August 1998. Databases searched: • MEDLINE only Keywords not given No algorithm given *n* = 21	Studies on the PAPNET English language only Objective 1 *n* = 5 Objective 2 *n* = 3 Objective 3 *n* = 13
AHTAC, 1998[53] PAPNET, AutoPap, AutoCyte SCREEN	(1) To what extent do the new technologies have the potential to reduce the incidence of, and morbidity and mortality from, cervical cancer? (2) What is the potential of the new technologies to increase the sensitivity and specificity of Pap smear screening?	1990 to July 1997 Databases searched: • MEDLINE • CancerLit • Manufacturers • Conference papers • FDA trials • Internet websites Keywords used in search given No algorithm given	No explicit inclusion criteria Criteria for quality of studies listed, but not for inclusion. Indications on the decision process behind including a study are given by the following two statements: "an overall judgement was made of which studies might contribute evidence about the potential of the technology to improve Pap smear screening in the Australian context" "Where there was sufficient evidence from well-designed studies, a critical appraisal of the associated technology was conducted" PAPNET *n* = 23, AutoPap *n* = 4, AutoCyte SCREEN *n* = 1

continued

TABLE 84 *Automation: detail of clinical effectiveness reviews (Set A) (cont'd)*

Study/device	Objective	Search	Inclusion criteria
Broadstock (NZHTA), 2000[62] AutoPap	(1) To assess the clinical effectiveness of semi-automated and automated cervical screening systems (2) To determine the applicability of the evidence in the context of the national screening programme in New Zealand	Updated search on AHTAC (1998) 1997 to May 2000 Databases searched: • MEDLINE • EMBASE • Current Contents • HealthStar • Science Citation index • HTA database • CancerLit • Econ Lit • Cochrane Library • DARE • NHS EED In addition, library catalogues, and websites were interrogated. Handsearching and contact with manufacturers not done Algorithm given	Studies included if: • after January 1997 • automated and semi-automated cervical screening systems • used a reference standard of either histology (or negative colposcopy), or cytology by adjudicated panel review where discrepancies are resolved by consensus diagnosis made by an independent panel of cytology professionals • FDA approved commercially available devices, i.e. AutoPap only • English articles only • not cited by AHTAC $n = 2$
Brown and Garber, 1999[63] AutoPap, PAPNET	To estimate the proportional increase in TPR of (1) manual primary screening and new technology-assisted rescreening, compared with (2) manual primary screening and 10% random rescreening. This was done to provide input data to a model	January 1987 to December 1997 Databases searched: • MEDLINE only Handsearching of major journals over same period, and contacted manufacturers for unpublished articles Text words and keywords used in search given	Studies included if: • reported the number and cytological results of all slides • reported on the FDA-approved use of one of the technologies for cervical screening • used biopsy or review of discrepant results by a panel of at least three cytopathologists to validate all positive cytology • included slides with a validated diagnosis of LSIL or more severe AutoPap $n > 2$, but for analysis only one study used; PAPNET $n = 7$, but for analysis $n = 4$
Mango and Radensky, 1998[59] PAPNET	To provide effectiveness metrics for the clinical utility of PAPNET	From ? to August 1997 Used their own database (Mango was the vice president and Medical Director of NSI). It consists of all published manuscripts, abstracts, text chapters and trade journal articles. It is periodically updated by MEDLINE searches and bibliography reviews No search algorithm or keywords given	Inclusion criteria not made explicit Studies on PAPNET $n = 22$

continued

TABLE 84 *Automation: detail of clinical effectiveness reviews (Set A) (cont'd)*

Study/device	Objective	Search	Inclusion criteria
McCrory *et al.*, 1999[12] AutoPap, PAPNET	To estimate the accuracy of the Pap test and the new technologies	Searched inception to March 1998. Databases searched: • MEDLINE (1966) • EMBASE (1980) • HealthSTAR (1975) • CINAHL (1983) • CancerLit (1983) • EconLit (1969) Algorithms given	Done as a step-by-step algorithm. Articles answering 'no' to any of the questions were excluded at that step. Step 1: 1. Was cervical cytology evaluated as a screening test? 2. Was the screening test for primary screening only, rescreening only, or primary screening with rescreening? 3. Was the reference standard histology, histology or negative colposcopy, or cytology? Step 2: 1. Did the study use a reference standard? 2. Was the reference standard histology or colposcopy? 3. For studies comparing histology or colposcopy with cytology as a screening test, were these tests reasonably concurrent, i.e. up to 3 months apart? 4. Can all cells of a 2-by-2 table be completed? Because there was only one article included at this stage, a step 3 in the algorithm was developed to reassess those articles that had been excluded owing to an inadequate reference standard and being unable to calculate sensitivity/specificity. Step 3: 1. Did the study use a two-armed design? 2. Were discordant results from the two study arms adjudicated by an independent panel of cytologists? 3. Were the majority of those testing positive for HSIL verified with histology or colposcopy? 4. Did the study design allow for separate analyses of sensitivity (or relative TPR) and specificity (or relative FPR)? Even after step 3, for AutoPap and PAPNET $n = 0$, but step 3.4 was relaxed, resulting in six studies of AutoPap and 11 of PAPNET

continued

TABLE 84 *Automation: detail of clinical effectiveness reviews (Set A) (cont'd)*

Study/device	Objective	Search	Inclusion criteria
Nanda et al., 2000[65] Any, including AutoPap, PAPNET	To evaluate the accuracy of conventional and new methods of Papanicolaou testing when used to detect cervical cancer and its precursors	Searched from (dates given below) to October 1999 Databases searched: • MEDLINE (from 1966) • EMBASE (from 1980) • HealthSTAR (from 1975) • CINAHL (from 1983) • CancerLit (from 1983) Handsearching of newly published journals, bibliographies of included studies and systematic reviews Unpublished studies located by contacting professional societies and manufacturers Algorithms given	Studies were included if they were in the English language, FDA-approved technologies (AutoPap, PAPNET) and satisfied step 3 of McCrory *et al.* (1999), namely: 1. Did the study use a two-armed design? 2. Were discordant results from the two study arms adjudicated by an independent panel of cytologists? 3. Were the majority (>50%) of those testing positive for HSIL verified with histology or colposcopy? 4. Did the study design allow for separate analyses of sensitivity (or relative TPR) and specificity (or relative FPR)? AutoPap $n = 0$, PAPNET $n = 0$
Noorani *et al.* (CCOHTA), 1997[52] AutoPap, PAPNET	(1) To examine the effectiveness of the Pap test (2) To consider different strategies for improving the effectiveness of the Pap test	? 1985 to ?1997 Databases searched: • MEDLINE • CancerLit • EMBASE • Health Planning and Administration • Pascal • CCOHTA Library database • FDA Online database • ECRINet • Current Contents • DIALOG PTS Newsletter Questionnaires circulated to manufacturers Keywords given No algorithm given	English articles only Other inclusion criteria not given

TABLE 85 *Automation: details of clinical effectiveness reviews (Set B)*

Study/device	Results and outcomes	Concluding remarks	Comments
Abulafia and Sherer, 1999[57] AutoPap	Definition of abnormal not given Threshold for positive not given Outcome: sensitivity Meta-analysis Primary screening mode with 50% review rate: sensitivity ranges between 85 and 100% Rescreening mode: average sensitivity = 37% (95% CI 34 to 40%) Note: a range of thresholds for positive was used in both analyses	A core group of authors was responsible for the majority of the publications (13/14). With the independence of these studies being called into question, the authors indicated that "any meta-analysis of this collection of studies should be interpreted with caution". There is a relative paucity of data on the AutoPap 300 QC	The search keywords or algorithm were not made explicit, making reproducibility difficult Confining the search to one database runs a substantial risk for excluding including valuable studies The inclusion criteria were not sufficiently explicit It appears the primary studies were not screened for an acceptable reference standard, giving the potential for inflated estimates of performance Only four studies provided complete data on the AutoPap as a primary screener, with only five studies as a quality control device. The remaining studies were incomplete in their data. The authors made no attempt to contact the research groups in question Specificity was not considered
Abulafia and Sherer, 1999[58] PAPNET	Definition of abnormal not given Threshold for positive not given Outcome: odds ratios (OR) Meta-analysis using Mantel–Haenszel method (1) Odds of detecting abnormalities with PAPNET primary screener vs conventional screening OR = 1.19 (95% CI 1.13 to 1.26) (2) Odds of false negatives with PAPNET primary screener vs conventional screening OR = 0.41 (95% CI 0.25 to 0.67) (3) PAPNET as a rescreening device. Wide range of values made meta-analysis difficult. PAPNET reclassified 0.1–5% of negatives from manual arm, and 20–90% of known false-negatives Note: a range of thresholds for positive was used in both analyses	Two studies were rejected for being discordant with the others "We conclude that compared with manual screening, PAPNET identifies 20% more abnormal, has two-fold less false negative, and re-classifies as abnormal one third of manually screened false negative slides"	The search keywords or algorithm were not made explicit, making reproducibility difficult Confining the search to one database runs a substantial risk of excluding valuable studies The inclusion criteria were not sufficiently explicit. Restricting studies to English language is dangerous in this instance, as a number of foreign researchers have assessed the PAPNET It appears that the primary studies were not screened for an acceptable reference standard, giving the potential for inflated estimates of performance Explanations for the two discordant studies' heterogeneity were not sought

continued

TABLE 85 *Automation: details of clinical effectiveness reviews (Set B) (cont'd)*

Study/device	Results and outcomes	Concluding remarks	Comments
AHTAC, 1998[53] PAPNET, AutoPap, AutoCyte SCREEN	Definition of abnormal and threshold for positive smear: considered and highlighted as deficiency in most studies Outcomes: sensitivity, specificity (with confidence intervals), additional cases detected PAPNET (rescreener): Increased detection of LSIL by 0–7% Increased detection of HSIL by 3–6% (threshold for positive smear not given for either of these estimates) AutoPap: Probable improvement over 10%: random rescreening, but too little evidence to quantify AutoCyte SCREEN: Too little evidence to evaluate performance	The following deficiencies in evidence were noted by the authors: • a limited number of studies • extensive manufacturer involvement in the studies, absence of RCTs • lack of cytological threshold for positive and negative results • no consistent definition of a positive smear • few studies with biopsy confirmation of results • no definition of gold standard for negative results • sensitivity and specificity generally not reported • tests of statistical significance often not undertaken or reported • lack of consistent comparator • reviewers not always blinded to outcome • study populations displayed spectrum bias • given all of these, the new technologies could not be recommended at the time of the review	Essentially only two major databases were searched, MEDLINE and CancerLit. This is not wide enough to ensure completeness of coverage The inclusion of studies for appraisal was too subjective to be free from selection bias Although problems with positive smear and disease definitions were highlighted, the reviewers did not indicate how these matters should be addressed Comprehensive coverage was given to the problems faced by reviewers in appraising studies in the field of cervical cytology
Broadstock, 2000[62] AutoPap	Definition of abnormal = HSIL+ Threshold for positive = HSIL+ Outcomes: sensitivity, specificity Probable small increase in sensitivity over the Pap test, for low-grade lesions only There is inadequate evidence concerning the specificity of the AutoPap	Sensitivity and specificity could not be reliably determined. No difference was found in the detection of high-grade lesions. Increase in sensitivity probably confined to low-grade lesions More research required "The majority of studies appraised were at least partially funded by the industry producing the devices considered" The vast majority of missed lesions in the existing programme would be detected in subsequent screening rounds Increases in sensitivity may come at the cost of decreased specificity Automation cannot be recommended for the New Zealand national cervical screening programme	A limited number of studies ($n = 2$) There was no consideration of the importance of a two-armed design in the inclusion criteria. This was arguably more important than an acceptable reference standard If strict inclusion criteria are to be applied, then a sensitivity analysis on the inclusion criteria should also be carried out The search was confined to English language articles only

continued

TABLE 85 *Automation: details of clinical effectiveness reviews (Set B) (cont'd)*

Study/device	Results and outcomes	Concluding remarks	Comments
Brown and Garber, 1999[63] AutoPap, PAPNET	Definition of abnormal = LSIL+ Threshold for positive not given Outcome: sensitivity (TPR) For overall sensitivity, a conventional screening sensitivity of 80% is assumed AutoPap 300 QC at 20% review: Sensitivity of rescreening = 77% (n = 1) Overall sensitivity of screening + rescreening = 95.4% (n = 1) PAPNET rescreening: Sensitivity ranged from 19.6 to 100% (n = 7) Mean sensitivity of rescreening = 85.9% (n =4) Overall sensitivity of screening + rescreening > 97% (n = 4)	"Our estimates are subject to uncertainty because the literature on the effectiveness of the 3 technologies (includes ThinPrep) reviewed here is incomplete and sometimes contradictory" "The highest quality studies suggest that the technologies increase the TPR by a modest amount, especially in a laboratory that is already highly accurate"	A limited number of studies (n = 1) for the AutoPap Threshold of test not given Evaluations of specificity not given Only one database searched, giving the potential of excluding articles of value
Mango and Radensky, 1998[59] PAPNET	Definition of abnormal: Ranged from CIN1 to CIN3 Threshold for positive: ranged from ASCUS+ to HSIL+. Outcomes: a number of sensitivity-based metrics With an assumed conventional sensitivity of 85%, the sensitivity of the PAPNET exceeded 89% Relative yield of the rescreening effectiveness of the PAPNET ranges from 1.08 to 1.49	There is a relatively extensive evidence base for the PAPNET, which suggests that its sensitivity for abnormalities exceeds that of unassisted screening Sensitivity used as the main outcome because the patients' safety depends on this metric more than any other	The periodic update of the database by MEDLINE is insufficient: MEDLINE is not a comprehensive database Use of their own database reduces transparency and makes it difficult to reproduce the work Lack of independence of the authors will always produce suggestions of favourable bias Unlike in a number of other reviews, reference standards and definition of positive were addressed Specificity not considered as an outcome measure. Oddly, the justification was that such estimates affect the economics of the screening programme, not the well-being of the patient. This appears to ignore the impact that a false-positive diagnosis has on a patient's life

continued

TABLE 85 *Automation: details of clinical effectiveness reviews (Set B) (cont'd)*

Study/device	Results and outcomes	Concluding remarks	Comments
McCrory *et al.*, 1999[12] AutoPap, PAPNET	Definition of abnormal: Ranged from CIN1 to CIN 3. Threshold for positive: ranged from ASCUS to HSIL Outcomes: sensitivity, specificity PAPNET: only one study gave values for both sensitivity and specificity. As a rescreening device: Se (LSIL, CIN1) = 38% Sp (LSIL, CIN 1) = 92% Se (LSIL, CIN2/3) = 41% Sp (LSIL, CIN2/3) = 83% but this was a narrow spectrum of patients, 160 ASCUS/AGUS Other studies allowed only for calculation of sensitivity. As a primary screening system: Se (PAPNET) 86% vs Se (Manual) 77% AutoPap: as above, little information to estimate specificity; five studies on the AutoPap 300 QC system	"The values reported for sensitivity and specificity in the few studies that use histological or colposcopic reference standards are well within the range of sensitivity reported for the conventional Pap test" "However, including studies that directly compare these new technologies with conventional Pap smear testing (screening or re-screening) using a cytological reference standard results in significant improvements in sensitivity" The authors concluded that the evidence on the new technologies was insufficient for two reasons: • There is little evidence on the specificity • Most of the estimates of sensitivity are based on a surrogate reference standard: cytology. Independent consensus agreement by a panel is often confined to the discordant samples only, with biopsy confirmation of high grade lesions often lacking Both lead to an overestimation of diagnostic performance	In terms of thoroughness of review techniques, this probably represents the best attempt at systematically reviewing the field of cervical cytology to date The use of sensitivity analysis to demonstrate the key criterion that excluded studies is powerful and departs from the attempts of forerunners This is also one of the few reviews that clearly separates the threshold for test positive from the threshold (or definition) of abnormal as measured by the reference standard The conventional Pap test was also appraised by meta-analysis using summary receiver operator characteristic curves; this has not been covered here
Nanda *et al.*, 2000[65] Any, including AutoPap, PAPNET	Definition of abnormal: ranged from CIN1 to CIN 3. Threshold for positive: ranged from ASCUS to HSIL Outcomes: sensitivity, specificity PAPNET: No included studies AutoPap: No included studies	There were three main deficiencies in the methodologies of the excluded studies: • Many of the studies reviewed did not apply the new technology and the conventional Pap test prospectively to the same sample • Where cytology was used as the reference standard, only discordant results were verified, concordant results remaining unverified • There was little evidence on the value of the specificity for any of the new technologies Other deficiencies applied to all studies and included: • the presence of spectrum bias in the study population • the problems associated with no recognised reference standard • the presence of verification bias in study methodology	This study also assessed the efficacy of the conventional Pap test and the ThinPrep. There were 94 studies included on the former and three on the latter (see below) It is essentially an update on the report provided for the AHCPR. But, unlike in the earlier report, the investigators chose not to relax criterion 3.4, hence the exclusion of all articles on the AutoPap and the PAPNET It would have been interesting to see after relaxing 3.4 whether more articles would have been included than in McCrory *et al*. Criterion 3.1 allowed for the comparison of the test with the reference standard alone. Unfortunately, this does not provide for the hypervigilance (and thus enhanced performance) experienced under test conditions

continued

TABLE 85 *Automation: details of clinical effectiveness reviews (Set B) (cont'd)*

