

D-10-01014R2

S1470-2045(10)70264-3

Automation assisted versus manual reading of cervical cytology (MAVARIC): a randomised controlled trial

Henry C Kitchener, Roger Blanks, Graham Dunn, Lionel Gunn, Mina Desai, Rebecca Albrow, Jean Mather, Durgesh N Rana, Heather Cubie, Catherine Moore, Rosa Legood, Alastair Gray, Sue Moss

Summary

Background The standard for reading cervical cytology is for a cytoscreener to manually search across an entire slide for abnormal cells using a conventional microscope. Automated technology can select fields of view to assess abnormal cells, which allows targeted reading by cytoscreeners. In the Manual Assessment Versus Automated Reading In Cytology (MAVARIC) trial, we compared the accuracy of these techniques for the detection of underlying disease.

Methods For this randomised controlled trial, women aged 25–64 years undergoing primary cervical screening in Manchester, UK, were randomly assigned (1:2) to receive either manual reading only or paired reading (automation-assisted reading and manual reading), between March 1, 2006, and Feb 28, 2009. In the paired arm, two automated systems were used—the ThinPrep Imaging System and the FocalPoint GS Imaging System. General practices and community clinics were randomised to either ThinPrep or to SurePath (for the FocalPoint system) liquid-based cytology with block randomisation stratified by deprivation index. Samples were then individually randomised to manual reading only or paired reading only. Laboratory staff were unaware of the allocation of each slide and concealment was maintained until the end of the reporting process. The primary outcome was sensitivity of automation-assisted reading relative to manual reading for the detection of underlying cervical intraepithelial neoplasia grade 2 or worse (CIN2+) in the paired arm. This trial is registered, number ISRCTN 66377374.

Findings 73 266 liquid-based cytology samples were obtained from women undergoing primary cervical screening; 24 688 allocated to the manual-only arm and 48 578 to the paired-reading arm. Automation-assisted reading was 8% less sensitive than manual reading (relative sensitivity 0.92, 95% CI 0.89–0.95), which was equivalent to an absolute reduction in sensitivity of 6.3%, assuming the sensitivity of manual reading to be 79%. Specificity of automation-assisted reading relative to manual reading increased by 0.6% (1.006, 95% CI 1.005–1.007).

Interpretation The inferior sensitivity of automation-assisted reading for the detection of CIN2+, combined with an inconsequential increase in specificity, suggests that automation-assisted reading cannot be recommended for primary cervical screening.

Funding National Institute for Health Research Health Technology Assessment programme

Introduction

Since the introduction of cervical cytology, cervical screening has relied on cytoscreeners to manually scan entire slides, on which exfoliated cervical cells are fixed and stained, with conventional microscopy. Identification of an abnormality prompts either immediate referral to colposcopy for identification of high-grade changes or repeat cytology for low-grade changes. Triage with human papillomavirus (HPV) testing^{1–3} is being used instead of repeat cytology. Traditionally, exfoliated cells were spread onto the slide before fixation, but liquid-based cytology,^{4,5} which can achieve cleaner preparations and provide a medium for reflex HPV testing, has become more common, and since 2003 has been rolled out across the UK's national cervical screening programmes. Because of its added sensitivity, HPV testing as a primary screen is being considered in several countries.^{6–8}

During the past 20 years, technology has been developed that automates the process of reading cytology by identifying abnormal cells and presenting a restricted

number of fields of view to the cytoscreener on a computer screen. This technology has the potential to reduce screening errors and increase productivity by reducing the time needed to read slides. Two commercially available automated screening systems are available that have been approved for primary cervical screening by the US Food and Drug Administration (FDA)—the BD FocalPoint GS Imaging System (BD Diagnostics, Franklin Lakes, NJ, USA) and the ThinPrep Imaging System (Hologic, Bedford, MA, USA). The FocalPoint system uses SurePath liquid-based cytology and the ThinPrep imager uses ThinPrep liquid-based cytology. FDA approval has been based on such automated systems being capable of detecting an equivalent or higher proportion of high-grade cytological abnormalities than does manual reading.^{9,10}

Findings from a systematic review of the effectiveness of automated and semi-automated cervical screening devices¹¹ showed that the evidence base was not strong enough for reliable conclusions to be made; it recommended further trials with robust reference

Lancet Oncol 2010; 11: ???–???

School of Cancer and Enabling Sciences (Prof H C Kitchener MD, R Albrow MPH), Health Sciences Research Group, School of Community Based Medicine (Prof G Dunn PhD), The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK; Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton, UK (R Blanks PhD, L Gunn BSc, S Moss PhD); Manchester Cytology Centre, Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK (M Desai FRCPath, J Mather MIBMS, D N Rana FRCPath); Specialist Virology Centre, Royal Infirmary of Edinburgh, Edinburgh, UK (Prof H Cubie PhD, C Moore BSc); Health Services Research Unit, London School of Hygiene and Tropical Medicine, London, UK (R Legood PhD); and Health Economics Research Centre, University of Oxford, Oxford, UK (Prof A Gray PhD)

Correspondence to: Prof Henry C Kitchener, School of Cancer and Enabling Sciences, The University of Manchester, Manchester Academic Health Science Centre, Research Floor (5th Floor), St Mary's Hospital, Oxford Road, Manchester, M13 9WL
henry.c.kitchener@manchester.ac.uk

standards. Several studies have had methodological weaknesses, including outdated technology, use of split samples to produce two slides, use of cytological outcomes as opposed to histopathological outcomes, and retrospective readings of the same slide set.

