Published in final edited form as:

Int J Cancer. 2009 December 1; 125(11): 2489–2496. doi:10.1002/ijc.24774.

How to evaluate emerging technologies in cervical cancer screening?

Marc Arbyn^{a,b}, Guglielmo Ronco^c, Jack Cuzick^d, Nicolas Wentzensen^e, and Philip E. Castle^e

^a Unit of Cancer Epidemiology, Scientific Institute of Public Health, Brussels, Belgium

^b ECCG (European Cooperation on development and implementation of Cancer screening and prevention Guidelines), IARC, Lyon, France

^c Unit of Cancer Epidemiology, Centro per la prevenzione Oncologica, Turin Italy

^d Cancer Research UK Centre for Epidemiology, Department of Mathematics and Statistics, Wolfson Institute of Preventive Medicine and Cancer Research, UK, Queen Mary University of London, London, UK

^e Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, USA

Abstract

Excellent recommendations exist for studying therapeutic and diagnostic questions. We observe that good guidelines on assessment of evidence for screening questions are currently lacking. Guidelines for diagnostic research (STARD), involving systematic application of the reference test (gold standard) to all subjects of large study populations, are not pertinent in situations of screening for disease that is currently not yet present. A five-step framework is proposed for assessing the potential use of a biomarker as a screening tool for cervical cancer: 1) correlation studies establishing a trend between the rate of biomarker expression and severity of neoplasia; 2) diagnostic studies in a clinical setting where all women are submitted to verification by the reference standard; 3) biobank-based studies with assessment in archived cytology samples of the biomarker in cervical cancer cases and controls; 4) prospective cohort studies with baseline assessment of the biomarker and monitoring of disease; 5) randomised intervention trials aiming to observe reduced incidence of cancer (or its surrogate, severe dysplasia) in the experimental arm at subsequent screening rounds.

The 5-phases framework should guide researchers and test developers in planning assessment of new biomarkers and protect clinicians and stakeholders against premature claims for insufficiently evaluated products.

Keywords

cervical cancer screening; human papillomavirus; HPV; biomarker; evaluation of diagnostic test; guidelines; health technology assessment

Principle of cytology-based screening

The rationale of cervical cancer cytological screening is to identify and treat high-grade cervical intraepithelial neoplasia^a (CIN) (precancerous lesion) to prevent its progression to invasive cancer⁴. Programme sensitivity is a convenient metric of assessing cancer reduction and population effectiveness although it does not account for the impact of false positives on cost-effectiveness, the negative consequences of over-screening, and the occurrence of side effects⁵. Programme sensitivity depends on the sensitivity of the chosen screening test, the compliance with further follow-up and the sensitivity of triage and diagnostic work-up, the natural history of the disease, and the screening policy (the target age group, screening interval, clinical thresholds for follow-up and treatment)⁶. The essential elements in the natural evolution of the disease are the rates of onset of precursor lesions, the progression and regression rates of these precursor lesions and the distribution of their sojourn times. The mean sojourn time (period from detectability of a lesion until it develops into a clinically manifest cancer) generally is believed to be in the order of 10 years or more with cytology and the probability of detection increases as the preclinical phase progresses^{7,8}. Sojourn times of cancer precursors are usually not observable because of treatment and are therefore only estimable by modelling. A unique (unethical) experience in New-Zealand, where CIN3 lesions were left untreated, allowed observation of the natural history. The 30-year cumulative incidence of invasive cancer among women with CIN3 was 30% and among women with persistent CIN3 was 50%. Because of the long natural history of precursors, repetition of a moderately sensitive screen test, such as the Pap smear, can achieve high programme sensitivity and thereby reduce incidence of and mortality from cervical cancer to a low residual level⁹. The International Agency for Cancer Research estimated that well-organised cytological screening for cervical cancer precursors every 3–5 years between the ages of 35 and 64 years reduces the incidence of cervical cancer by 80% or more among the women screened^{8,10,11b}. The success of screening depends essentially on the participation of the target population, the quality of the screening test and further on the compliance with follow-up and the efficacy of treatment of screen-detected lesions. The efficiency of screening decreases in subsequent rounds because successive sensitive screening followed by appropriate therapy reduces the endemicity of precursors over time. The lesions still found are smaller lesions with less invasive potential.