Study/device	Results and outcomes	Concluding remarks	Comments
Noorani *et al.*, 1997[52] AutoPap, PAPNET	Definition of abnormal not given Threshold for positive not given Outcomes: not given Estimates of the test performance of the automated devices could not be derived	Operating characteristics of automated systems could not be estimated There is a "lack of a common definition of the gold standard for Pap smear re-screening" The new techniques may increase the effectiveness of the Pap test Resources should not be diverted from recruitment (of subjects), information systems, training and quality control requirements for laboratories to promote the new technologies	Probably owing to limited evidence at the time of the review, the sensitivity and specificity of the automated systems could not be ascertained It is not clear from the review what type of study qualified for appraisal The problems of defining test positives and disease before calculating test performance were not addressed Possible effectiveness outcomes were not considered

TABLE 86 *LBC: detail of clinical effectiveness reviews (Set A)*

Study/device	Objective	Search	Inclusion criteria
AHTAC, 1998[53] AutoCyte PREP, ThinPrep	See *Table 84*	See *Table 84*	See *Table 84* AutoCyte PREP $n = 4$, ThinPrep $n = 13$
Austin and Ramzy, 1998[66] ThinPrep, AutoCyte Prep	To compare the clinical effectiveness of LBC with conventional screening, by measuring the increased number of abnormalities detected	Search from ? to November 1997 Databases searched not given Indication of sources used given in abstract, "published studies, reported on at professional meetings or in the press and from the FDA pre-market trials" No search algorithm or keywords given	Studies were included if: • split-sample design (one sample taken from subject but divided between the new technology and conventional smear) • ThinPrep 2000 or ThinPrep beta system • AutoCyte Prep or CytoRich beta system ThinPrep 2000 $n = 5$, ThinPrep beta system $n = 14$, AutoCyte Prep $n = 10$
Broadstock, 2000[62] AutoCyte PREP, ThinPrep	(1) To assess the clinical effectiveness of LBC cervical screening systems (2) To determine the applicability of the evidence in the context of the national screening programme in New Zealand	See *Table 84*	See *Table 84*, but FDA-approved commercially available devices are ThinPrep and AutoCyte PREP in this instance ThinPrep $n = 10$, AutoCyte PREP $n = 3$

continued

TABLE 86 *LBC: detail of clinical effectiveness reviews (Set A) (cont'd)*

Study/device	Objective	Search	Inclusion criteria
Brown and Garber, 1999[63] ThinPrep 2000 system	To estimate the proportional increase in sensitivity (TPR) of (1) manual primary screening with ThinPrep and 10% random rescreening, compared with (2) manual primary screening with Pap smear and 10% random rescreening	See *Table 84* *n* = 200	See *Table 84* ThinPrep 2000 *n* = 0, earlier versions of ThinPrep system *n* = 3
McCrory *et al.*, 1999[12] ThinPrep	See *Table 84*	See *Table 84*	See *Table 84* ThinPrep *n* = 8
Nanda *et al.*, 2000[65] ThinPrep	See *Table 84*	See *Table 84*	See *Table 84* ThinPrep *n* = 3
Payne *et al.* (NICE), 2000[38] CYTOSCREEN, LABONORD Easy Prep, AutoCyte PREP, ThinPrep	What is the effectiveness of LBC for cervical screening compared with conventional screening?	Searched 1996 to November 1999 Databases searched • MEDLINE • EMBASE • Science Citation Index • Cochrane Library • HealthSTAR • NHS CRD: DARE, NEED and HTA • NRR Algorithms given	All HTA and related secondary research studies included Primary studies included that compared LBC cytology with conventional smears using any of the following outcome measures: • sensitivity and/or specificity • categorisation of specimens • % of inadequate or unsatisfactory specimens • specimen interpretation times Number of studies included not given (counted to be around 40)

TABLE 87 *LBC: detail of clinical effectiveness reviews (Set B)*

Study/device	Results and outcomes	Concluding remarks	Comments
AHTAC, 1998[53]	See *Table 85*	See *Table 85*	See *Table 85*
AutoCyte PREP, ThinPrep	ThinPrep: May increase the detection of biopsy proven abnormalities by 6–11% This includes a 5–6% increase in the detection of high-grade abnormalities Screening time was reported as being shorter "There is a reduction in the proportion of smears rated unsatisfactory for evaluation when ThinPrep is used" AutoCyte PREP: Fewer studies Probably has similar benefits to the ThinPrep, in terms of sensitivity, reduction in unsatisfactory smears and shorter screening time, but too few studies to quantify	In addition, • All study designs for the assessment of both LBC technologies have been prospective and used the split-sample technique. This may disadvantage LBC, as this is prepared after the conventional smear. • There appears to be a trade-off between the conventional Pap test and LBC, as each technology detects abnormalities that the other fails to detect • There is a significant learning period before becoming competent at monolayer screening	
Austin and Ramzy, 1998[66] ThinPrep, AutoCyte Prep	Definition of abnormal: not given Threshold for positive: LSIL+ Outcome: yield = (difference in no. LSIL+ detected)/(no. LSIL+ detected by Pap test) ThinPrep yield: ranged from –18% to +105%, mean yield = +14.1% (no. of cases 83,000) AutoCyte Prep yield: ranged from –10% to +117%, mean yield = +10.0% (no. of cases 16,000)	The data show an overall increase in the detection of LSIL for both LBC technologies, compared with conventional techniques Comparisons made to detect the impact of collecting device, however, demonstrated that LBC detects fewer abnormalities than conventional when the Ayre's wooden spatula is used. The residue is used for the LBC system This has led the authors to suggest that direct-to-vial studies would "ultimately be much more relevant than currently available split-sample data in judging the true potential of liquid-based methods to enhance detection"	The review has a number of deficiencies, including • no search strategy • no clear list or referencing of the data sources used • no explicit inclusion criteria • no assessment of quality • no clear reference standard used All of the above shortcomings led to problems with reproducibility, selection bias and overestimation of performance metrics
Broadstock, 2000[62] AutoCyte PREP, ThinPrep	See *Table 85* Sensitivity and specificity could not be reliably determined owing to inadequate verification	See *Table 85* Inadequate verification of the new tests against a suitable reference standard was a consistent failing; only discordant results were verified and concordant positives were often not subject to any verification The introduction of LBC cannot be recommended for the New Zealand national cervical screening programme	See *Table 85*

continued

TABLE 87 *LBC: detail of clinical effectiveness reviews (Set B) (cont'd)*

Study/device	Results and outcomes	Concluding remarks	Comments
Brown and Garber, 1999[63] ThinPrep 2000 system	See *Table 85* Increase in sensitivity ranged from 9.4 to 20.9% Mean proportional increase of sensitivity on initial screening = 14.9% Overall sensitivity of initial screening with rescreening = 92.6%	See *Table 85*	A limited number of studies (n = 3) of an earlier system Threshold of test not given Evaluations of specificity not given Only one database searched, giving the potential for excluding articles of value
McCrory *et al.*, 1999[12] ThinPrep	See *Table 85* ThinPrep The main results are the same as those detailed in Nanda *et al.*[65] (see below)	See *Table 85*	See *Table 85*
Nanda *et al.*, 2000[65] ThinPrep	See *Table 85* One study in which true sensitivity and specificity could be estimated Positive threshold = LSIL+ and abnormal = CIN2/3 ThinPrep vs conventional: Se 94.2% vs 84.6% Sp 57.7% vs 37.0% Note, however, that small samples were used here (54 patients in ThinPrep arm and 89 in conventional arm) For the other two studies relative TPR and FPR could be calculated: ThinPrep vs conventional: Relative TPR 1.13 to 1.19 Relative FPR 1.12 to 2.05 i.e. ThinPrep shows increased sensitivity but reduced specificity compared with conventional screening	See *Table 85*	See *Table 85* The demonstration that the increased sensitivity of the ThinPrep over the conventional Pap test could be at the expense of specificity confirms the fears of Payne *et al.*,[38] and reaffirms the necessity to consider the two metrics together and not separately

continued

TABLE 87 *LBC: detail of clinical effectiveness reviews (Set B) (cont'd)*

Study/device	Results and outcomes	Concluding remarks	Comments
Payne *et al.* (NICE), 2000[38] CYTOSCREEN, LABONORD Easy Prep, AutoCyte PREP, ThinPrep	Definition of abnormal not given Threshold for positive: ranges from LSIL+ to HSIL+ Outcome: sensitivity, specificity and interpretation times Sensitivity could not be estimated, but LBC technology probably detects more LSIL+. Improvements were small or not statistically significant Specificity could not be assessed Screening times were probably lower, but few studies (3 vs 4–6 minutes)	LBC would lead to: • a decrease in the % of inadequate specimens • an improvement in sensitivity • a probable decrease in interpretation times • a potential for easier use with other technologies, such as HPV and automation However, the following deficiencies were in evidence: • no RCTs • specificity is largely unknown and may be worse than the conventional Pap test • few studies had a gold-standard comparator	In estimates of sensitivity/specificity, the threshold for disease was not specified (i.e. CIN 1, 2, 3 or invasive) No attempt was made to calculate relative TPR and relative FPR, when a cytological reference standard was used The results of the search and inclusion could have been made more explicit to facilitate reproducibility

TABLE 88 *HPV: detail of clinical effectiveness reviews (Set A)*

Study/device	Objective	Search	Inclusion criteria
Cuzick *et al.* (NHSHTA), 1999[41,67] All possible devices reviewed, but only two were suitable for mass screening: MY09/11 Consensus PCR, GP+5/+6 Consensus PCR and Hybrid Capture II (Digene)	To evaluate the available data: (1) concerning the role of HPV testing in primary screening, either alone or as an adjunct to cytology (2) to improve the management of women with low-grade cytological abnormalities (3) to improve the accuracy of follow-up after treatment of preinvasive or early invasive lesions To review the methods available for HPV testing and determine their appropriateness for widespread implementation To determine what future research is required to obtain more reliable answers about its use in screening	Date of searches not given Eight databases apparently searched, but only the following are listed: • MEDLINE • EMBASE • Cochrane Database of Systematic Reviews Abstracts from 16th and 17th International Papillomavirus Conferences were searched Scanning for citations listed in reviews Ongoing and unpublished studies were considered based on personal knowledge consultation with experts in the field *n* = 2100	Studies were included if they: • were written in English • evaluated an HPV assay that could be used in screening • could be applied to smear or lavage material • estimated sensitivity/specificity and repeatability • could be implemented on a large scale and/or used in conjunction with automation Specifically, all articles had to: • provide a direct comparison of ≥ 2 of the technologies considered • use a sampling technique that is applicable to a cervical cancer screening programme • have a sample size ≥ 75

TABLE 89 *HPV: details of clinical effectivenees reviews (Set B)*

Study	Results and outcomes	Concluding remarks	Comments
Cuzick *et al*. (NHSHTA), 1999[41,67]	*Application of test to a solution of HPV and human DNA* Definition of abnormal: presence of HPV DNA Threshold for positive: HPV DNA No reference standard given Outcome: no. of HPV genomes required for test to detect presence of HPV Both PCR systems = 1–500 (n = 10) Hybrid Capture II = 5000 (1.0 pg ml^{-1}) (n = 0, based on the manufacturer's recommendations) *Comparisons of tests on clinical samples* Definition of abnormal: presence of HPV DNA Threshold for positive: HPV DNA No reference standard given Outcome: relative TPR and relative sensitivity (different definition to that used in the present report) HCII equivalent to PCR MY09/11 (n = 1) when cut-off for HPV <1.0 pg ml^{-1} in HCII system PCR MY09/11 equivalent to PCR GP5+/GP6+ (n = 1) *Comparisons of tests for the identification of cervical disease* Definition of abnormal: HSIL Threshold for positive: HPV DNA No reference standard given Outcome: sensitivity/specificity HCII: Se = 88.9%, Sp = 67.1% PCR MY09/11 – Digene Sharp assay: Se = 75.6%, Sp = 34.9% Cytology: Se = 62% Sp not given (n = 1) Note: these results are based on one study, and the lead author is the same lead author of the HPV review, and has declared funding from Digene. Further, it was acknowledged that the PCR performance in this study was probably not representative of the PCR generally	All three technologies had similar performance characteristics. Sensitivity and NPV were superior to other technologies used to detect HPV However, as the authors are keen to point out, from a practical point of view if a national programme were to adopt the technology, "it is important to note that the only technology that is currently available as a commercial 'off the shelf' kit is the Digene HCII assay" As a primary screening test, "HPV testing is more sensitive than cytology for detecting CIN II/III", however, "the specificity is substantially lower" With borderline and low-grade smears testing for high-risk types of HPV "greatly improves the specificity and positive predictive value"	The search for evidence resulted in a large number of articles being retrieved. However, the authors were not sufficiently explicit about the period covered by the search and all the databases searched for others to be able to reproduce the work Lack of independence of authors: one was a consultant for, and another received a grant from Digene, the manufacturer of the only commercially available HPV detection system The value of the true test performance is difficult to discern, as the study sample has had no recognised gold standard applied. This is particularly the case in an HPV-positive but otherwise cervical dysplasia-free population The total number specimens found positive by either or both tests was assumed to quantify the prevalence of HPV in the population. Unless otherwise stated, even the "status of discrepant samples was not established" Note: the natural history and prevalence of HPV were also reported in this review

Appendix 5

Search algorithm used on MEDLINE to identify articles on cost and cost-effectiveness of automated image analysis devices

Database: Medline (1993 to Present)
Search Strategy (You Saved Citations 1–50 From Set 47):

1	cervix neoplasms/	9007
2	cervical intraepithelial neoplasms/	1759
3	cervix dysplasia/	685
4	exp diagnosis computer assisted/	3696
5	exp image processing computer assisted/	26692
6	automat$.mp.	21743
7	"((papnet$ or (pap adj net))".tw.	0
8	papnet.tw.	87
9	(pap adj net).tw.	1
10	8 or 9	87
11	(autopap$ or (auto adj pap$)).tw.	26
12	(autocyte$ or (auto adj cyte$)).tw.	22
13	(thinprep$ or (thin adj prep$)).tw.	102
14	or/4-13	49190
15	((pap or papan$) and (smear$ or test$)).tw.	2282
16	(cervical adj cytology).tw.	453
17	(cervical adj screening).tw.	335
18	vaginal smears/	3350
19	mass screening/	14999
20	or/15-19	18464
21	dyskaryo$.mp.	155
22	1 or 2 or 3 or 21	9469
23	14 and 20 and 22	234
24	economics/	480
25	"exo costs and cost analysis"/	0
26	exp "costs and cost analysis"/	33627
27	cost of illness/	2759
28	exp health care costs/	10074
29	economic value of life/	331
30	exp economics medical/	986
31	exp economics hospital/	2866
32	economics pharmaceutical/	494
33	exp "fees and charges"/	3994
34	(cost or costs or costed or costly or costing).tw.	49923
35	(economic$ or pharmacoeconomic$ or price$ or pricing).tw.	21272
36	or/24-35	81146
37	23 and 36	33
38	limit 37 to yr=1996-2002	28
39	from 38 keep 1-28	28
40	exp decision support techniques/	16246
41	exp models statistical/	40910
42	monte carlo method/	2908

43	survival analysis/	21095
44	or/40-43	75442
45	36 or 44	51506
46	23 and 45	60
47	limit 46 to yr=1996-2002	50
48	23 and 44	34
49	limit 48 to yr=1996-2002	29
50	from 47 keep 1-50	