Since the launch of the Manual Assessment Versus Automated Reading In Cytology (MAVARIC) study, a systematic review by the UK Health Technology Assessment programme¹² also concluded that previous studies had not been of sufficient quality to allow reliable recommendations. The MAVARIC trial was designed to achieve a rigorous, prospective, unbiased comparison of manual and automation-assisted reading that had been powered to show non-inferiority in terms of sensitivity to detect cervical intraepithelial neoplasia grade 2 or worse (CIN2+), which worldwide represents the threshold for proceeding to treatment. Other objectives were to compare specificity of automation-assisted screening with manual screening, to incorporate both automated systems, and to assess the reliability of slides defined by the FocalPoint system as needing no further review for exclusion of CIN2+. A robust economic analysis to determine cost-effectiveness has also been done and will be reported elsewhere.¹³

Methods

Participants

This randomised controlled trial was designed to incorporate procedures used in routine national cervical screening practice in the UK to study detection of high-grade disease in liquid-based cytology samples. Between March 1, 2006, and Feb 28, 2009, 174 general practices and community clinics in Greater Manchester took part in the study. Coverage by the screening programme in Manchester is similar to that for England as a whole (78%). Cervical samples from women aged 25–64 years were obtained during routine primary cervical screening as part of the national screening programme. Samples were randomly allocated to either manual reading alone or automated reading paired with manual reading. In the paired arm, automated reading of ThinPrep samples was done with the ThinPrep Imaging System and SurePath samples with the FocalPoint GS Imaging System.

Initially, written informed consent was needed because HPV triage was not standard practice, but during the study HPV triage became UK National Health Service (NHS) Cervical Screening Programme standard protocol in Greater Manchester and signed consent was then not needed. The study received ethical approval from the central Manchester local research ethics committee (04/Q1407/318).

Randomisation and masking

There were two randomisation stages in the study. Cervical samples were initially randomised between the two imaging technologies when obtained at the general practice stage. This initial randomisation was to ensure

that the two systems had a similar population of women at similar risk. Randomisation was therefore stratified by primary care trust to take account of variation in Townsend deprivation score¹⁴ and in ethnic minority composition. Sources within each trust were assumed to have close levels of these risk indicators. Community clinic sources were included as a separate stratum. Sources were ordered by decreasing size (number of women anticipated to participate based on previous numbers of samples sent to the laboratory) within each stratum because of variation in population size. A sequence of random digits was used to allocate blocks of four general practice sources to the six possible combinations of the two technologies to ensure that similar populations from each primary care trust were allocated to each technology. [A: rewording ok?]

The second stage of randomisation took place between the manual only and paired arms; samples received in the laboratory were individually randomly allocated, with the use of preprepared spreadsheets produced by the Cancer Screening Evaluation Unit (Institute of Cancer Research, Sutton, UK). Each liquid-based cytology system had a separate randomisation spreadsheet with unique numbers to randomise each sample to either manual reading only or manual reading paired with automation-assisted reading. Initially randomisation was 1 to 1, but after a third of the samples had been obtained, at a slower than expected rate of accrual, the ratio was changed to 1 to 3 in favour of the paired arm to accelerate the accrual of samples for paired reading. The objective was to achieve a final ratio of 1 to 2, which would preserve the planned size and power of the paired arm (50 000 samples).

Laboratory routines were developed to ensure that staff were masked to both the study arm and the automated result at the time of manual reading. Slides for automated screening were screened with the review scopes; no marks were made on the slides to show any abnormal cells and results were entered into the randomisation list. The list was removed and passed to the laboratory coordinator before the slides were put back into routine screening by numerical order, thus ensuring the manual screener was masked to the result of the automated read.

Procedures

The same stain was used for both manual reading and automation-assisted reading, so in the case of ThinPrep we therefore needed to do manual reading with the Imager stain. Before the study began, manual reading with the Imager stain passed the northwest regional technical external quality assurance assessment. Both automated systems use similar screening methods, which were followed throughout the trial; the ThinPrep Imaging System presented 22 fields of view to the cytoscreeners and the FocalPoint selected ten. Fields of view were selected by the machines as areas of the slide most likely to contain an abnormality. If any abnormalities

For full protocol see
www.hta.ac.uk/1462

were detected in these fields of view, the slide received a full screen on the review scope, as per the manufacturer's protocol. The FocalPoint system also has the ability to rank slides into quintiles according to the likelihood of the slide containing an abnormality, with quintile one consisting of slides with the most abnormalities. In addition to the production of quintiles the machine can also designate up to 25% of slides in each run as needing no further review (ie, they can be archived without further interpretation). Slides designated as needing no further review during the trial had this result recorded. After automated screening, the list and slides were passed to another screener for rapid review.

Manual screening (in both arms) was done according to routine laboratory protocols, including the practice of marking areas of interest on the slide. In the paired arm, automated reading was undertaken first, followed by the manual read. After an initial read by the cytoscreener, whether automated or manual, a manual rapid review was done for every sample according to standard practice for UK screening programmes. If a slide was positive it was re-read by a skilled practitioner or cytopathologist, which determined either a final manual or automation-assisted result. In the paired arm, management was based on which result was worse in terms of abnormality.