Shortcomings of cytological screening

The cross-sectional test accuracy of cervical cytology is highly variable since it depends on the availability of an adequately collected and prepared sample taken from the transformation zone and well-trained and motivated cyto-technicians for microscopic interpretation of the morphologic changes. By good quality assurance, a reasonably high sensitivity for high-grade CIN can be reached (>70%) but low sensitivity values (<50%) are not exceptional¹².

Because of low sensitivity, reported in several settings, alternative screening methods have been developed. We can distinguish four new methods of screening: a) alternative forms of cytology e.g., liquid-based cytology [see ref¹³ for a systematic review], automated or computer-assisted cytology; b) molecular detection of DNA or RNA of high-risk types of human papillomavirus (HPV), the virus causing cervical cancer¹⁴; and c) biomarkers associated with a progressive HPV infection such as immuno-staining of certain cell cycle regulating proteins whose expression has been altered, or maybe in the future, proteomic, transcriptomic or methylomic signatures of transforming HPV infections^{15,16} and d) biophysical changes

^aIn this paper “CIN” (cervical intraepithelial neoplasia) is used for histologically confirmed lesions, while the SIL (Bethesda) terminology is used to describe cytological findings, as recommended in recent international guidelines^{1–3}.

^bIt must be remarked that this estimate implies 100% compliance of screened women and that cancer occurring in women with lesions when screening starts are excluded from the estimate of 80% reduction.

identifiable by spectroscopy^{17,18}. In the rest of the paper, we discuss how such new techniques should be evaluated using – where possible - established methods to assess evidence of efficacy.

We first propose a methodology to rank evidence from published studies already performed. Subsequently, we propose a comprehensive framework for setting up new studies through which evaluation of biomarkers should pass to generate evidence on their potential application as a cancer screening test.

Levels of evidence of efficacy derived from published studies

Strength of evidence of screening effectiveness

A list of indicators for screening effectiveness, assessed by different study methods, is enumerated in Table 1 and ranked from high to low according to the level of evidence that such studies provide.

Randomised clinical trials (RCTs) designed to demonstrate a reduction in invasive cervical cancer provide the highest level of evidence of efficacy of screening. Observation of a lower incidence of cervical cancer in the trial arm where a new screening test is applied provides the proof that the new method (including the management of screen positives) is more effective than the control method. Nevertheless, conducting such studies requires enormous financial resources and huge study populations to be followed for many years including a high risk of contamination between the experimental and control arms^c. Meanwhile, during the lengthy interval to validate the new method, it may no longer be available or have become obsolete. Therefore, it is often proposed to study intermediate or surrogate outcomes (for instance outcomes 4 to 6 in Table 1) and to simulate the most likely outcomes relevant to public health using mathematical models. CIN3 is the direct precursor of invasive cancer and therefore, reduced incidence of CIN3+ is considered as an acceptable a proxy outcome of trials evaluating new preventive strategies^{19,20}. Prospective cohort studies do not allow obtaining more rapid results than randomised trials and suffer from several potential biases. Retrospective evaluation of previously identified cohorts can speed evaluation but not reduce bias. Case-control studies, comparing screening histories in women with and without cervical cancer are appropriate to evaluate effectiveness retrospectively but are also prone to several selection and information biases. Changes over time (secular trends) or geographical differences in incidence or mortality can be interpreted as screening effects but can only be accepted as indication of screening effectiveness when no other factors can plausibly explain the observed changes.

It must be stressed that the aim of screening is to prevent cervical cancer, not simply detect pre-invasive lesions. A new screen test allowing detection of more high-grade CIN does not necessarily result in more pronounced reduction of cancer incidence since just additional non-progressive lesions might be detected.