Appendix 6

Search strategies for systematic reviews of evidence on clinical effectiveness

Database: MEDLINE 1997 to December 2000

1. cervix neoplasms/
2. cervical intraepithelial neoplasms/
3. cervix dysplasia/
4. exp diagnosis computer assisted/
5. exp image processing computer assisted/
6. automat$.mp
7. (papnet$ or (pap adj net$)).tw
8. (autopap$ or (auto adj pap$)).tw
9. (autocyte$ or (auto adj cyte$)).tw
10. (autocyte adj screen$).tw
11. (thinprep$ or (thin adj prep$)).tw
12. or/4-11
13. ((pap or papan$) and (smear$ or test$)).tw
14. (cervical adj cytology).tw
15. (cervical adj screening).tw
16. vaginal smears/
17. mass screening/
18. or/13-17
19. dyskaryo$.mp
20. 1 or 2 or 3 or 19
21. 12 and 18 and 20
22. limit 21 to yr=1998 –2000

Database: EMBASE 1980 to present

1. exp uterine cervix cancer/
2. uterine cervix dysplasia/
3. uterine cervix tumor/
4. dyskaryo$.mp
5. or/1-4
6. computer assisted diagnosis/
7. image processing/
8. automat$.mp
9. (autocyte adj screen$).tw
10. (papnet$ or (pap adj net$)).tw
11. (autopap$ or (auto adj pap$)).tw
12. (autocyte$ or (auto adj cyte$)).tw
13. (thinprep$ or (thin adj prep$)).tw
14. or/6-13
15. papanicolaou test/
16. ((pap or papan$) and (smear$ or test$)).tw
17. uterine cervix cytology/
18. (cervical adj cytology).tw
19. (cervical adj screening).tw
20. vaginal smear/
21. screening/
22. screening test/
23. mass screening/
24. or/15-23
25. 5 and 14 and 24
26. limit 25 to yr=1998-2000

Appendix 7

Included and excluded studies in systematic review of test performance

TABLE 90 *Studies from update search reaching step 2 (n = 45)*

Authors	Description of study	Standard	Inclusion/exclusion[a]
Anon.	PRISMATIC trial: primary screener PAPNET vs conventional in SE England. Five centres, 21,700 smears. UK	Independent adjudication of discordant smears by panel of more than two experts. All abnormals and random sample of negatives independently reviewed	Excluded: no biopsy confirmation of HSIL+ [4]
Bibbo and Hawthorne	Primary screener AutoPap review of 5865 smears. AutoPap negatives reviewed by rapid then detailed review. USA	Panel of three experts adjudicated discordant results. No HSIL+ smears	Excluded: not a two-armed design [1]
Bibbo, Hawthorne and Zimmerman	AutoPap screening of 5865 cases to test correlation of ranking with severity of abnormality. USA	Cytology diagnosis for normal and abnormal slides, with cytopathologist, reviewing the latter	Excluded: not a two-armed design [1]
Bishop, Cheuvront and Elston	AutoCyte SCREEN. Tested the use of residual material from cervical samples for later analysis. 99 cases. LBC and DNA profiling. USA	Combined cytological diagnosis from original LBC smear and results of DNA profile	Excluded: ineligible reference standard [2]
Bishop, Kaufman and Taylor	AutoCyte SCREEN + PREP vs PREP. 1676 smears. USA	Discordant results, all abnormal 5% normals evaluated independently by one cytopathologist	Excluded: reference standard. No consensus opinion from two or more cytologists [3]
Bishop, Cheuvront and Sims	AutoCyte SCREEN + PREP vs PREP. 1992 smears. USA	Two cytopathologists reviewed cytology, but second cytopathologist's view taken as reference diagnosis	Excluded: reference standard. No consensus opinion from two or more cytologists [3]
Brotzman *et al.*	PAPNET rescreen vs original diagnosis from 10% random rescreen. 1200 smears. USA	Cytology: one cytopathologist's review of abnormal smears identified by PAPNET	Excluded: not a two-armed design [1]
Chhieng *et al.*	PAPNET rescreen of 91 cases of AGUS. USA	Histological biopsy of all smears within 1 year	Excluded: not a two-armed design [1]

continued

TABLE 90 *Studies from update search reaching step 2 (n = 45) (cont'd)*

Authors	Description of study	Standard	Inclusion/exclusion[a]
Cosentino *et al.*	PAPNET vs manual (original diagnosis). Primary screening. 397 smears. Imola, Italy	Cytology: three cytopathologists reviewed subset of discordant cases, when one arm abnormal and the other negative. Discordance between classes of abnormality not verified	Excluded: reference standard. Not all discordant slides verified [3]
Doornewaard, van der Schouw, van der Graaf, Bos, Habbema *et al.*	PAPNET Primary screener vs conventional. Cohort of 6063 smears. The Netherlands	Normal smears had 7 years of follow-up as negative. Abnormal cases nearly all had biopsy in 7 years of follow-up	Included
Doornewaard, van der Schouw, van der Graaf, Bos and van den Tweel	PAPNET vs PAPNET vs Conventional testing of interobserver and intraobserver error using 6063 smears. The Netherlands	Normal smears had 7 years of follow-up as negative. Abnormal cases nearly all had biopsy in 7 years of follow-up	Excluded: unable to calculate both sensitivity and specificity [5]
Doornewaard, van der Schouw & van der Graaf	PAPNET vs PAPNET at different times. Test devices ability to reproduce information from the same set of slides. 196 smears. The Netherlands	Normal smears had 7 years of follow-up as negative. Abnormal cases nearly all had biopsy in 7 years of follow-up	Excluded: unable to calculate both sensitivity and specificity [5]
Duggan	PAPNET primary screening vs conventional. 2200 archived smears. Canada	Consensus diagnosis by more than two cytopathologists on discrepant results between the two methods	Excluded: not a two-armed design [1]
Fetterman *et al.*	AutoPap 300 QC rescreening of 35,143 smears from a 3-month period compared with conventional over previous 6 months. Historical trial USA	All HSIL+ had biopsy, otherwise single cytopathologist's diagnosis on cytology	Excluded: not a two-armed design [1]
Ghidoni *et al.*	PAPNET in primary screening vs original diagnosis from routine practice. 1654 smears. Imola, Italy	Discrepant cases reviewed by three cytologists. Diagnosis = two concordant opinions out of the three. Concordant results unverified	Excluded: not a two-armed design [1]
Ghidoni *et al.* [Italian]	PAPNET rescreening vs original diagnosis. 1309 negative smears from original report of primary screen. Imola, Italy	Cytology. Consensus opinion of three cytologists on discrepant slides between PAPNET and original report	Excluded: not a two-armed design [1]
Halford *et al.*	PAPNET +manual + rapid rescreen vs manual screen + rapid rescreen of negative cases. 25,656 smears. Australia	All high-grade and less than one-third of low-grade lesions had biopsy	Excluded: not a two-armed design [1]

continued

TABLE 90 *Studies from update search reaching step 2 (n = 45) (cont'd)*

Authors	Description of study	Standard	Inclusion/exclusion[a]
Hølund *et al.*	1500 slides prescreened by PAPNET, with negatives rescreened manually. Further 1500 negatives from manual routine screen, rescreened by PAPNET. Denmark	Not explicitly described, although some (quantity unspecified) had histology follow-up. Negatives not verified	Excluded: not a two-armed design [1]
Howell *et al.*	AutoCyte SCREEN + Prep vs Prep. 856 smears. USA	No clear consistent reference standard, but one reviewing cyotopathologist used in some cases	Excluded: no clear reference standard [2]
Huang *et al.*	AutoPap with LGS vs manual. 400 smears (52 then excluded owing to process problems). Taiwan	Discrepant smears resolved by one cytopathologist	Excluded: reference standard. No consensus opinion from two or more cytologists [3]
Kok, Habers, Schreiner-Kok and Boon	PAPNET vs manual. 2000 smears. Both methods used to select a subset of ASCUS smears (n = 168) which were then subject to further PAPNET screen using alternative criteria for diagnosing ASCUS. The Netherlands	Biopsy on ASCUS cases only	Excluded: not a two-armed design [1]
Kok, Boon, Schreiner-Kok and Koss	PAPNET vs conventional in primary screening. 245,527 vs 109,104 smears in statistically equivalent populations. RCT. The Netherlands	Biopsy confirmation when high-grade lesion reported on cytology	Excluded: not a two-armed design [1]
Kok, Schreiner-Kok, van der Veen and Boon	Rescreening of 40 true-positive SCCs identified by PAPNET and eight false-negative SCCs missed by conventional screening. Analysis of images. The Netherlands	Histology	Excluded: not a two-armed design [1]
Lee *et al.*	AutoPap LGS vs reference standard. 683 smears, with 532 successfully processed. USA	Neopathology cytotechnician screened slides and compared with original laboratory report. Where discordant, slide adjudicated by one neopath cytopathologist	Excluded: not a two-armed design [1]
Lerma *et al.*	PAPNET vs manual. Rescreening of 163 cases of ASCUS. Spain	Colposcopy: if positive then biopsy (=111/163). If negative then followed up by three consecutive annual smears	Excluded: unable to calculate both sensitivity and specificity [5]
Losell and Dejmek	PAPNET vs original diagnosis from routine screening on 1000 consecutive smears. Sweden	Discrepant cases subject to re-evaluation by four senior cytotechnicians	Excluded: not a two-armed design [1]
Mackin *et al.*	Deconvolution and segmentation classifiers of 808 nuclei from 31 ThinPrep smears. USA	Original cytological diagnosis	Excluded: not a two-armed design [1]

continued

TABLE 90 *Studies from update search reaching step 2 (n = 45) (cont'd)*

Authors	Description of study	Standard	Inclusion/exclusion[a]
Mango and Valente	PAPNET vs manual. Rescreening of negative smears with seeding of high-grade smears. 2293 PAPNET vs 13761 conventional. USA	Biopsy-confirmed high-grade cases	Excluded: not a two-armed design [1]
Marshall *et al.*	AutoPap 300 QC vs manual + 10% random rescreen. Historical trial of rescreening. March to August 1996 (35,027 smears) vs March to August 1997 (31,951). USA	Pathologist's diagnosis	Excluded: not a two-armed design [1]
Minge *et al.*	AutoCyte system (Prep + Screen) vs conventional smears. 2156 paired samples. USA	With discordant results the more abnormal finding taken as the diagnosis after one cytopathologist's review. Biopsy on 134/2156 cases	Excluded: reference standard. No consensus opinion from two or more cytologists [3]
Mitchell and Medley	PAPNET review of biopsy- and cytology-proven abnormalities by three cytotechnicians to measure interobserver variation. 164 smears. Australia	Biopsy of all smears reviewed	Excluded: not a two-armed design [1]
Mitchell and Medley	PAPNET rescreen of 20,000 smears. 19,805 were negative after two screens. 195 abnormal. Machine repeatability measured on 2690 smears. Australia	Biopsy of high-grade abnormalities. Discordant smears adjudicated by fewer than three cytopathologists	Excluded: not a two-armed design [1]
Mitchell and Medley	As above. Australia	Biopsy of high-grade abnormalities. Discordant smears adjudicated by fewer than three cytopathologists	Excluded: not a two-armed design [1]
O'Leary *et al.*	PAPNET review of 5478 negative smears after manual + 10% random rescreen. Outcomes: effectiveness and cost. USA	Panel of three cytopathologists reviewed discrepant cases. No HSIL cases	Excluded: not a two-armed design [1]
Romeo *et al.*	Infrared absorption spectra of 120 smears using neural networks to assist diagnosis. Australia	Cytological diagnosis	Excluded: not a two-armed design [1]
Schneider *et al.*	Conventional cytology vs HPV vs colposcopy. 5455 slides with PAPNET used as quality assessment of 1314 slides. Germany	Additional screening at 4–8 months for those negative to the three methods. Abnormal slides referred for biopsy or curettage, interpreted by one pathologist. High-grade abnormalities had consensus opinion of two pathologists in addition to the one above	Excluded: not a two-armed design [1]

continued

TABLE 90 *Studies from update search reaching step 2 (n = 45) (cont'd)*

Authors	Description of study	Standard	Inclusion/exclusion[a]
Shaw *et al.*	Infrared absorption spectra with automated feature extraction algorithms. 800 spectra vs gold standard. Canada	Cytology and histology: for cytological diagnosis not stated how many cytopathologists were involved. Biopsy also taken, but discrepancy between cytology and histology	Excluded: not a two-armed design [1]
Sherman *et al.*	PAPNET vs manual vs HPV. 7327 women subjected to a split-sample smear, DACRON swab and cervigram at initial examination. Costa Rica	Based on all pathology material available. Colposcopic referral after positive pelvic examination, cytology or cervigram. Biopsy on 93% HSIL, 100% cancer and 39% LSIL	Included
Sturgis *et al.*	PAPNET vs reference standard. 61 smears of varied glandular abnormalities from archives. USA	Diagnosis from cytology, and biopsy where appropriate	Excluded: not a two-armed design [1]
Takahashi *et al.*	AutoCyte system (Prep + SCREEN) vs conventional/Prep. 583 smears. Japan	Cytology: discrepant cases subject to review by one cytotechnologist	Excluded: reference standard. No consensus opinion from two or more cytologists [3]
van Ballegooijen *et al.*	PAPNET primary screen followed by conventional rescreen vs conventional primary screen followed by PAPNET rescreen. 500 in each arm taken from routine programme. The Netherlands	Cytopathologist's diagnosis	Excluded: ineligible reference standard [2]
Veneti *et al.*	PAPNET rescreening vs manual rescreen of 24 negative smears which later developed precancerous lesions. Greece	Biopsy, but 11/24 cases were up to 2 years after cytology	Excluded: unable to calculate both sensitivity and specificity [5]
Wertlake	AutoPap QC rescreening (February 1997 to April 1998) vs 10% random rescreening of negative cases (October 1995 to January 1997). Historical trial. 550,076 smears vs 591,837 smears. USA	Cytopathologist's diagnosis in referred cases	Excluded: not a two-armed design [1]
Wilbur *et al.*, 1998	AutoPap (primary + QC) vs manual + QC (10% random rescreen). 25,124 smears. USA	Discrepant cases reviewed by a panel of three cytopathologists and consensus opinion taken as diagnosis	Excluded: no biopsy confirmation of HSIL [4]
Wilbur *et al.*, 1999	As above. USA	As above, with 27/70 HSIL+ having biopsy confirmation	Excluded: biopsy confirmation of HSIL <50% [4]

[a] The stage of step 2 where the study failed is indicated in brackets.
SCC, squamous cell carcinoma.

TABLE 91 *Studies from foreign-language search reaching step 2* (n = *0)*

Authors	Description of study	Standard	Inclusion/exclusion
No studies reaching step 2			

TABLE 92 *Studies from McCrory* et al. *'near-misses' reaching step 2* (n = *12)*

Authors	Description of study	Standard	Inclusion/exclusion[a]
Colgan *et al*.	AutoPap QC rescreen vs manual rescreen. 3487 normal slides. Canada	Cytology: abnormal cases found by manual arm only, reviewed by blinded independent panel. Abnormal cases from automated arm not reviewed	Excluded: not a two-armed design. Two arms not treated equally [1]
Patten *et al*.	AutoPap 300 QC vs manual (original reading). 2339 slides prospectively recruited with 3028 slides taken from the archives. Diverse case-mix. USA	No reference standard described	Excluded: not a two-armed design [1]
Stevens *et al*.	AutoPap QC 10, 20, 30% review rate vs computer-simulated manual random rescreen 10, 20, 30% of 1840 normal smears. AutoPap QC 10, 20, 50% rescreen of archived abnormal (*n* = 139) and FN (*n* = 40) smears. Australia	Slides discordant with original diagnosis reviewed by panel of two experts. Unreviewed negatives rescreened; false negatives found subject to panel review. Abnormals and false negatives verified by biopsy in a large proportion	Excluded: not a two-armed design [1]
Wilbur *et al*.	Rescreening of 86 archived cases of HSIL by AutoPap QC at 10% and 20% review rates. USA	Histology: biopsy on all 86 cases of HSIL	Excluded: not a two-armed design [1]
Ashfaq, Liang and Saboorian	PAPNET rescreening vs manual rescreening. 2238 negatives rescreened by PAPNET vs 2000 reviewed manually. USA	Cytology: unsatisfactory or atypical slides by PAPNET reviewed by a single cytopathologist	Excluded: not a two-armed design [1]
Ashfaq, Saliger, Solares *et al*.	PAPNET primary screening vs manual prospective. 5170 smears. USA	Cytology: discordant slides reviewed by a single cytopathologist	Excluded: reference standard. No consensus opinion from two or more cytologists [3]
Duggan and Brasher	PAPNET primary screen vs manual in blind comparison. 5037 consecutive smears. Canada	One pathologist reviewed all abnormal smears. Discordant cases reviewed by a panel of two experts	Excluded: no biopsy confirmation of HSIL+ [4]

continued

TABLE 92 *Studies from McCrory* et al. *'near-misses' reaching step 2 (*n = *12) (cont'd)*

Authors	Description of study	Standard	Inclusion/exclusion[a]
Farnsworth *et al*.	PAPNET rescreening vs manual screening. 54,658 normal and 1022 abnormal slides. Australia	Cytology: verification by a single cytopathologist. Histological follow-up on a small subset of abnormalities	Excluded: reference standard. No consensus opinion from two or more cytologists [3]
Halford *et al*.	PAPNET rescreening vs rapid manual rescreening. 1020 slides (1000 normal and 20 seeded abnormals). Australia	No reference standard described	Excluded: no reference standard [2]
Jenny *et al*.	PAPNET rescreening vs PAPNET rescreening. 1200 smears, to measure interobserver variation. Switzerland	All abnormals (n = 516) had biopsy, negative smears had 2-year follow-up	Excluded: unable to calculate sensitivity and specificity [5]
Kaufman *et al*.	PAPNET rescreening of 160 ASCUS cases vs reference standard. USA	Histology: all 160 cases had biopsy	Excluded: not a two-armed design [1]
Slagel, Zaleski and Cohen	PAPNET primary screening vs original diagnosis. 500 slides. USA	Cytology: discordant cases were reviewed by two pathologists	Excluded: not a two-armed design [1]

[a] The stage of step 2 where the study failed is indicated in brackets.