Final reading of each slide was done only once by a doctor if it was abnormal and as a result there are no discrepancies between abnormalities for final readings. Discordant final readings arose from inadequate or negative results paired with an abnormal result. High-grade cytological abnormality prompted referral to colposcopy, and low-grade abnormalities were triaged by Hybrid Capture 2 high-risk HPV testing (Qiagen, Crawley, UK), with HPV-positive cases referred to colposcopy. Women with negative cytology and those with HPV-negative low-grade abnormalities were returned to routine recall. These low-grade abnormalities included both borderline (atypical squamous cells of undetermined significance) and mild dyskaryosis (low-grade squamous intraepithelial lesion); even though the mild dyskaryoses have a high prevalence of HPV, a negative triage allows a few women to avoid colposcopy referral. The reference standard was histopathology obtained at colposcopy from either a colposcopically directed punch biopsy or loop excision. When both procedures had been done the most severe grade was used. Abnormalities were examined by specialist gynaecological pathologists who were blinded to the arm of the study.

The trial was designed to study detection of high-grade disease in liquid-based cytology samples. The primary outcome was sensitivity of automation-assisted reading relative to manual reading for the detection of underlying CIN2+. Secondary outcomes were relative specificity, the same outcomes for cervical intraepithelial neoplasia grade 3 or worse (CIN3+), sensitivity of the two

automated technologies relative to manual screening and to each other, and the reliability of slides defined as needing no further review to exclude underlying CIN2+.

Statistical analyses

Absolute sensitivity of manual reading or automated reading could not be calculated because the number of cases of CIN2+ in samples negative according to both methods was unknown. However, an estimation of the ratio of the two sensitivities (the missing count cancels out) and an assessment of confidence intervals and statistical significance was possible. The ratio is the number of samples of CIN2+ that were screen positive with automated screening divided by the number of samples of CIN2+ that tested positive with manual screening. Similarly, relative specificity was calculated roughly as the ratio of the number of samples of cervical intraepithelial neoplasia of grade 1 or less that were negative on automated reading, to the number that were negative on manual reading, on the assumption that the

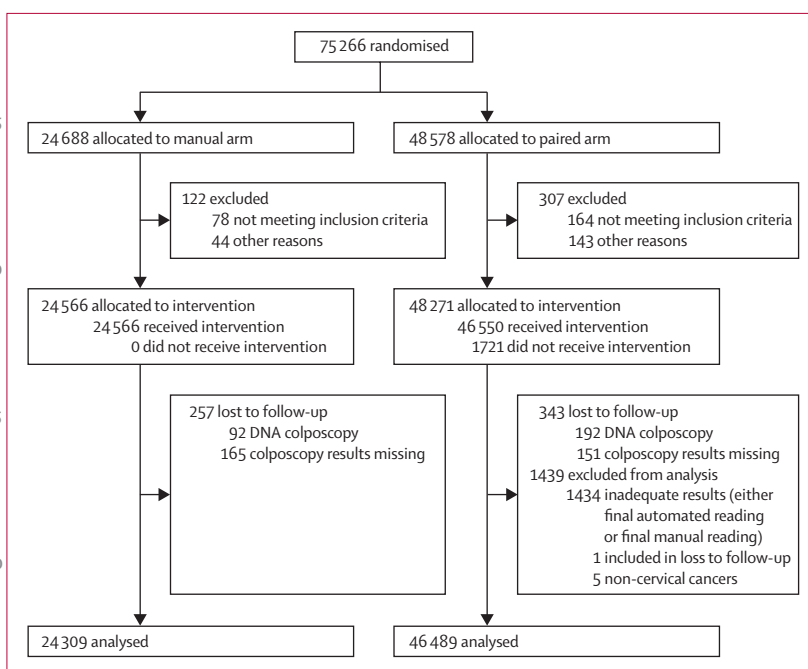


Figure 1: Study profile

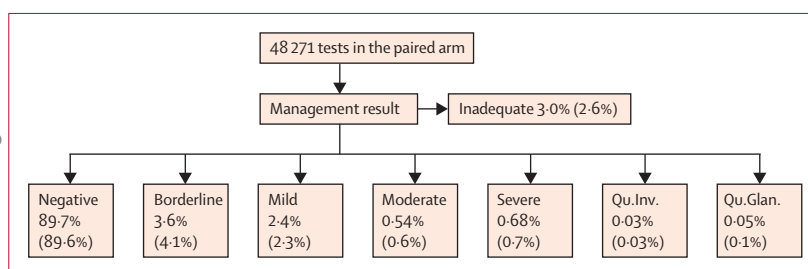


Figure 2: Management results of samples in the paired arm

Results in parentheses are related corresponding national figures.¹⁴ Qu. Inv.=query invasive neoplasia. Qu. Glan.=query glandular neoplasia.

number of samples of CIN2+ not detected by either 1 screening method is zero.