Cross-sectional test accuracy, threshold of disease

For screening, an accurate test is needed²¹: this means that it is positive when CIN2/3 is present and negative when CIN2/3 is not present. In other words, a screen test must have a good clinical sensitivity and specificity. The severity of CIN must be explicitly defined when assessing the accuracy of a test. CIN1 is the histopathologic manifestation of a carcinogenic or non-carcinogenic HPV infection that rarely progresses on a per event basis to cancer^{22,23}. Its detection is not clinically useful, possibly leading to over-treatment, and should not be targeted by any screening test. On the other hand, CIN2 and especially CIN3 indicate a considerable

^cContamination means that study subjects enrolled to participate to a trial arm do not follow the procedures foreseen in the study protocol. For instance: women randomised to screening with cytology are screened with an HPV test in the context of opportunistic screening.

risk of developing cancer and should therefore not be missed by a screen test. CIN2 is an intermediate condition, which contains overcalled CIN1 (caused by both carcinogenic and non-carcinogenic HPV), and under-called CIN3^{24–28}. CIN2 is a more regressive²⁹ and less reproducible histological diagnosis than CIN3²⁸. Thus, while a CIN2 diagnosis is typically the clinical threshold for triggering excisional or ablative treatment, its inclusion as an endpoint for evaluation of a screening test may exaggerate the overall impact of a screening test. The observation that a new screen test is more sensitive than the conventional test in detecting CIN3 provides more convincing evidence that its use in screening will result in a higher reduction in cancer incidence than the detection of CIN2/3, which can be artificially elevated due to the detection of low-risk CIN2 destined to regress (over-diagnosis). Whether detection of more CIN2 with a new method corresponds (at least partly) with either progressive or regressive disease, cannot be assessed from cross-sectional studies. However, observing, at the second screening round among women with a negative first screen test, less CIN3+, in the experimental compared to the control arm of a trial, indicates that at least a part of the additionally detected CIN2 was not regressive. The total amount of CIN2+ cases in first and second screening arm in the experimental over the conventional arm, represent a measure of over-diagnosis, not a measure of efficacy.

Therefore, future authors should be recommended to report cross-sectional accuracy separately for both CIN2+ and CIN3+.

Incomplete application of the gold standard, verification bias

The most comprehensive design for evaluating the cross-sectional accuracy of screen tests is the independent application of all the tests to a screening population followed by verification in all study subjects, irrespective of the screen test results, using a valid gold standard assessed without prior knowledge of the screen test results. Under these conditions unbiased estimation of the test sensitivity and specificity is possible. We invite readers to consult STARD guidelines³⁰ for good diagnostic research and QUADAS guidelines for evaluation of the quality of individual studies included in systematic reviews of diagnostic studies³¹.

Often, even in a research context (because of cost and/or ethical concerns), only women with positive screen tests and none or only a few with negative screen tests are verified and this situation results in verification bias yielding inflated sensitivity and underestimated specificity. Nevertheless, if multiple tests are evaluated and at least one test is very sensitive, the extent of verification bias is reduced, because virtually all women with CIN2/3 or CIN3 undergo diagnostic evaluation. Verification bias can be adjusted for if a random fraction of screen-negatives are referred for the application of the gold standard^{32–36}. Also long term follow-up can be used to capture missed disease²⁹.

When 2 screen tests are applied to the same study subjects and all subjects, positive for one or both tests, are verified with an acceptable gold standard, unbiased estimation of the test positive predictive value, the relative sensitivity and detection rate of true positives is possible^{37, 38d,e}. Thus, while the true absolute sensitivity cannot be determined, test performance can be ranked in an unbiased fashion. The same is true for randomised clinical trials, where different tests are applied to subjects in two or more study arms. For this reason, we believe that the Cochrane Collaboration should consider including such studies in systematic reviews (see further below). The reader should be warned that correction for verification bias by additional verification of test negative cases can yield erroneous results (sometimes even more biased

^dThe same is true when different tests are studied in different populations as long as the prevalence of disease can be assumed to be the same (e.g. in randomised trials)⁴.

^eWhen not all screen-positives are verified and the selection of verified positive cases is not random, verification bias still can occur at the level of the PPV, detection rate and relative sensitivity.

than the original verification bias) if subjects are not selected at random, see ref³⁹ for an example.

When the prevalence of disease is low (which always is the case in a screening setting) and only test-positive cases are verified, an *approximated test specificity* can be computed, (see formula).

$$Specificity_{approx} = \frac{Number\ of\ test\ negatives}{Total\ number\ tested - Number\ of\ true\ positives}$$

This *approximated test specificity* does not suffer from verification bias.