Appendix 8

Using verification of discordant cytology as a reference standard

Theorem

In the analysis of cervical smears, if a cytological reference standard is used where only discordant slides are verified, and

> Measured sensitivity of test 1 ≥ Measured sensitivity of test 2

then

> Absolute sensitivity of test 1 ≥ Absolute sensitivity of test 2

where the absolute sensitivity is that obtained using a perfect gold standard to the whole population.

Proof

Let a, b, c and d be defined by their position in the following 2×2 contingency table

		Test 2	
		+	−
Test 1	+	a	b
	−	c	d

where a = assumed true positives for reference standard (i.e. no concordant verification) and d = assumed true negatives of reference standard (i.e. no concordant verification). Only discordant cells (i.e. b and c) were verified.

Notation

TP, TN, FP and FN are all values of the underlying population, that is, the truth if a perfect gold standard had been applied. Note: for the purpose of this analysis verification of discordant cytology is considered perfect, that is, there are no errors in the verification process.

A superfix refers to the position of a variable 2×2 table and a suffix to either test 1 or test 2. For example, the variable

$$TP_1^b$$

refers to the number of true positives of test 1 in cell b. These happen to be known because it is a discordant cell.

Let an overscore imply that the value is not known (i.e. unverified by the reference standard). Thus, because concordant positives are not verified (hence overscore), then for both test 1 and test 2 we have

$$a = \overline{TP}_1^a + \overline{FP}_1^a = \overline{TP}_2^a + \overline{FP}_2^a$$

But in any particular cell, because a true positive for one test cannot be a false positive for another, with respect to the perfect gold standard, then

$$\overline{TP}_1^a = \overline{TP}_2^a$$
$$\overline{FP}_1^a = \overline{FP}_2^a$$

As cell b consists of discordant cytology, all slides are known with respect to the reference standard (hence no overscore)

$$b = TP_1^b + FP_1^b = TN_2^b + FN_2^b$$

But since being a true positive of one test with respect to the perfect gold standard is mutually exclusive, it cannot be a true negative of the other test as well, then it must equate to the false negatives of the other test. Hence,

$$TP_1^b = FN_2^b$$
$$FP_1^b = TN_2^b$$

For similar arguments we have

$$c = TN_1^c + FN_1^c = TP_2^c + FP_2^c$$
$$TN_1^c = FP_2^c$$
$$FN_1^c = TP_2^c$$

and

$$d = \overline{TN}_1^d + \overline{FN}_1^d = \overline{TN}_2^d + \overline{FN}_2^d$$
$$\overline{TN}_1^d = \overline{TN}_2^d$$
$$\overline{FN}_1^d = \overline{FN}_2^d$$

Let

Measured sensitivity of test1 ≥ Measured sensitivity of test 2

The usual formula for sensitivity is

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{false negatives}}$$

With respect to discordant cytology verification as the reference standard, the measured total of true positives for test 1 is

Total true positives = all positives in cell a + those true positives verified in cell b.

Since in cell a the slides are unverified, we cannot distinguish absolute true positives from absolute false positives (with respect to a perfect gold standard), whereas in cell b we can, since they are verified.

The measured false negatives of test 1 have to be, by definition, those that are measured as negative by test 1 but are in fact positive with respect to the measured reference standard (verified discordant cytology), that is

FN_1^c

Note that the false negatives of test 1 in cell d (a concordant cell) cannot be measured using discordant cytology verification as the reference standard.

Hence, the left-hand side of the equation below is the measured sensitivity for test 1; similarly, the right-hand side is the measured sensitivity for test 2

$$\frac{(\overline{TP}_1^a + \overline{FP}_1^a) + TP_1^b}{(\overline{TP}_1^a + \overline{FP}_1^a + TP_1^b) + FN_1^c} \geq \frac{(\overline{TP}_2^a + \overline{FP}_2^a) + TP_2^c}{(\overline{TP}_2^a + \overline{FP}_2^a + TP_2^c) + FN_2^c} \geq \frac{(\overline{TP}_1^a + \overline{FP}_1^a) + TP_2^c}{(\overline{TP}_1^a + \overline{FP}_1^a + FN_1^c) + TP_1^b}$$

$$\Rightarrow \overline{TP}_1^a + TP_1^b \geq \overline{TP}_1^a + TP_2^c$$

$$\Rightarrow TP_1^b \geq TP_2^c$$

$$\therefore \frac{(\overline{TP}_1^a + TP_1^b)}{(\overline{TP}_1^a + TP_1^b) + (FN_1^c + \overline{FN}_1^d)} \geq \frac{(\overline{TP}_1^a + TP_2^c)}{(\overline{TP}_1^a + TP_1^b) + (FN_1^c + \overline{FN}_1^d)}$$

But

$$\overline{TP}_1^a = \overline{TP}_2^a$$
$$TP_1^b = FN_2^b$$
$$FN_1^c = TP_2^c$$

and

$$\overline{FN}_1^d = \overline{FN}_2^d$$

$$\therefore \frac{(\overline{TP}_1^a + TP_1^b)}{(\overline{TP}_1^a + TP_1^b) + (FN_1^c + \overline{FN}_1^d)} \geq \frac{(\overline{TP}_2^a + TP_2^c)}{(\overline{TP}_2^a + FN_2^b) + (TP_2^c + \overline{FN}_2^d)}$$

that is,

Absolute sensitivity of test 1 ≥ Absolute sensitivity of test 2

since in the perfect world all true and false positives and negatives are known.

Appendix 9

General considerations concerning achieving representativeness of literature included in systematic reviews and health technology assessment

When a review's search has been confined to the published literature of a new technology say, it forces the reviewer to consider the following question: Are the results from the published literature a fair representation of the performance of the new technology?

To answer this question it is useful to break it down into three separate problems:

1. to estimate theoretically the extent of unpublished studies on the new technologies
2. to prove their existence, by a method that systematically identifies the unpublished studies
3. to evaluate the effects of those studies identified on the conclusion of performance drawn from the published studies.

Each of these problems is addressed under appropriate subheadings in the next section.

Theoretical considerations

1. Theoretical techniques

The first problem involves the application of published techniques, such as funnel plots to derive a theoretical estimate of the level of unpublished studies. Theoretical estimates of the effects of these studies have also been proposed, as have already been discussed. These techniques are comprehensively dealt with in a number of reviews and are not expanded upon here.

2a. Existence and completeness

The second question requires development. To date there is no recognised general methodology to tackle the problem of proving the existence of unpublished studies. Specific problems have been tackled, however, with admirable efforts.[103]

Any area of study will have a number of sources of publication worldwide. As such, theoretically all countries should be trawled for evidence of unpublished studies before concluding on the exact level of publication bias. This would represent a formidable task and could be regarded as impossible in practical terms. If the exact answer will never be known, then that raises the question of what is the merit in searching for unpublished studies in less than the total number of possible sources?

First, to prove the existence of unpublished studies, only one study needs to be identified to satisfy this objective, and this does not usually require an all-encompassing search.

Second, the motivation behind striving for completeness relates to evaluating the aggregate effect of the studies on the review. To gain a measure of this effect does not necessarily require sifting all data sources. Unpublished studies identified from a subtotal search still give some measure of the effect that they have on the mean results from the published data. The associated confidence interval of the mean statistic is wider than would be expected from a complete search, but it must be borne in mind that the complete trawl is a theoretical ideal that is never achieved.

2b. Coverage

To give the study statistical power and reduce the level of uncertainty as far as possible, the issue of coverage needs addressing. If it is assumed that the set of unpublished studies and the published studies are drawn from the same population, then a starting hypothesis is that the countries responsible for the majority of published studies are also responsible for the unpublished studies. An assumed monotonically increasing relationship between the two would result in efforts being directed towards those countries that produced the most publications on the subject. Such an approach is borne out by at least one major study.[103]

A search for primary studies in the published literature and categorising the results by country would determine the most productive countries. However, the larger countries with the highest number of journals will consistently appear at the top of the list (e.g. the USA), but may not be the most accessible in terms of identifying unpublished studies.

Thus, maximum coverage may not always be achievable, and the participation of a particular country could well be determined by externally imposed constraints such as practicality, rather than coverage. In such instances targeted searches have been used.[103]

2c. Accessibility and manageability

Success in the investigation of publication bias is probably more dependent on adopting a systematic approach than on any other factor. For a trawl through a country's unpublished literature to be systematic relies to a large degree on the pre-existing systems and infrastructure. For instance, a country that has a clear partition of its health system, with each subunit having a well-defined division of labour that is reproduced with fidelity across all subunits, will greatly facilitate a coordinated approach and maximise the chances of total coverage. For a given healthcare system, it is likely that some areas of health will have better infrastructures than others, depending on a number of factors such as funding, priority and length of time for which the system has been operating. Nevertheless, the lack of an organised system with good communication lines will make a search for any unpublished literature more difficult.

Even in the presence of a good infrastructure, there may be other reasons for not including a particular country as part of the search. The sheer size of the task associated with a particular country may prohibit an attempt, based on resource and time constraints. The success of organising a search that extends to a number of countries may be ultimately tied to the cooperation of key figures in those countries. If the scale of a country is such that the number of key individuals becomes large, then ensuring good cooperation across all these individuals will be more difficult, reducing the chances of wide coverage. It may be assumed, owing to its vastness, that the USA holds such problems.

3a. Study quality

Following identification of a set of unpublished studies, the next stage is to evaluate them. Although feedback and peer review tend to come much later than at the design stage, those articles that are eventually published (and those that are not) have undergone a process of selection, albeit of unknown objectivity. To this extent, one may expect to find a larger representation of flawed designs in the unpublished literature, as rejected articles are a particular subset of the unpublished literature.[156] Nonetheless, to maintain equity, such studies should be subject to the same rigorous critique as the published studies. This may mean in some cases that no conclusion could be drawn from the study.

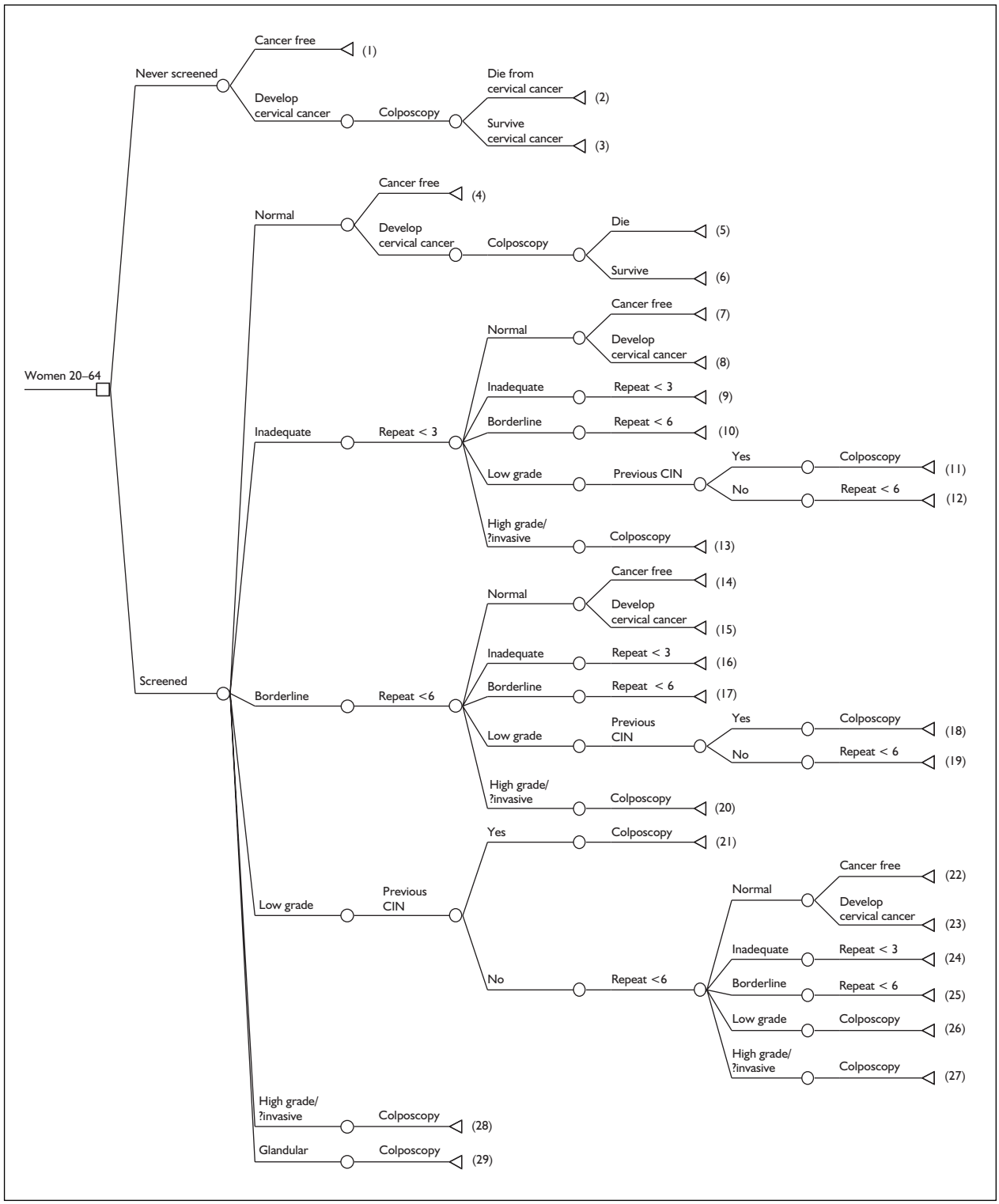
Some have suggested that "authors do not submit their work for publication because they have become aware of its limitations", and they go on to suggest that such literature should not be considered at all in meta-analyses if accurate answers are sought.[157,158] There may be some truth in the first part of the statement. However, the exclusion of such studies before evaluation and critique, on the basis of the a priori assumption that such studies are in some way limited in methodology, would run the risk of leaving out valuable information.