Sample size was based on a test of non-inferiority of 2 the automated technology in terms of its sensitivity relative to manual reading, using data from the paired 5

	Manual arm		Paired arm	
	BD SurePath (n=12 195)	ThinPrep (n=12 371)	BD SurePath (n=23 404)	ThinPrep (n=24 867)
CIN1	117 (9.59)	94 (7.60)	195 (8.33)	159 (6.39)
CIN2	93 (7.63)	87 (7.03)	159 (6.79)	144 (5.79)
CIN3	109 (8.94)	81 (6.55)	181 (7.73)	182 (7.32)
CGIN	7 (0.57)	5 (0.40)	15 (0.64)	7 (0.28)
Cancer*	10 (0.82)	6 (0.49)	11† (0.47)	8† (0.32)
CIN2+	219 (2%)	179 (1%)	366 (2%)	341 (1%)
CIN3+	126 (1%)	92 (1%)	207 (1%)	197 (1%)

Data are number (rate per 1000) or n (%). CIN=cervical intraepithelial neoplasia. CGIN=cervical glandular intraepithelial neoplasia. CIN2+=cervical intraepithelial neoplasia 2 or worse. CIN3+=cervical intraepithelial neoplasia 3 or worse. *30 samples stage 1a, 5 stage 1b. †14 were manual positive and automated positive, two were automated positive only, two manual positive only, and one was manual negative and automated negative.

Table 1: Histology results for all ages

observations. Non-inferiority was based on an absolute 3 difference of not more than 5%. With 630 CIN2+ lesions in the paired arm and a paired test with a 0.025 one-sided significance, there was 80% power to reject the null hypothesis that the sensitivities are not equivalent (ie, 0.05 or further from zero, on the assumption of an estimated upper limit in the proportion of discordant pairs of 20%). With an estimated 3% of the screened population showing CIN2+ 46 000 paired readings would be needed, so an accrual target of 50 000 was set in the paired arm. To achieve these targets, we originally planned to accrue 100 000 samples, 50 000 in the manual only and 50 000 in the paired arm. Because accrual was initially slower than anticipated, the accrual target was 15 reduced after 23 months to 75 000 and the randomisation ratio changed to 1 to 3, which would achieve the 50 000 target in the paired arm and 25 000 in the manual arm. Such targets would maintain the study's power and provide a sufficiently large manual arm, the main 20 purpose of which was to mask cytoscreeners to the arm in which samples for manual reading belonged. This change was approved by the independent data monitoring and ethics committee and the independent trial steering committee. An absolute difference of 5% for the

	Final automated reading						Total
	Inadequate	Negative	HPV positive	HPV negative	HPV result not known	Modified [A: correct?]	
Final manual reading (both systems)							
Inadequate	810	556	1366
Negative	69	43 284	125	101	56	12	43 647
HPV positive	..	317	900	1217
HPV negative	..	350	..	334	684
HPV result not known	..	217	523	..	740
Modified [A: correct?]	..	47	570	617
Total	879	44 771	1025	435	579	582	48 271
Final manual reading (ThinPrep)							
Inadequate	456	284	740
Negative	26	22 237	91	75	34	8	22 471
HPV positive	..	165	503	668
HPV negative	..	189	..	191	380
HPV result not known	..	88	228	..	316
Modified [A: correct?]	..	17	275	292
Total	482	22 980	594	266	262	283	24 867
Final manual reading (FocalPoint)							
Inadequate	354	272	626
Negative	43	21 047	34	26	22	4	21 176
HPV positive	..	152	397	549
HPV negative	..	161	..	143	304
HPV result not known	..	129	295	..	424
Modified [A: correct?]	..	30	295	325
Total	397	21 791	431	169	317	299	23 404

Borderline or mild abnormalities include human papillomavirus (HPV) positive, HPV negative, and HPV result not known. [A: additional wording correct?] At final reading each slide was seen only once by a doctor if it was abnormal and as a result there are no discrepancies between abnormalities. HPV=human papillomavirus.

Table 2: Final automated read versus final manual read of borderline or mild abnormalities

definition of non-inferiority would need a relative difference of at least 6.5% on the assumption of 79% sensitivity of liquid-based cytology to detect CIN2+.⁴ All analyses were done with Stata version 10.1. This study is registered as an International Standard Randomised Controlled Trial, number ISRCTN66377374.

Role of the funding source

The National Institute for Health Research Health Technology Assessment programme reviewed and approved the study design. Data collection, analysis, interpretation, and write-up were done independently by the authors, all of whom had access to the raw data. HCK had full access to all of the data and had final responsibility to submit for publication.

Results

Figure 1 shows the trial profile. 35812 specimens were assessed with BD SurePath and 37454 with ThinPrep. 429 samples were excluded, most of which were either vault cytology errors or processing errors. 124 (71%) of 174 randomised general practices provided samples as randomised, 22 provided samples using the alternative technology, and 28 did not contribute at all. Some non-randomised practices were included to increase the number of samples and to raise the proportion of high-grade cytology; a small proportion of samples were included from women attending colposcopy clinics. Mean age of women in the study was 39 years in the manual and paired arms. Mean Townsend deprivation score was very similar between arms and liquid-based cytology systems—manual arm, SurePath (3.84) and ThinPrep (3.99), and paired arm, SurePath (3.64) and ThinPrep (3.85).

Fewer samples were obtained from women aged 45–64 years (21231; 29.1%) than were obtained from women aged 25–44 years (47987; 65.7%) because women aged 50–64 years are invited every 5 years for screening, whereas women aged 25–49 years are invited every 3 years. In real life, some women outside these age ranges are screened, and exclusion of these samples was felt to be inappropriate. There were 3619 (5.0%) slides from women outside the screening age range; 3103 from women aged less than 25 years and 606 from women aged 65 years or older.