Reproducibility

The reliability or reproducibility of a test, including intra-batch and inter-batch reproducibility as well as intra-laboratory and inter-laboratory reproducibility, expresses the capacity to obtain the same test result – correct or not – when the screening test is repeated on the same individual. The reliability depends on the definition of distinct test criteria that can be applied by skilled personnel. Poor reproducibility automatically yields low average sensitivity and specificity. Reproducibility can be enhanced by training. Evaluation of new screening tests requires reproducibility experiments, preferentially including field circumstances.

Quality of the gold standard

Assessment of the gold standard, knowing the screen test result, includes a serious risk of overestimation of both the sensitivity and specificity. Therefore, in diagnostic research, where the objective is to evaluate the cross-sectional accuracy of a screen test, verification should be performed independently. This can be difficult when the screen test and the gold standard are based on the same principle, for instance in case of VIA screening (visual inspection of the cervix after application of acetic acid), validated using colposcopy^{40–42}.

It is usually assumed that histological examination of material obtained by colposcopically directed biopsy, loop excision or endocervical curettage, and – in absence of biopsy - a negative colposcopic impression provide a valid ascertainment of the true disease status. Recent data indicate that this assumption might not always be true^{43,44}.

Colposcopy performance has been challenged by results from prospective studies suggesting that up to 50% of prevalent precancers may be missed during colposcopy⁴⁵. The visual assessment of the cervix in colposcopy has a high inter-observer variability^{46,47,47}. It has been demonstrated that the sensitivity of colposcopy is not related to the experience of the colposcopist, but to the number of biopsies taken⁴³. In random biopsies from normal appearing regions on the cervix substantial disease has been identified⁴⁸. Again, follow-up can be used to compensate partially for the lack of sensitivity of colposcopy. As a consequence, one-time colposcopic-directed biopsy as it has been practiced should be considered an imperfect referent standard.

Currently, studies are underway that aim at analyzing better colposcopic procedures and at determining how many biopsies are necessary to improve disease ascertainment. Meanwhile, a combined endpoint including histology and cytology results can improve the disease ascertainment⁴⁹.

Longitudinal sensitivity

Once again, it must be repeated that the observation of increased cross-sectional sensitivity of a new test for histologically confirmed CIN2/3 or CIN3 does not necessarily imply that its inclusion in a screening programme will yield a reduction in incidence of lethal cervical cancer with respect to conventional cytological screening^f. Nevertheless, when biological and epidemiological arguments justify the assumption that the lesions detected in excess by the new method have a substantial chance of progression (acceptable longitudinal positive predictive value) and that screen negatives have a substantially lower chance to develop cancer in the future (higher longitudinal negative predictive value), planning of evaluating the new test in a randomised population- based trial in an organised setting can be considered⁵⁰. Audits of screening effectiveness, including linkages with screening and cancer registries, that allow picking up missed disease detected beyond the timelines of studies, are a particularly useful tool of evaluation^{51,52}. Finally, simulation models must help in identifying best choices but also in orienting the most influential issues to be addressed in future studies.

Costs of screening

Until now we studied essentially programme effectiveness, stressing test sensitivity. Cervical cancer screening involves large populations and therefore can be extremely costly. Costs are mostly determined by the test cost and specificity. An overview of the cost components attributed to screening is presented in Table 2.

Since the prevalence of progressive cervical precursors is very low the number of false positive cases results from the false positive rate applied to nearly the entire target population. Therefore even a small decrease in specificity can have serious consequences on costs, if the next step involves a complicated or invasive procedure. Nevertheless, the loss in specificity of a screen test can be limited by raising the screening interval, by increasing the age at onset of screening and by raising the cut-off for test positivity. Mathematical models can be used to estimate the final outcome per unit of cost, but rely on accurate estimates of the screening performance, which are not always available.

Comprehensive framework for setting up new studies for evaluation of biomarkers potentially applicable as a cancer screening test

The Cochrane Collaboration

The Cochrane Collaboration is a world-wide not-for-profit and independent organisation, dedicated to making up-to-date, accurate information about the effects of healthcare readily available worldwide. It produces and disseminates systematic reviews of healthcare interventions and promotes the search for evidence in the form of clinical trials and other studies of interventions. The Cochrane Collaboration essentially addresses therapeutic questions or effects of interventions, assessed by randomised clinical trials (conducted following the rules of good research practice: CONSORT guideline⁵³), and has developed a rigorous method for assessing and pooling of such trials (based on the QUORUM guidelines)⁵⁴. In 2007, at the Cochrane Colloquium in Sao Paulo, the Cochrane Diagnostic Test Accuracy Working Group officially launched the implementation of systematic reviews of diagnostic test accuracy in its Library. The original studies should involve testing subjects for the presence of a target disease with two (or more) tests (for instance a conventional and a new test) and, subsequently, submitting all tested subjects with a valid gold standard method (STARD guideline)³⁰. All