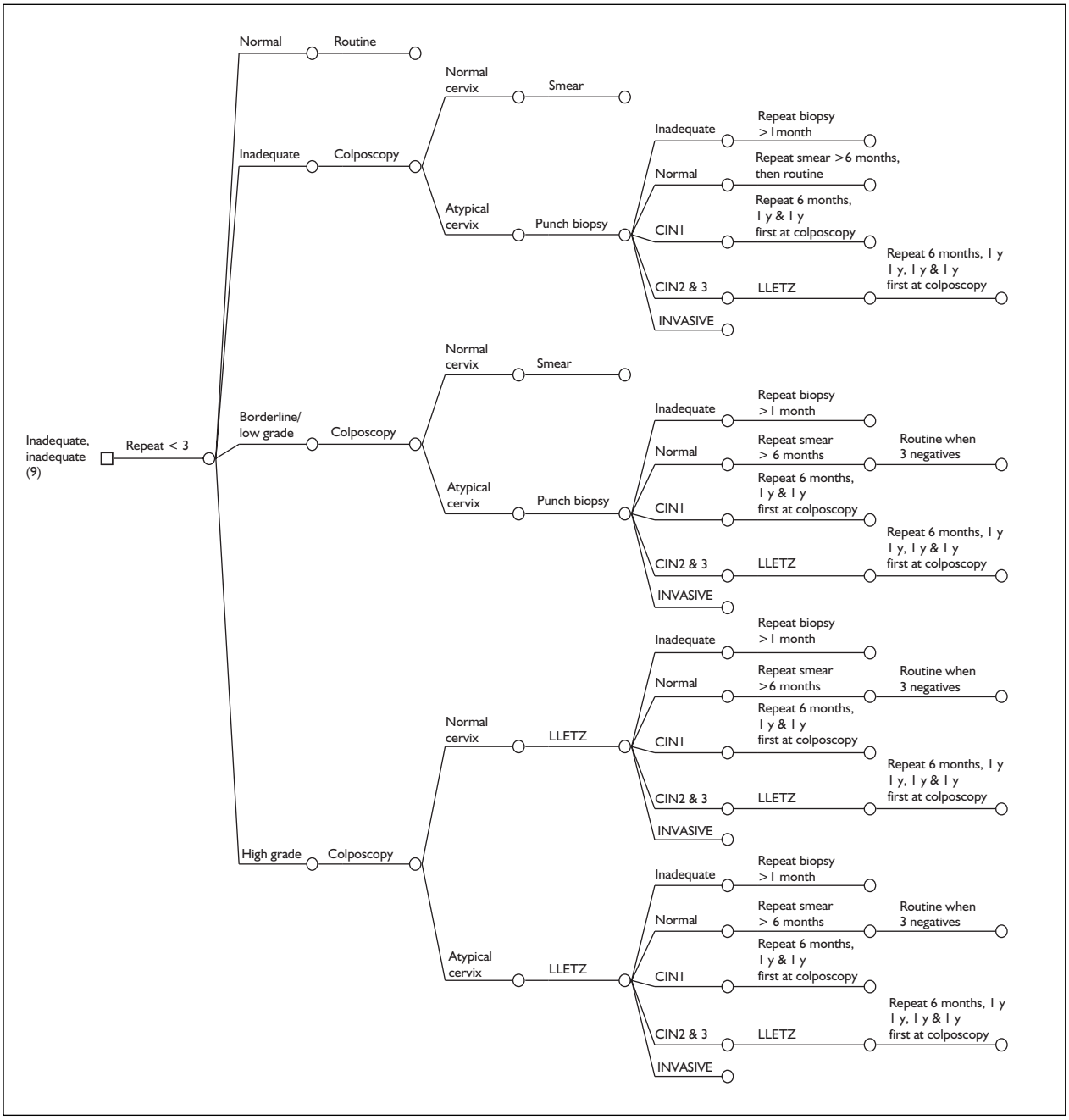
3b. Performance metric

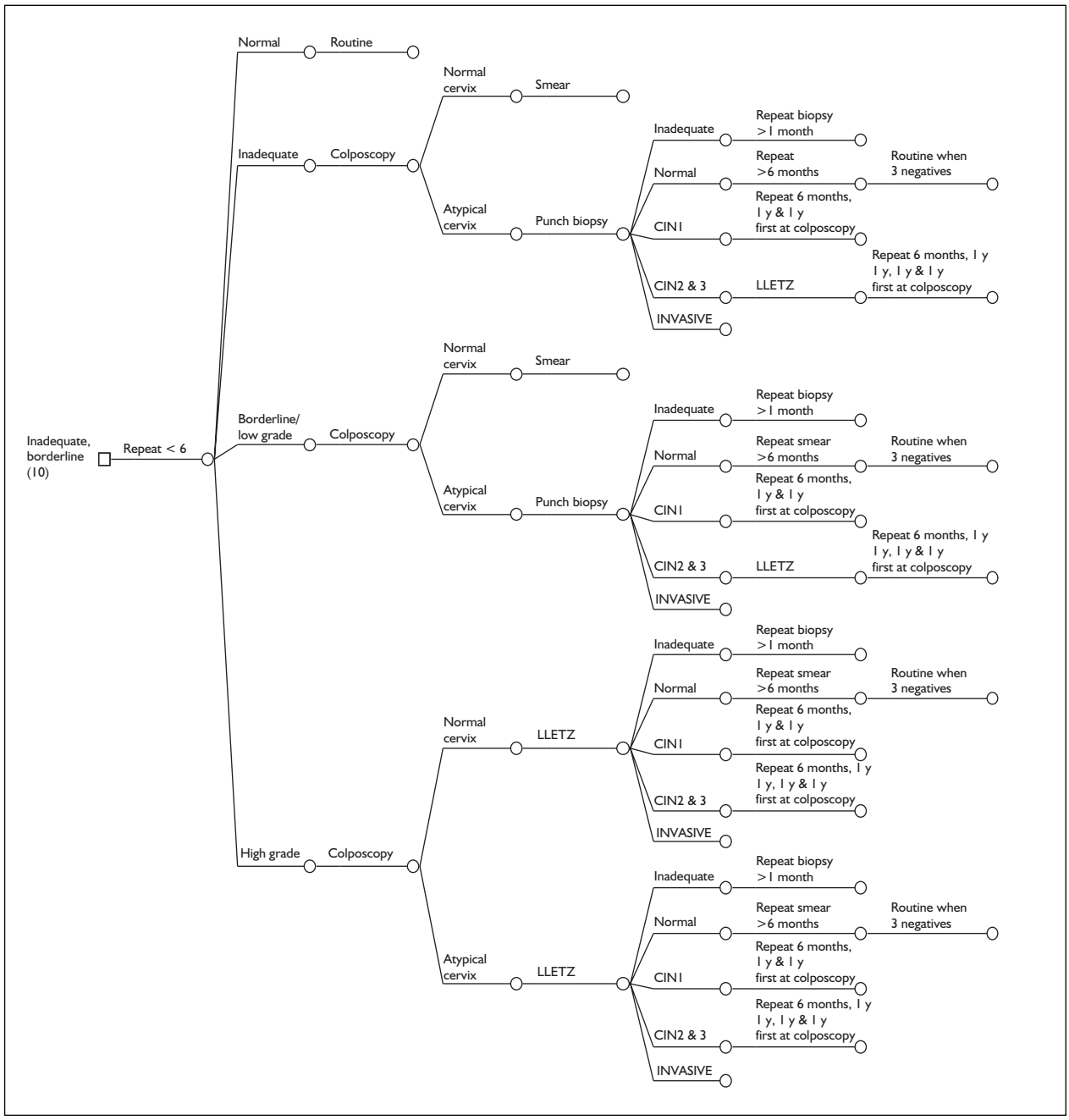
If there are sufficient data present in the unpublished work to perform basic data analysis, then the same metrics should be used for comparison as used for the published literature.

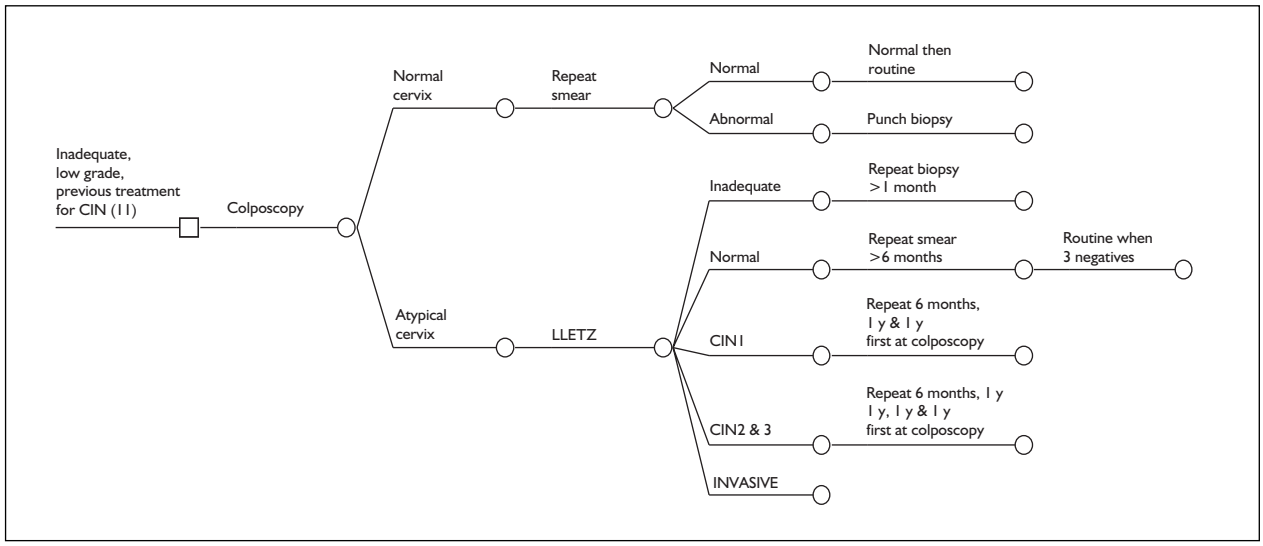
Appendix 10

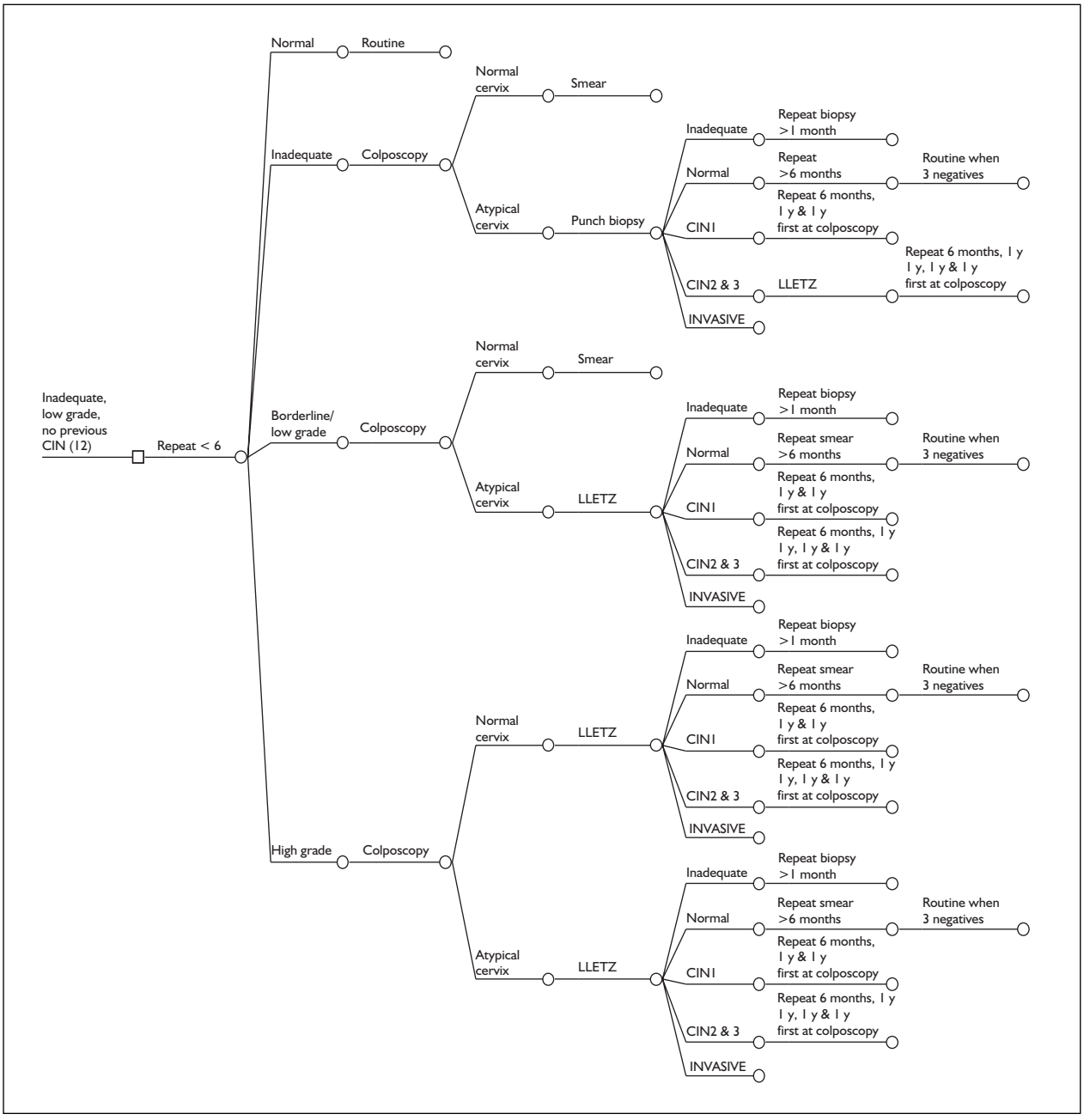
Patient walk-through for the NHS Cervical Screening Programme

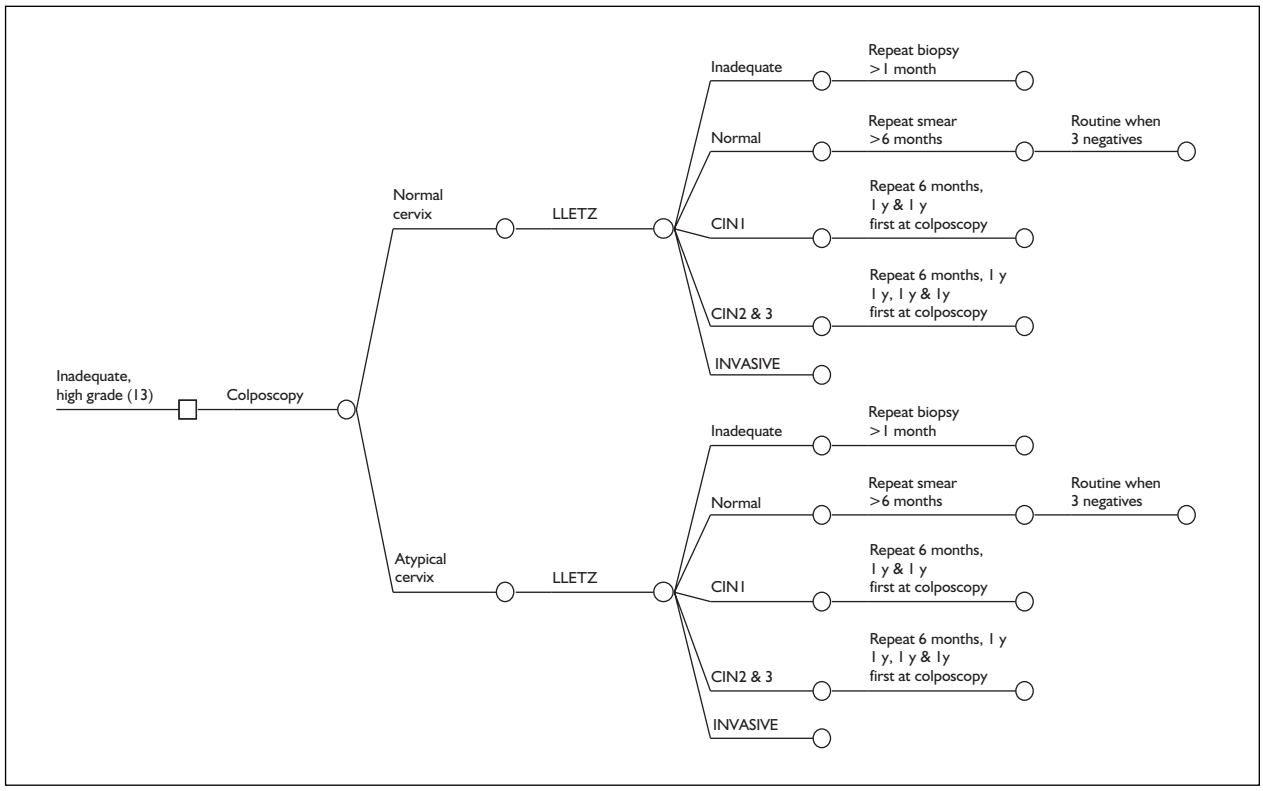

Women 20–64
Never screened
Cancer free
(1)
Develop cervical cancer
Colposcopy
Die from cervical cancer
(2)
Survive cervical cancer
(3)
Screened
Normal
Cancer free
(4)
Develop cervical cancer
Colposcopy
Die
(5)
Survive
(6)
Inadequate
Repeat < 3
Normal
Cancer free
(7)
Develop cervical cancer
(8)
Inadequate
Repeat < 3
(9)
Borderline
Repeat < 6
(10)
Low grade
Previous CIN
Yes
Colposcopy
(11)
No
Repeat < 6
(12)
High grade/ ?invasive
Colposcopy
(13)
Borderline
Repeat <6
Normal
Cancer free
(14)
Develop cervical cancer
(15)
Inadequate
Repeat < 3
(16)
Borderline
Repeat < 6
(17)
Low grade
Previous CIN
Yes
Colposcopy
(18)
No
Repeat < 6
(19)
High grade/ ?invasive
Colposcopy
(20)
Low grade
Previous CIN
Yes
Colposcopy
(21)
No
Repeat <6
Normal
Cancer free
(22)
Develop cervical cancer
(23)
Inadequate
Repeat < 3
(24)
Borderline
Repeat < 6
(25)
Low grade
Colposcopy
(26)
High grade/ ?invasive
Colposcopy
(27)
High grade/ ?invasive
Colposcopy
(28)
Glandular
Colposcopy
(29)