MAVARIC screening results are representative of those across the national programme (figure 2).¹⁵ When results of manual readings from manual and paired arms were compared, they were almost identical in terms of negative rates; 90.04% (manual) and 90.41% (paired). Corresponding rates for borderline or mild dyskaryosis were 6.01% (manual) and 5.47% (paired) and for moderate dyskaryosis or worse, 1.36% (manual) and 1.27% (paired). Comparison of these results before and after the change in randomisation from 1 to 1 to 1 to 3 showed that the proportions of abnormality in all adequate samples in each arm was very similar during

each period (8.32% vs 8.31% before change, and 6.71% vs 6.44% after change). The fall in proportions was due to cessation of accrual of samples from colposcopy clinics. There was no evidence that the less than perfect compliance with the randomisation of allocated technology or with the change in randomisation ratio affected the study. Table 1 shows the histopathology results, which indicate the similarity between the manual and paired arms for each technology.

		Final automated reading			
		CIN2+, positive	CIN2+, negative	CIN3+, positive	CIN3+, negative
Final manual reading					
Positive		577 (362)*	83 (59)	340 (225)†	39 (28)
Negative		31 (22)	16 (2)	21 (13)	4 (0)

Numbers in parentheses show data restricted to ages 25–64 years and routine screening samples only. CIN2+=cervical intraepithelial neoplasia grade 2 or worse. CIN3+=cervical intraepithelial neoplasia grade 3 or worse. *Relative sensitivity, based on matched pairs 0.92; 95% CI 0.89–0.95 (0.91; 0.87–0.95). †Relative sensitivity based on matched pairs 0.95; 95% CI 0.91–0.99 (0.94; 0.89–0.99).

Table 3: Relative sensitivity for final automated and final manual readings in the paired arm

		Final automated reading			
		CIN1-, positive	CIN1-, negative	CIN2-, positive	CIN2-, negative
Final manual reading					
Positive		1120 (665)*	358 (225)	1357 (802)†	402 (256)
Negative		98 (73)	44 206 (35763)	108 (82)	44 218 (35765)

Numbers in parentheses show data restricted to age 25–64 years and routine samples only. CIN1-=cervical intraepithelial neoplasia grade 1 or less. CIN2-=cervical intraepithelial neoplasia grade 2 or less. *Relative sensitivity based on matched pairs 1.006; 95% CI 1.005–1.007 (1.004; 1.003–1.005). †Relative sensitivity based on matched pairs 1.007; 95% CI 1.006–1.008 (1.005; 1.004–1.006).

Table 4: Relative specificity for final automated and final manual readings in the paired arm

		Final automated reading							
		Final automated read BD FocalPoint				Final automated read ThinPrep Imaging System			
		CIN2+, positive	CIN2+, negative	CIN3+, positive	CIN3+, negative	CIN2+, positive	CIN2+, negative	CIN3+, positive	CIN3+, negative
Final manual read									
Positive		176*	31	106 †	18	186‡	28	119§	10
Negative		11	1	7	0	11	1	6	0

Age restricted to 25–64 years and routine samples only are included to avoid bias, because there were more colposcopy samples for the SurePath system than the ThinPrep system. CIN2+=cervical intraepithelial neoplasia grade 2 or worse. CIN3+=cervical intraepithelial neoplasia grade 3 or worse. *Relative sensitivity based on matched pairs 0.90 (95% CI 0.85–0.96). †Relative sensitivity based on matched pairs 0.91 (95% CI 0.84–0.99). ‡Relative sensitivity based on matched pairs 0.92 (95% CI 0.87–0.98). §Relative sensitivity based on matched pairs 0.97 (95% CI 0.91–1.03).

Table 5: Relative sensitivity for final automated readings and final manual readings for the BD FocalPoint GS Imaging System and the ThinPrep Imaging System

	No further review	Quintile 5	Quintile 4	Quintile 3	Quintile 2	Quintile 1	Total
Not referred	4452 (3714)	3381 (2869)	3190 (2686)	3118 (2634)	3016 (2584)	2439 (2041)	19 596 (16 528)
Total referred	117 (26)	93 (41)	113 (59)	152 (83)	222 (124)	589 (370)	1286 (703)
CIN2+	10 (2)	18 (9)	18 (12)	29 (20)	41 (27)	205 (152)	321 (222)
CIN3+	4 (1)	10 (6)	10 (6)	16 (10)	28 (20)	116 (90)	184 (133)
Total number of samples	4569 (3740)	3474 (2910)	3303 (2745)	3270 (2717)	3238 (2708)	3028 (2411)	20 882 (17231)

Numbers in parentheses show data restricted to routine samples only. CIN2+=cervical intraepithelial neoplasia grade 2 or worse. CIN3+=cervical intraepithelial neoplasia grade 3 or worse.

Table 6: Analysis of BD FocalPoint GS Imaging System for samples allocated to quintiles and for samples designated as needing no further review, by suggested management outcomes

	Number of automated low-grade abnormalities (HPV positive) vs manual negative results	Number of automated high-grade abnormalities vs manual low-grade abnormalities or lower (HPV negative)	Total	Number of manual low-grade abnormalities (HPV positive) vs automated negative results*	Number of manual high-grade abnormalities vs automated low-grade abnormalities or lower (HPV negative)†	Total
Interpretation error						
Manual	23	8	31	0	0	0
Auto	0	0	0	29	17	46
Automated location error	NA	NA	NA	12‡	3§	15
Total	23	8	31	41	20	61

Six slides were not available for review. HPV=human papillomavirus. *One ThinPrep slide in this category could not be reviewed because of bad optical character reading (ie, the machine could not read the slide's barcode and display the corresponding fields of view) [A: ok] plus eight BD SurePath slides had been signed out on the machine in error, preventing the reviewers from accessing and reviewing the fields of views. †Seven BD SurePath slides could not be reviewed because they had been signed out on the machine, preventing the reviewers from accessing and reviewing the fields of views. ‡Total includes three BD SurePath slides that were classified as needing no further review. §Total included three BD SurePath slides that were classified as needing no further review.