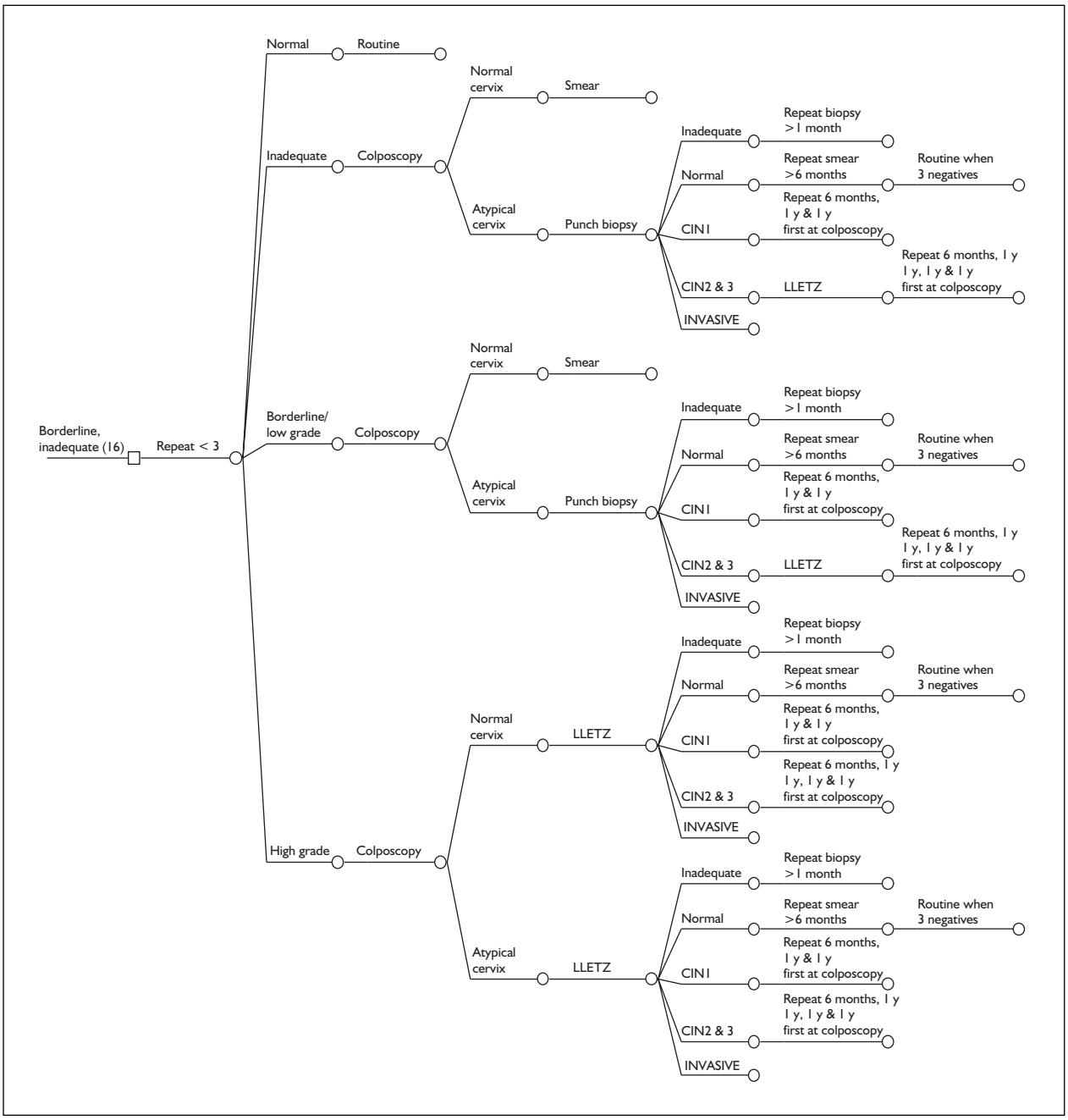


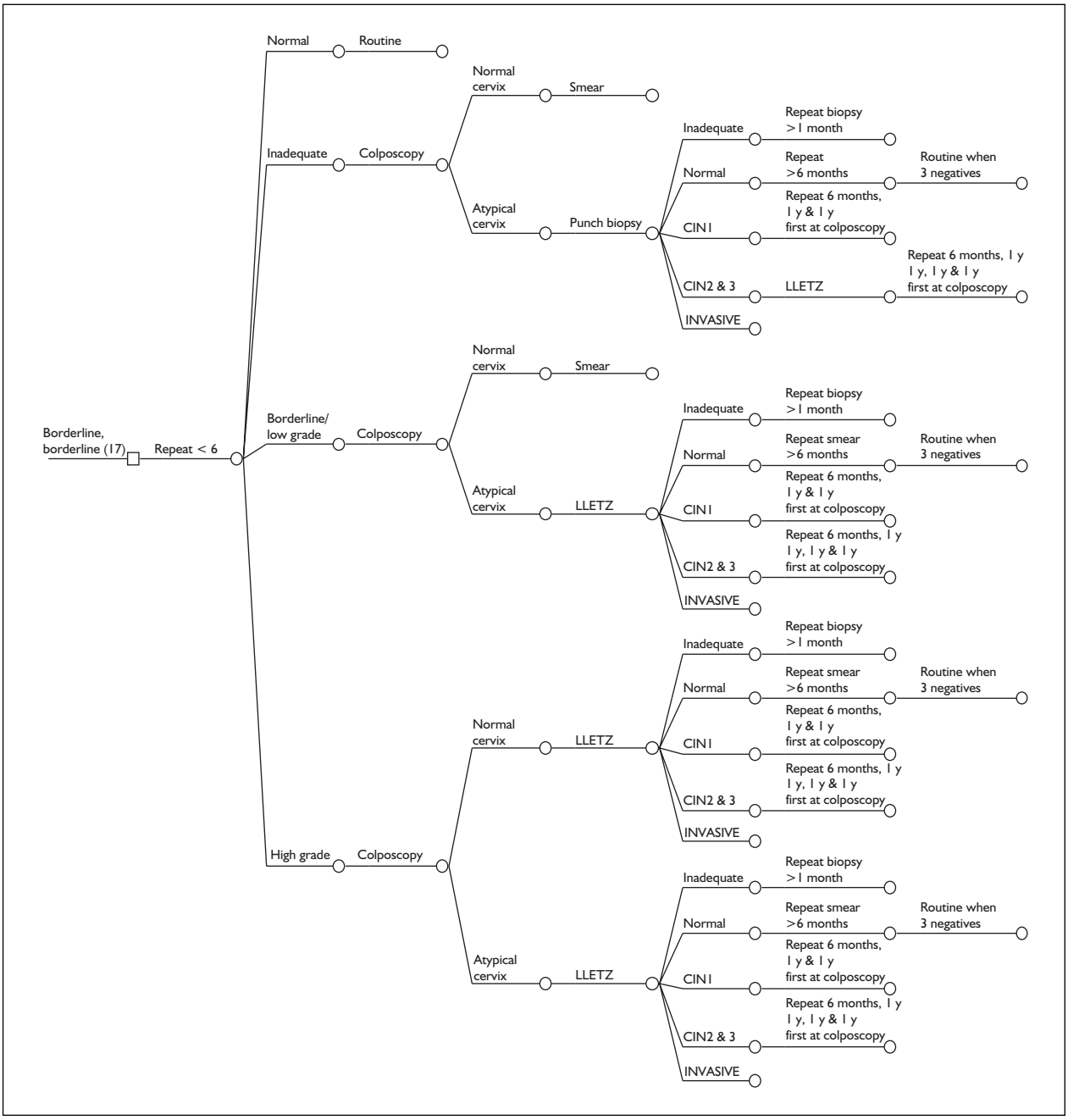

Inadequate, borderline (10)
Repeat < 6
Normal
Routine
Inadequate
Colposcopy
Normal cervix
Smear
Atypical cervix
Punch biopsy
Inadequate
Repeat biopsy >1 month
Normal
Repeat >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
LLETZ
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
Borderline/ low grade
Colposcopy
Normal cervix
Smear
Atypical cervix
Punch biopsy
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
LLETZ
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
High grade
Colposcopy
Normal cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE

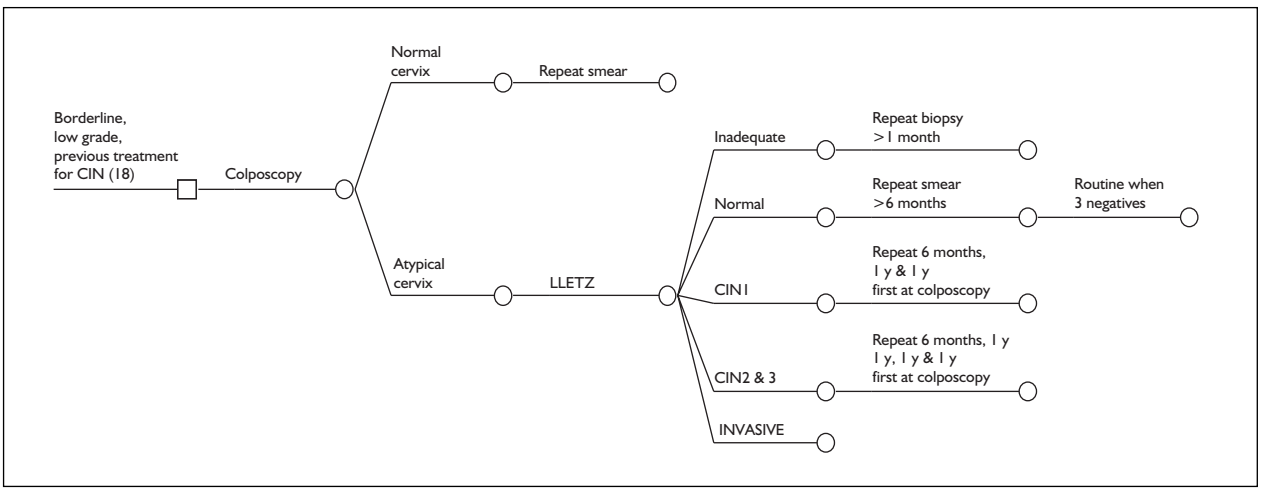

Inadequate, low grade, previous treatment for CIN (11)
Colposcopy
Normal cervix
Repeat smear
Normal
Normal then routine
Abnormal
Punch biopsy
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE

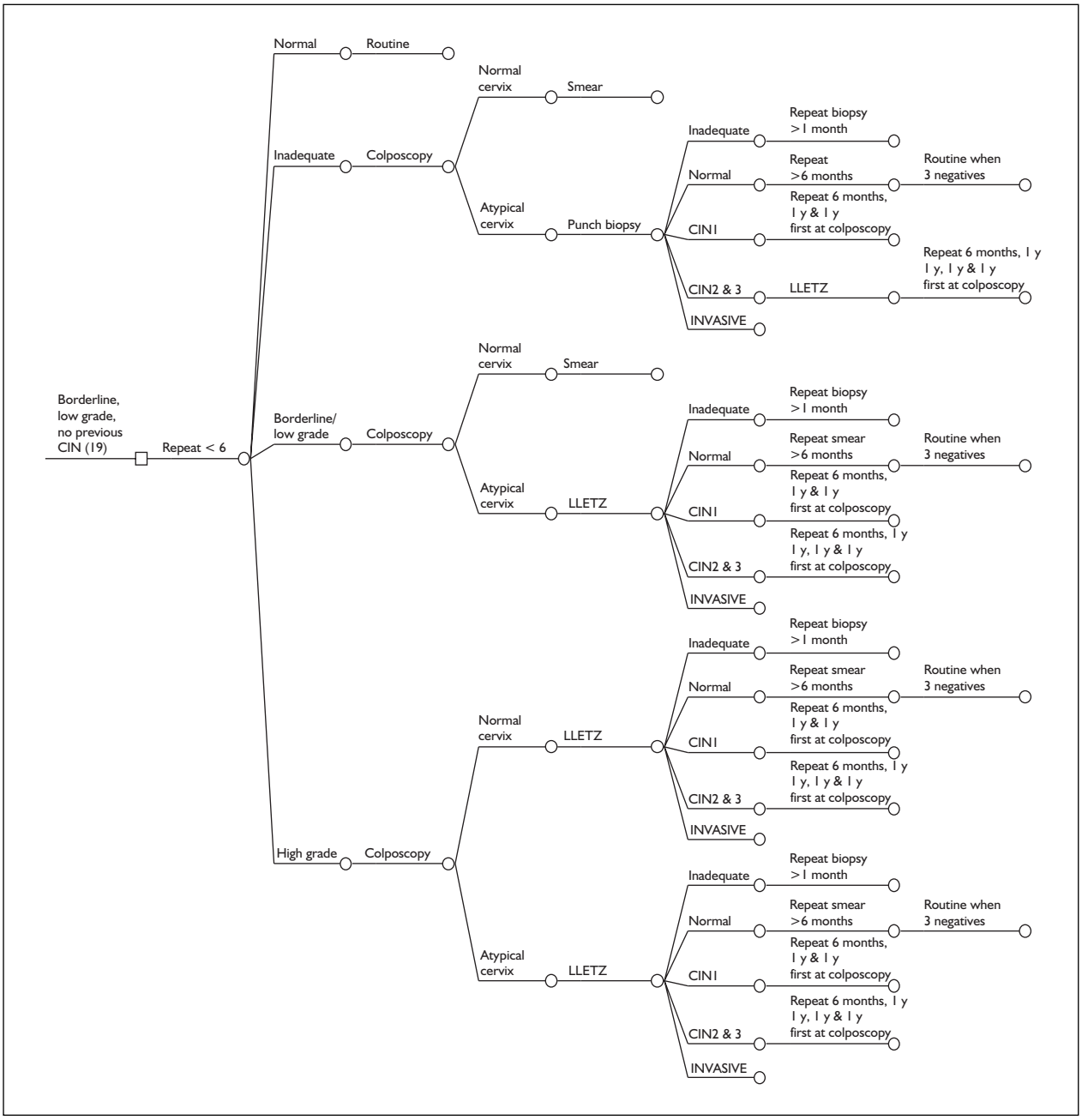


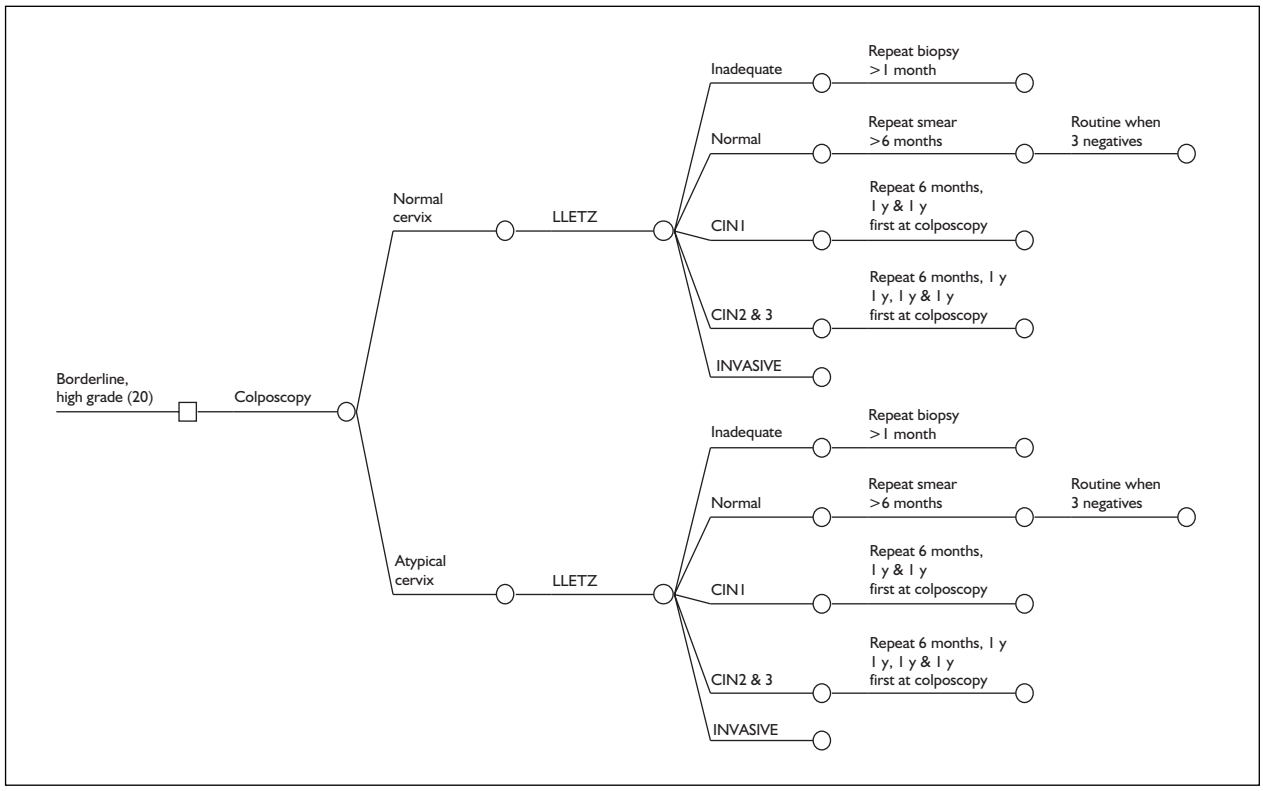

Inadequate, high grade (13)
Colposcopy
Normal cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1y first at colposcopy
INVASIVE
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1y first at colposcopy
INVASIVE

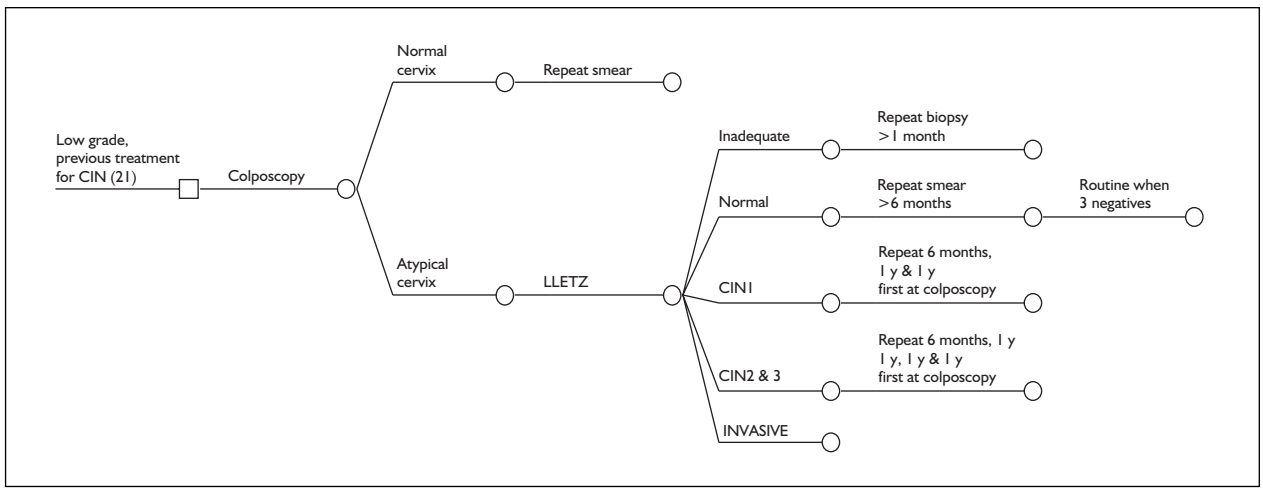


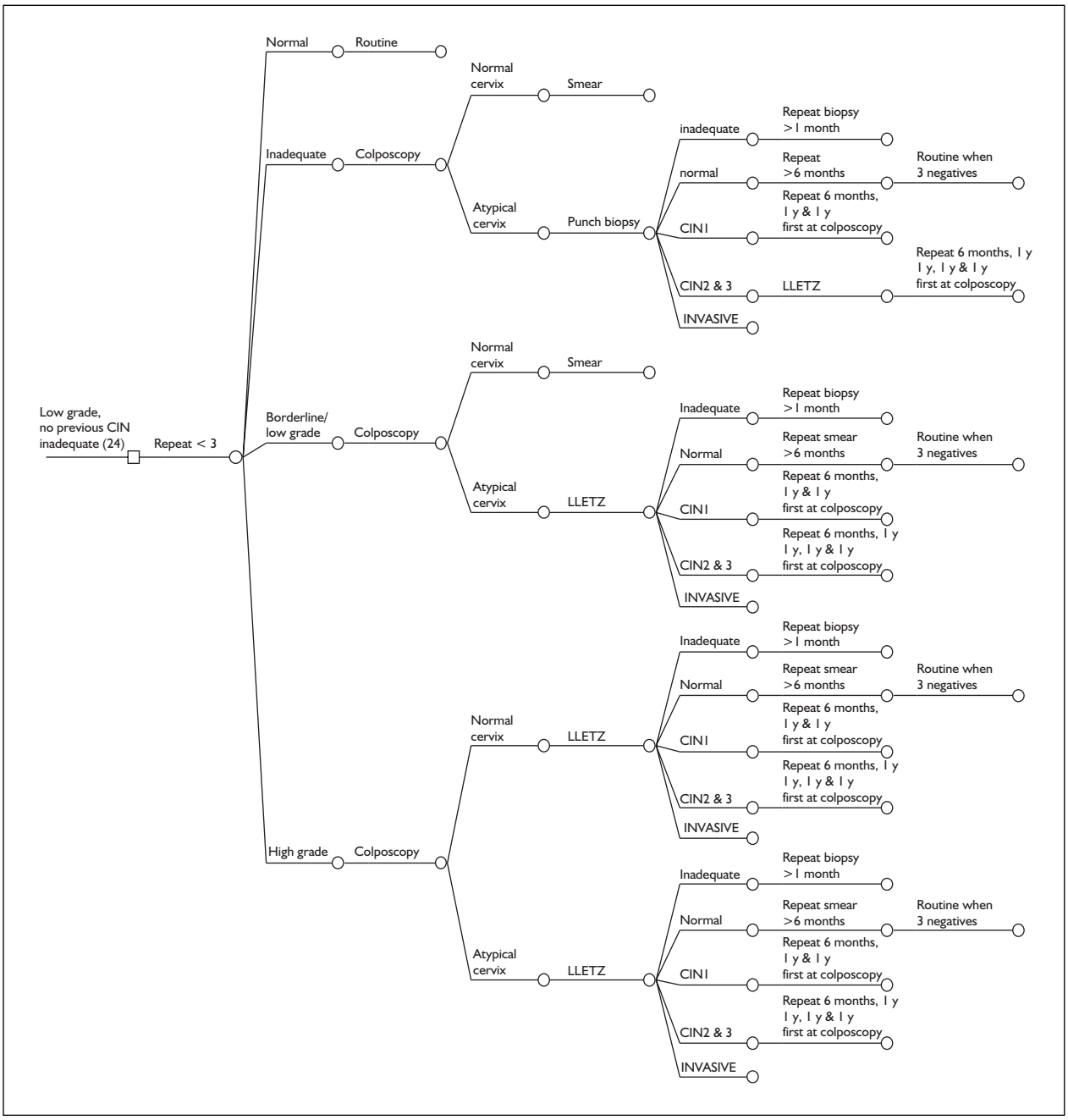

Borderline, borderline (17)
Repeat < 6
Normal
Routine
Inadequate
Colposcopy
Normal cervix
Smear
Atypical cervix
Punch biopsy
Inadequate
Repeat biopsy >1 month
Normal
Repeat >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
LLETZ
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
Borderline/ low grade
Colposcopy
Normal cervix
Smear
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
High grade
Colposcopy
Normal cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE

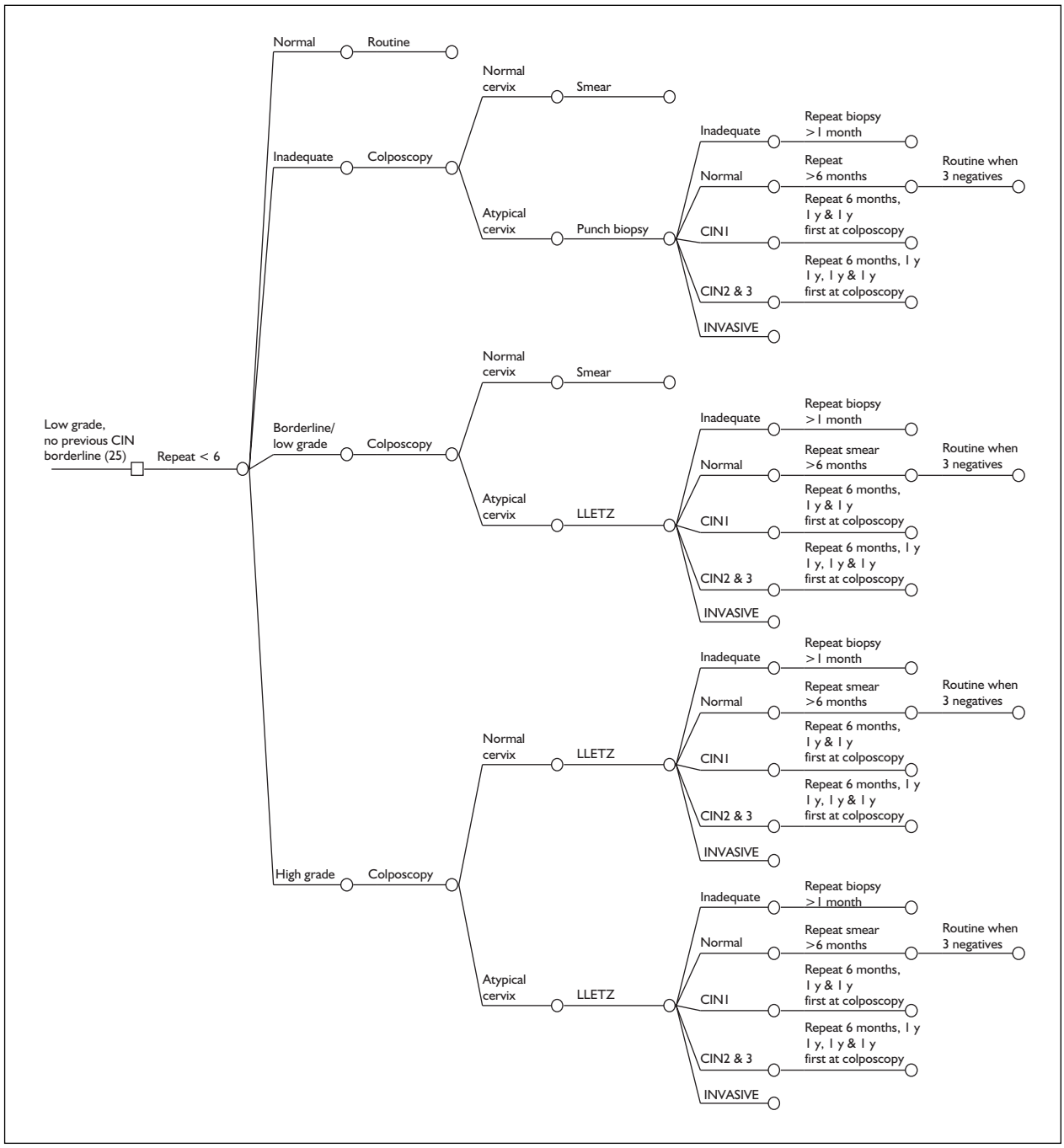

Borderline, low grade, previous treatment for CIN (18)
Colposcopy
Normal cervix
Repeat smear
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE

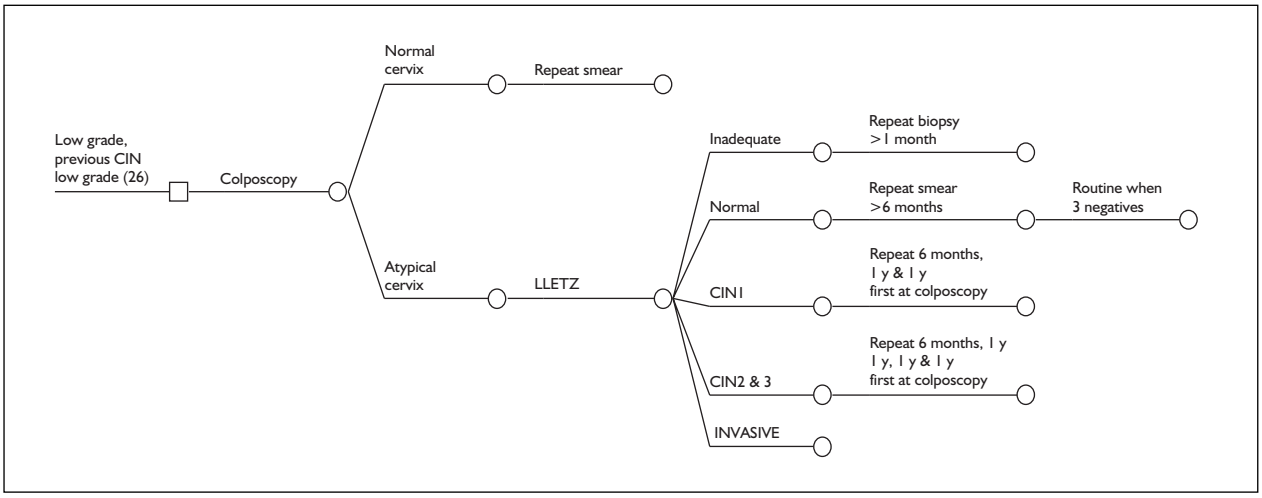


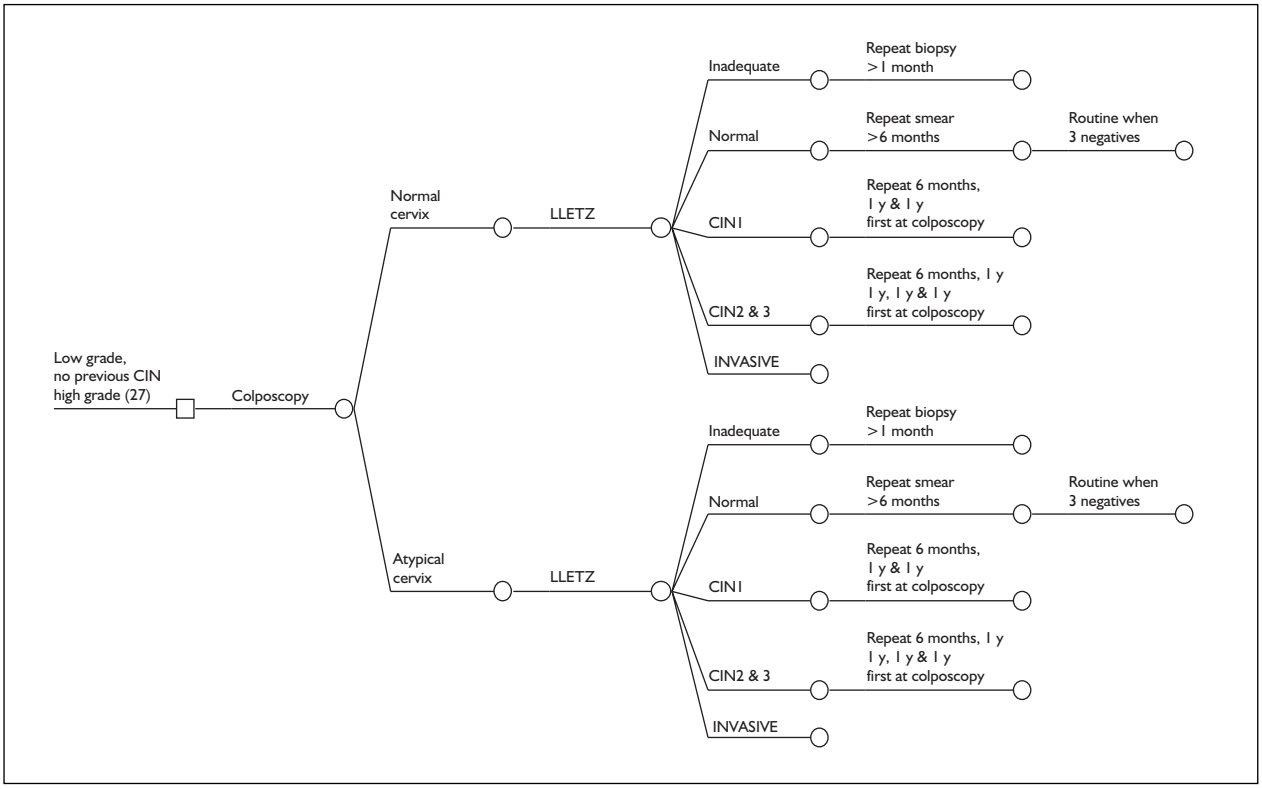

Borderline, high grade (20)
Colposcopy
Normal cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE


Low grade, previous treatment for CIN (21)
Colposcopy
Normal cervix
Repeat smear
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE




Low grade, no previous CIN borderline (25)
Repeat < 6
Normal
Routine
Inadequate
Colposcopy
Normal cervix
Smear
Atypical cervix
Punch biopsy
Inadequate
Repeat biopsy >1 month
Normal
Repeat >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
Borderline/ low grade
Colposcopy
Normal cervix
Smear
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
High grade
Colposcopy
Normal cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE
Atypical cervix
LLETZ
Inadequate
Repeat biopsy >1 month
Normal
Repeat smear >6 months
Routine when 3 negatives
CIN1
Repeat 6 months, 1 y & 1 y first at colposcopy
CIN2 & 3
Repeat 6 months, 1 y 1 y, 1 y & 1 y first at colposcopy
INVASIVE





Health Technology Assessment Programme

Prioritisation Strategy Group

Members

Chair,
Professor Tom Walley,
Director, NHS HTA Programme, Department of Pharmacology & Therapeutics, University of Liverpool

Professor Bruce Campbell, Consultant Vascular & General Surgeon, Royal Devon & Exeter Hospital

Professor Shah Ebrahim, Professor in Epidemiology of Ageing, University of Bristol

Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Radcliffe Hospital, Oxford

Dr Ron Zimmern, Director, Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

HTA Commissioning Board

Members

Programme Director,
Professor Tom Walley,
Director, NHS HTA Programme, Department of Pharmacology & Therapeutics, University of Liverpool

Chair,
Professor Shah Ebrahim,
Professor in Epidemiology of Ageing, Department of Social Medicine, University of Bristol

Deputy Chair,
Professor Jenny Hewison,
Professor of Health Care Psychology, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds School of Medicine

Dr Jeffrey Aronson
Reader in Clinical Pharmacology, Department of Clinical Pharmacology, Radcliffe Infirmary, Oxford

Professor Ann Bowling, Professor of Health Services Research, Primary Care and Population Studies, University College London

Professor Andrew Bradbury, Professor of Vascular Surgery, Department of Vascular Surgery, Birmingham Heartlands Hospital

Professor John Brazier, Director of Health Economics, Sheffield Health Economics Group, School of Health & Related Research, University of Sheffield

Dr Andrew Briggs, Public Health Career Scientist, Health Economics Research Centre, University of Oxford

Professor Nicky Cullum, Director of Centre for Evidence Based Nursing, Department of Health Sciences, University of York

Dr Andrew Farmer, Senior Lecturer in General Practice, Department of Primary Health Care, University of Oxford

Professor Fiona J Gilbert, Professor of Radiology, Department of Radiology, University of Aberdeen

Professor Adrian Grant, Director, Health Services Research Unit, University of Aberdeen

Professor F D Richard Hobbs, Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham

Professor Peter Jones, Head of Department, University Department of Psychiatry, University of Cambridge

Professor Sallie Lamb, Research Professor in Physiotherapy/Co-Director, Interdisciplinary Research Centre in Health, Coventry University

Professor Julian Little, Professor of Epidemiology, Department of Medicine and Therapeutics, University of Aberdeen

Professor Stuart Logan, Director of Health & Social Care Research, The Peninsula Medical School, Universities of Exeter & Plymouth

Professor Tim Peters, Professor of Primary Care Health Services Research, Division of Primary Health Care, University of Bristol

Professor Ian Roberts, Professor of Epidemiology & Public Health, Intervention Research Unit, London School of Hygiene and Tropical Medicine

Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh

Professor Mark Sculpher, Professor of Health Economics, Centre for Health Economics, Institute for Research in the Social Services, University of York

Professor Martin Severs, Professor in Elderly Health Care, Portsmouth Institute of Medicine

Dr Jonathan Shapiro, Senior Fellow, Health Services Management Centre, Birmingham

Ms Kate Thomas, Deputy Director, Medical Care Research Unit, University of Sheffield

Professor Simon G Thompson, Director, MRC Biostatistics Unit, Institute of Public Health, Cambridge

Ms Sue Ziebland, Senior Research Fellow, Cancer Research UK, University of Oxford

Current and past membership details of all HTA 'committees' are available from the HTA website (www.ncchta.org)

Diagnostic Technologies & Screening Panel

Members

Chair,
Dr Ron Zimmern, Director of the Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

Ms Norma Armston, Freelance Consumer Advocate, Bolton

Professor Max Bachmann Professor Health Care Interfaces, Department of Health Policy and Practice, University of East Anglia

Professor Rudy Bilous Professor of Clinical Medicine & Consultant Physician, The Academic Centre, South Tees Hospitals NHS Trust

Dr Paul Cockcroft, Consultant Medical Microbiologist/Laboratory Director, Public Health Laboratory, St Mary's Hospital, Portsmouth

Professor Adrian K Dixon, Professor of Radiology, Addenbrooke's Hospital, Cambridge

Dr David Elliman, Consultant in Community Child Health, London

Professor Glyn Elwyn, Primary Medical Care Research Group, Swansea Clinical School, University of Wales Swansea

Dr John Fielding, Consultant Radiologist, Radiology Department, Royal Shrewsbury Hospital

Dr Karen N Foster, Clinical Lecturer, Dept of General Practice & Primary Care, University of Aberdeen

Professor Antony J Franks, Deputy Medical Director, The Leeds Teaching Hospitals NHS Trust

Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London

Dr Edmund Jessop, Medical Adviser, National Specialist Commissioning Advisory Group (NSCAG), Department of Health, London

Dr Jennifer J Kurinczuk, Consultant Clinical Epidemiologist, National Perinatal Epidemiology Unit, Oxford

Dr Susanne M Ludgate, Medical Director, Medical Devices Agency, London

Dr William Rosenberg, Senior Lecturer and Consultant in Medicine, University of Southampton

Dr Susan Schonfield, CPHM Specialised Services Commissioning, Croydon Primary Care Trust

Dr Margaret Somerville, Director of Public Health, Teignbridge Primary Care Trust

Professor Lindsay Wilson Turnbull, Scientific Director, Centre for MR Investigations & YCR Professor of Radiology, University of Hull

Professor Martin J Whittle, Head of Division of Reproductive & Child Health, University of Birmingham

Dr Dennis Wright, Consultant Biochemist & Clinical Director, Pathology & The Kennedy Galton Centre, Northwick Park & St Mark's Hospitals, Harrow

Pharmaceuticals Panel

Members

Chair,
Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Oxford Radcliffe Hospital

Professor Tony Avery, Professor of Primary Health Care, University of Nottingham

Professor Stirling Bryan, Professor of Health Economics, Health Services Management Centre, University of Birmingham

Mr Peter Cardy, Chief Executive, Macmillan Cancer Relief, London

Dr Christopher Cates, GP and Cochrane Editor, Bushey Health Centre

Professor Imti Choonara, Professor in Child Health, University of Nottingham, Derbyshire Children's Hospital

Mr Charles Dobson, Special Projects Adviser, Department of Health

Dr Robin Ferner, Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Dr Karen A Fitzgerald, Pharmaceutical Adviser, Bro Taf Health Authority, Cardiff

Mrs Sharon Hart, Managing Editor, *Drug & Therapeutics Bulletin*, London

Dr Christine Hine, Consultant in Public Health Medicine, Bristol South & West Primary Care Trust

Professor Stan Kaye, Professor of Medical Oncology, Consultant in Medical Oncology/Drug Development, The Royal Marsden Hospital

Ms Barbara Meredith, Project Manager Clinical Guidelines, Patient Involvement Unit, NICE

Dr Frances Rotblat, CPMP Delegate, Medicines Control Agency, London

Professor Jan Scott, Professor of Psychological Treatments, Institute of Psychiatry, University of London

Mrs Katrina Simister, New Products Manager, National Prescribing Centre, Liverpool

Dr Richard Tiner, Medical Director, Association of the British Pharmaceutical Industry

Dr Helen Williams, Consultant Microbiologist, Norfolk & Norwich University Hospital NHS Trust

Therapeutic Procedures Panel

Members

Chair,
Professor Bruce Campbell,
Consultant Vascular and General Surgeon, Royal Devon & Exeter Hospital

Dr Mahmood Adil, Head of Clinical Support & Health Protection, Directorate of Health and Social Care (North), Department of Health, Manchester

Dr Aileen Clarke, Reader in Health Services Research, Public Health & Policy Research Unit, Barts & the London School of Medicine & Dentistry, Institute of Community Health Sciences, Queen Mary, University of London

Mr Matthew William Cooke, Senior Clinical Lecturer and Honorary Consultant, Emergency Department, University of Warwick, Coventry & Warwickshire NHS Trust, Division of Health in the Community, Centre for Primary Health Care Studies, Coventry

Dr Carl E Counsell, Senior Lecturer in Neurology, University of Aberdeen

Dr Keith Dodd, Consultant Paediatrician, Derbyshire Children's Hospital

Professor Gene Feder, Professor of Primary Care R&D, Barts & the London, Queen Mary's School of Medicine and Dentistry, University of London

Professor Paul Gregg, Professor of Orthopaedic Surgical Science, Department of Orthopaedic Surgery, South Tees Hospital NHS Trust

Ms Bec Hanley, Freelance Consumer Advocate, Hurstpierpoint

Ms Maryann L. Hardy, Lecturer, Division of Radiography, University of Bradford

Professor Alan Horwich, Director of Clinical R&D, The Institute of Cancer Research, London

Dr Phillip Leech, Principal Medical Officer for Primary Care, Department of Health, London

Dr Simon de Lusignan, Senior Lecturer, Primary Care Informatics, Department of Community Health Sciences, St George's Hospital Medical School, London

Dr Mike McGovern, Senior Medical Officer, Heart Team, Department of Health, London

Professor James Neilson, Professor of Obstetrics and Gynaecology, Dept of Obstetrics and Gynaecology, University of Liverpool, Liverpool Women's Hospital

Dr John C Pounsford, Consultant Physician, North Bristol NHS Trust

Dr Vimal Sharma, Consultant Psychiatrist & Hon Snr Lecturer, Mental Health Resource Centre, Victoria Central Hospital, Wirrall

Dr L David Smith, Consultant Cardiologist, Royal Devon & Exeter Hospital

Professor Norman Waugh, Professor of Public Health, University of Aberdeen

Expert Advisory Network

Members

Professor Douglas Altman,
Director of CSM & Cancer Research UK Med Stat Gp, Centre for Statistics in Medicine, University of Oxford, Institute of Health Sciences, Headington, Oxford

Professor John Bond,
Director, Centre for Health Services Research, University of Newcastle upon Tyne, School of Population & Health Sciences, Newcastle upon Tyne

Mr Shaun Brogan,
Chief Executive, Ridgeway Primary Care Group, Aylesbury

Mrs Stella Burnside OBE,
Chief Executive, Office of the Chief Executive. Trust Headquarters, Altnagelvin Hospitals Health & Social Services Trust, Altnagelvin Area Hospital, Londonderry

Ms Tracy Bury,
Project Manager, World Confederation for Physical Therapy, London

Mr John A Cairns,
Professor of Health Economics, Health Economics Research Unit, University of Aberdeen

Professor Iain T Cameron,
Professor of Obstetrics and Gynaecology and Head of the School of Medicine, University of Southampton

Dr Christine Clark,
Medical Writer & Consultant Pharmacist, Rossendale

Professor Collette Mary Clifford,
Professor of Nursing & Head of Research, School of Health Sciences, University of Birmingham, Edgbaston, Birmingham

Professor Barry Cookson,
Director, Laboratory of Healthcare Associated Infection, Health Protection Agency, London

Professor Howard Stephen Cuckle,
Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Professor Nicky Cullum,
Director of Centre for Evidence Based Nursing, University of York

Dr Katherine Darton,
Information Unit, MIND – The Mental Health Charity, London

Professor Carol Dezateux,
Professor of Paediatric Epidemiology, London

Mr John Dunning,
Consultant Cardiothoracic Surgeon, Cardiothoracic Surgical Unit, Papworth Hospital NHS Trust, Cambridge

Mr Jonothan Earnshaw,
Consultant Vascular Surgeon, Gloucestershire Royal Hospital, Gloucester

Professor Martin Eccles,
Professor of Clinical Effectiveness, Centre for Health Services Research, University of Newcastle upon Tyne

Professor Pam Enderby,
Professor of Community Rehabilitation, Institute of General Practice and Primary Care, University of Sheffield

Mr Leonard R Fenwick,
Chief Executive, Newcastle upon Tyne Hospitals NHS Trust

Professor David Field,
Professor of Neonatal Medicine, Child Health, The Leicester Royal Infirmary NHS Trust

Mrs Gillian Fletcher,
Antenatal Teacher & Tutor and President, National Childbirth Trust, Henfield

Professor Jayne Franklyn,
Professor of Medicine, Department of Medicine, University of Birmingham, Queen Elizabeth Hospital, Edgbaston, Birmingham

Ms Grace Gibbs,
Deputy Chief Executive, Director for Nursing, Midwifery & Clinical Support Servs, West Middlesex University Hospital, Isleworth

Dr Neville Goodman,
Consultant Anaesthetist, Southmead Hospital, Bristol

Professor Alastair Gray,
Professor of Health Economics, Department of Public Health, University of Oxford

Professor Robert E Hawkins,
CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor F D Richard Hobbs,
Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham

Professor Allen Hutchinson,
Director of Public Health & Deputy Dean of ScHARR, Department of Public Health, University of Sheffield

Dr Duncan Keeley,
General Practitioner (Dr Burch & Ptnrs), The Health Centre, Thame

Dr Donna Lamping,
Research Degrees Programme Director & Reader in Psychology, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London

Mr George Levvy,
Chief Executive, Motor Neurone Disease Association, Northampton

Professor James Lindesay,
Professor of Psychiatry for the Elderly, University of Leicester, Leicester General Hospital

Professor Rajan Madhok,
Medical Director & Director of Public Health, Directorate of Clinical Strategy & Public Health, North & East Yorkshire & Northern Lincolnshire Health Authority, York

Professor David Mant,
Professor of General Practice, Department of Primary Care, University of Oxford

Professor Alexander Markham,
Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Dr Chris McCall,
General Practitioner, The Hadleigh Practice, Castle Mullen

Professor Alistair McGuire,
Professor of Health Economics, London School of Economics

Dr Peter Moore,
Freelance Science Writer, Ashtead

Dr Andrew Mortimore,
Consultant in Public Health Medicine, Southampton City Primary Care Trust

Dr Sue Moss,
Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton

Professor Jon Nicholl,
Director of Medical Care Research Unit, School of Health and Related Research, University of Sheffield

Mrs Julietta Patnick,
National Co-ordinator, NHS Cancer Screening Programmes, Sheffield

Professor Robert Peveler,
Professor of Liaison Psychiatry, University Mental Health Group, Royal South Hants Hospital, Southampton

Professor Chris Price,
Visiting Chair – Oxford, Clinical Research, Bayer Diagnostics Europe, Cirencester

Ms Marianne Rigge,
Director, College of Health, London

Dr Eamonn Sheridan,
Consultant in Clinical Genetics, Genetics Department, St James's University Hospital, Leeds

Dr Ken Stein,
Senior Clinical Lecturer in Public Health, Director, Peninsula Technology Assessment Group, University of Exeter

Professor Sarah Stewart-Brown,
Director HSRU/Honorary Consultant in PH Medicine, Department of Public Health, University of Oxford

Professor Ala Szczepura,
Professor of Health Service Research, Centre for Health Services Studies, University of Warwick

Dr Ross Taylor,
Senior Lecturer, Department of General Practice and Primary Care, University of Aberdeen

Mrs Joan Webster,
Consumer member, HTA – Expert Advisory Network

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278