Table 7: Reasons for discordance between automated and manual results

Table 2 shows paired cytological results, comparing the final automation results and final manual results in the paired arm. 1850 (4%) of 48 271 samples showed non-concordant results. Of these, 931 (50%) were abnormal with the manual method and negative with the automated method, with far fewer abnormal by the automated method but negative with the manual method (294 [16%] of 1850). The remainder of discordant pairs were negative or inadequate. A net difference of 35 high-grade results and 192 low-grade HPV-positive results in favour of manual reading was reported.

Our results show that a larger number of patients with CIN2+ lesions would probably be diagnosed with manual reading than with automated reading (table 3). 16 CIN2+ lesions that were negative on automated reading and final manual reading were diagnosed as a result of colposcopy and biopsy for women who might have been in follow-up for previous low-grade cytology, had previous treatment, or were referred for clinical reasons.

Automation-assisted reading was 8% less sensitive relative to manual reading for detection of CIN2+ (relative sensitivity 0.92, 95% CI 0.89–0.95), and 5% less sensitive for detection of CIN3+ (0.95, 0.91–0.99). When routine screening samples from women aged 25–64 years only were included in the analysis, automation-assisted reading was 9% (relative sensitivity 0.91, 95% CI

0.87–0.95) less sensitive relative to manual reading for detection of CIN2+ and 6% less for detection of CIN3+ (0.94, 0.89–0.99). Data for specificity (table 4) showed a difference of only 0.6% in favour of automation-assisted reading for all samples (relative specificity 1.006, 95% CI 1.005–1.007), and 0.4% for routine screening samples (1.004, 1.003–1.005) for cervical intraepithelial neoplasia grade 1 (CIN1) or less; values for CIN2 or less were 1.007 (1.006–1.008) for all samples and 1.005 (1.004–1.006) for routine screening samples.

On the assumption that absolute sensitivity of manual reading is 79%,⁴ then a relative sensitivity of 92% is equivalent to an absolute difference of 6.3%. There was no difference in sensitivity of each of the two automated systems relative to manual reading (relative sensitivity of BD FocalPoint 0.90, 95% CI 0.85–0.96 for CIN2+ and 0.91, 0.84–0.99 for CIN3+; relative sensitivity of ThinPrep 0.92, 0.87–0.98 for CIN2+ and 0.97, 0.91–1.03 for CIN3+), but the study was not sufficiently powered to show significant differences between the two (table 5).

Table 6 shows results from the BD FocalPoint GS Imaging System for samples allocated to quintiles and for samples designated as needing no further review, in which histopathological and ranking outcomes are correlated. This correlation was possible because the samples classified as needing no further review did not

determine management, which was based solely on the read result. As a result, a small number of CIN2+ and CIN3+ samples were classified as needing no further review; corresponding data restricted to routine screening samples for women aged 25–64 years showed even fewer were classified as needing no further review. 4569 (93%) of 4910 samples marked for no further review underwent rapid review. Of 26 samples deemed non-negative by the rapid reviewer, the most severe histological result obtained was CIN1. This result suggests that rapid review added nothing clinically significant to a sample classified as needing no further review and would be non-contributory.

A sample of discordant paired reads associated with underlying CIN2+ was reviewed. 15 of 61 cases showed no cytological abnormality, suggesting a location error by the automated technology (table 7). Most cases, however, showed interpretive errors. Few of these showed abnormal cells at the edge of the field of view, and there was a mixture of other abnormal features.

Discussion

The main finding of the MAVARIC trial was that sensitivity of automation-assisted reading was inferior to manual reading of cervical cytology (Panel). Although the aim of screening is to prevent cancer, we did not believe that a study with this endpoint was feasible within a reasonable timescale. CIN2+ was selected as an outcome measure because it is the internationally accepted histopathological abnormality that warrants treatment, and was viewed as a more meaningful and valid measure than a cytological abnormality, which is an intermediate measure in the diagnostic pathway.

Strengths of this study included the manual-only arm, which indicated that the manual reading in the paired arm was not different to that expected in everyday practice, and the real-life setting and adherence to routine protocol and quality-controlled practice. The procedures we used ensured blinding between paired readings, and although cytoscreeners were aware that an automated reading would be paired with a manual reading we have no reason to believe that their awareness would have affected the thoroughness of the read. Our study had sufficient power to show a predetermined limit of non-inferiority, and the use of HPV triage ensured a high degree of colposcopic verification of underlying disease without risking sample non-attendance for repeat cytology.

A study limitation could be that only one laboratory was used; however, the laboratory's reported rates of cytological abnormality and positive predictive values for CIN2+ for high-grade cytology are very much in the midrange for the English programme.¹⁴ The increased specificity achieved at the expense of reduced sensitivity was not sufficient to be clinically useful. In fact, the additional colposcopy triggered by the greater sensitivity of manual reading than that of automation-assisted

Panel: Research in context

Systematic review

Two systematic reviews of automated screening^{11,12} have both concluded that reliable conclusions could not be drawn because previous studies had not been of sufficient quality, and recommended studies with a robust reference standard and greater methodological strength. Studies of automated technologies have not addressed these issues. In our study we used liquid-based cytology and used a histological reference standard with automated readings paired with subsequent manual readings, one being independent of the other.

Interpretation

Our study provides a reliable estimate of the performance of automated reading of cervical cytology relative to manual reading. The finding that automated reading is less sensitive than manual reading in the detection of cervical intraepithelial neoplasia grade 2 or worse and grade 3 or worse means that we cannot recommend the adoption of automation-assisted cervical cytology in well organised screening programmes currently based on manual reading.

reading carried a positive predictive value of 19%, which was very much in line with that recorded for HPV-triaged borderline or mild abnormalities.²

Most discrepant cytological abnormalities associated with underlying CIN2+ were HPV-positive borderline or mild abnormalities (77 [68%] of 114), and arose because the rate of borderline or mild abnormality was lower (2039; 4.2%) for automation-assisted reading than for manual reading (2641; 5.5%). Notably, 317 discrepant samples had HPV-positive borderline or mild abnormalities on manual reading and were negative on automated reading; 47 (15%) had underlying CIN2+. This result shows that discrepant samples did not have less significant abnormalities than those identified within non-discrepant readings of low-grade abnormalities.

Samples of discordant pairs associated with underlying CIN2+ were reviewed. 15 (25%) cases had no abnormality seen on negative automated readings. This finding suggests that auto-location guided errors restricted the detection of abnormal cells to manual readings. In remaining cases, however, there were additional reasons for abnormal cells seen on review of negative automated readings, including that a substantial proportion of abnormal cells were at the periphery of the fields of view; which has also been reported by Halford and colleagues.¹⁶ Monotony was also reported by staff¹³ and could have been a contributing factor in reduced vigilance. Discrepant results associated with CIN2+ occurred throughout the study, indicating that discrepancies were not caused by improvements in staff skill over time. Productivity gains in terms of increased throughput of slides reported in other studies, and in the MAVARIC study,¹³ do not compensate for the magnitude of reduced sensitivity that we report.

Classification of samples as needing no further review by the BD FocalPoint system proved to be reliable in terms of negative predictive value, missing only 1% of CIN2+ lesions associated with routine screening samples. This system could be a valuable adjunct in primary screening because the module does not need the expensive workstations necessary for reading fields of view, and could reduce the number of slides requiring manual reading by 20%. Some consideration would have to be given to the need for rapid review of such cases, but rapid review is unlikely to be fruitful.

When the automated systems were individually compared with manual reading for routine screening samples there was no obvious difference in relative sensitivity, although the study was underpowered for a direct head-to-head comparison. This finding suggests that reduced sensitivity is a function of the reading of fields of view and that the systems have similar levels of accuracy in displaying abnormal cells to the reader.

Several studies of automation-assisted reading have been published while the MAVARIC study was underway. A Finnish trial¹⁷ compared outcomes for over half a million women randomised between manual reading and automation-assisted reading. A now obsolete PapNet system was used, with technical and logistics reasons preventing random allocation in 16% of cases. A slightly increased proportion of abnormal cytology was reported in the automated arm, but no increase was reported in detection rates of CIN2, CIN3, or cancer.¹⁷ A large study of the FocalPoint GS Imaging System in the USA,¹⁰ which used an adjudicated cytological outcome for each slide as the reference standard, reported a 20% increase in sensitivity for automation-assisted reading in the detection of high-grade cytology, but there was no association with histology. A similar study of the ThinPrep Imaging System in the USA,⁹ which used a cytology reference standard, reported a 6% increase in sensitivity of automation-assisted screening for atypical squamous cells of undetermined significance or worse, and equivalent sensitivity for low-grade squamous intraepithelial lesions or worse and high-grade squamous intraepithelial lesions or worse. In a study in Scotland,¹⁸ 110 000 samples were randomly allocated to either manual reading or automated reading using the ThinPrep Imaging System in six laboratories. The overall rate of high-grade abnormalities detected by manual reading was 1.38% and by automated reading was 1.45% ($p=0.512$), although rates ranged from 1% to 8%.

Our review of published works identified six studies that associated histological results with cytological results, but the published data did not allow determination of relative sensitivity or specificity. The clinical significance of cytological classification between manual reading and automation-assisted reading depends mainly on the consequent detection of disease. None of the published studies to date have allowed such a reliable comparison as did ours between manual reading and

automation-assisted reading for disease detection.

The MAVARIC trial results suggest that substantial changes to cervical screening programmes to accommodate automation-assisted reading cannot be justified. In terms of cervical screening policy, forward planning in several countries will increasingly include HPV testing. To achieve savings in terms of staff time, the FocalPoint system for classifying samples as needing no further review is, however, worthy of further consideration in programmes using liquid-based cytology.

Contributors

HCK, SM, RB, GD, MD, HC, RL, and AG contributed to the study design. HCK, SM, and GD wrote the report. SM, RB, and GD did the statistical analyses. HC and CM supervised and co-ordinated the virological assessments. MD, JM, and DNR supervised and co-ordinated the cytological assessments. HCK was the chief investigator. RA was the trial co-ordinator. LG was the data manager. All authors contributed revisions of the report.

Conflicts of interest

All authors' institutions received funding from the National Institute for Health Research Health Technology Assessment programme (NIHR HTA) to do the study. SM, RB, and CM received NIHR HTA funding to attend trial management group meetings. HC's institute received an M2000 system on loan from Abbott Diagnostics, and money from Abbott Diagnostics and Qiagen to reimburse travel expenses to scientific meetings. HCK received reimbursement for travel expenses from the International Academy of Cytology to present the results of this study at their 2010 scientific meeting.

Acknowledgments

This study was done under the guidance of an independent trial steering committee (David Torgerson, Maggie Cruickshank, and Karin Denton) and a data monitoring and ethics committee (Paula Williamson, John Smith, and Patrick Walker). We are grateful to Yvonne Hughes and the staff at the Manchester Cytology Centre for their cooperation and effort in accommodating the MAVARIC study. We acknowledge the use of a BD FocalPoint GS Imaging System provided free of charge by Medical Solutions for the initial 2 years of the study. We thank Qiagen for providing substantially discounted Hybrid Capture 2 kits. This project was funded by the National Institute for Health Research Health Technology Assessment programme (NIHR HTA; project number 03/04/02). The views and opinions expressed in this study are those of the authors and do not necessarily reflect those of the Department of Health. This work was supported by the NIHR Manchester Biomedical Research Centre.

References

- Arbyn M, Buntinx F, Van Ranst M, Paraskevaidis E, Martin-Hirsch P, Dillner J. Virologic versus cytologic triage of women with equivocal pap smears: a meta-analysis of the accuracy to detect high-grade intraepithelial neoplasia. *J Natl Cancer Inst* 2004; **96**: 280–93.
- Moss S, Gray A, Legood R, Vessey M, Patnick J, Kitchener H. Effect of testing for human papillomavirus as a triage during screening for cervical cancer: observational before and after study. *BMJ* 2006; **332**: 83–85.
- Schiffman M, Solomon D. Findings to date from the ASCUS-LSIL Triage Study (ALTS). *Arc Pathol Lab Med* 2003; **127**: 946–49.
- Arbyn M, Bergeron C, Klinkhamer P, Martin-Hirsch P, Siebers AG, Bulten J. Liquid compared with conventional cervical cytology—A systematic review and meta-analysis. *Obstet Gynecol* 2008; **111**: 167–77.
- National Institute for Clinical Excellence. Guidance on the use of liquid-based cytology for cervical screening (Technology Appraisal 69). London: National Institute for Clinical Excellence, London.
- Naucier P, Ryd W, Tornberg S, et al. Human papillomavirus and Papanicolaou tests to screen for cervical cancer. *N Engl J Med* 2007; **357**: 1589–97.

- 7 Bulkman NW, Berkhof J, Rozendaal L, et al. Human papillomavirus DNA testing for the detection of cervical intraepithelial neoplasia grade 3 and cancer: 5-year follow-up of a randomised controlled implementation trial. *Lancet* 2007; **370**: 1764–72.
- 8 Cuzick J, Clavel C, Petry K, et al. Overview of the European and North American studies on HPV testing in primary cervical cancer screening. *Int J Cancer* 2006; **119**: 1095–101.
- 9 Biscotti CV, Dawson AE, Dziura B, et al. Assisted primary screening using the automated ThinPrep Imaging System. *Am J Clin Pathol* 2005; **123**: 281–87.
- 10 Wilbur DC, Black-Schaffer WS, Luff RD, et al. The Becton Dickinson FocalPoint GS Imaging System: clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions. *Am J Clin Pathol* 2009; **132**: 767–75.
- 11 Broadstock M. Effectiveness and cost effectiveness of automated and semi-automated cervical screening devices: a systematic review of the literature. *N Z Med J* 2001; **114**: 311–13.
- 12 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**: 1–207.
- 13 Harris V, Sandridge A, Black R, Brewster D, Gould A. Cancer registration statistics, Scotland 1986-1995. Edinburgh: ISD Scotland Publications, 1998.
- 14 The Health and Social Care Information Centre. Cervical screening programme, England 2008-09. London, The Health and Social Care Information Centre, 2009.
- 15 Halford JA, Batty T, Boost T, et al. Comparison of the sensitivity of conventional cytology and the ThinPrep Imaging System for 1,083 biopsy confirmed high-grade squamous lesions. *Diagnostic Cytopathology* 2010; **38**: 318–26.
- 16 Kitchener HC, Blanks R, Cubie H, et al. MAVARIC—A comparison of automation assisted and manual cervical screening: a randomised controlled trial. *Health Technol Assess* (in press).
- 17 Nieminen P, Kotaniemi-Talonen L, Hakama M, et al. Randomized evaluation trial on automation-assisted screening for cervical cancer: results after 777,000 invitations. *J Med Screen* 2007; **14**: 23–28.
- 18 Scottish Cervical Cytology Review Group Feasibility Sub Group. Cervical Cytology ThinPrep Imager (TIS) Feasibility Study—Report from the Feasibility Sub Group to Cervical Cytology Review Group (2009). http://www.pathologyscotland.org/download/cervical_cytology_laboratory_provision/feasibility0911.pdf (accessed Aug 30, 2010).

20

25

30

35

40

45

50

55