^fIt is important to distinguish cross-sectional and longitudinal accuracy parameters. Increased detection with a new test of CIN2 that will largely regress, will result in a higher cross-sectional sensitivity which is clinically not useful (over-diagnosis). In contrast, a screen-positive woman who, currently, does not have colposcopically visible CIN can develop a high-grade CIN2 in the future. Such a case may initially be classified as false-positive, only to be re-classified subsequently as a true-positive with longitudinal surveillance.

tests should be applied independently and nearly simultaneously, in a setting representative for the situation where the tests will be used. The *hierarchical summary ROC curve analysis* is an adequate statistical tool that allows summarizing accuracy estimates accounting for the intrinsic negative correlation between sensitivity and specificity corresponding with different test cutoffs⁵⁵. In the evaluation of a new biomarker as a potential screening method, it often is unfeasible, unpractical and even unethical to apply the gold standard (for instance excision biopsies). Moreover, it is possible that such 'gold standard' verification is unreliable when the target disease, is not yet detectable or, if the procedure detects lesions which have a high chance of spontaneous regression (over-diagnosis).

We agree that strict application of the Cochrane methodology for reviewing and the STARD guidelines³⁰ for original diagnostic studies will result in tremendous improvements of the quality of the research on diagnosis for current clinical disease. Nevertheless, more appropriate methods and longitudinal study designs are needed for screening studies aimed at identifying cancer precursors, where the target disease is not yet developed and where management is restricted to screen-positive subjects. The conceptual five-step evaluation process (see Table 3, below) will be of guidance as a paradigm for screen test evaluation⁵⁶. In particular, biobank-based case-control studies exploring presence of biomarkers in samples, collected years to decades before the outcome, can provide a powerful research tool, but still require investigations with respect to feasibility. We refer readers to a more extensive discussion of the use of stored cervical cytology samples as a resource for molecular epidemiology⁵⁷.

Following Pepe⁵⁶, five phases can be distinguished in a straight forward evaluation of biomarkers with the purpose of use in screening (see Table 3).

It is the intention of the authors to work out this conceptual model for cervical cancer screening including triage of screen-positive women. A major outcome would be a concept and guideline for the design and conduct of biobank case-control studies as also proposed recently by Pepe *et al* 58. This concept will require thorough discussion and levels of approval by international methodologists.

As one example, the triage of LSIL (and its equivalent, hr HPV-positive ASCUS) offers an interesting opportunity to evaluate the capacity of biomarkers to distinguish between regressing and progressing abnormalities using a biobank-based design. High-risk (hr) HPV testing is considered insufficiently specific^{59,60}. One could select prior cases of LSIL archived in the biobank and follow these up with repeat testing and registration (different algorithms are possible). After two or more years certain cases will have progressed and others regressed. Subsequently, one can retrieve the stored original LSIL samples from cases that progressed to high-grade CIN and from matched disease-free controls and apply one or more biomarker assays. When the new biomarker assay requires fresh samples, such biobank-based studies must be designed prospectively with concealed testing at baseline⁵⁸.

Two examples: high-risk (hr)HPV testing, over-expression of p16

hrHPV testing

Cervical cancer screening using detection of DNA of hrHPV types passed through all phases of evaluation (as listed in Table 3), although some RCTs are still running. It was already known for many years that hrHPV testing is more sensitive but less specific than cervical cytology⁶¹. More recently, randomised population-based trials have demonstrated that hrHPV-negative women older than 30–34 years, are at 47–71% lower risk of developing CIN3 or worse (CIN3 +) than women who have a negative Pap smear over the next 5 years^{62–64}. This reduction in the CIN3+ burden can be regarded as a proxy for reduced incidence of invasive cancer¹⁴. A large RCT, conducted in India, demonstrated lower incidence of and mortality from cervical

cancer in women testing HPV-negative compared to not-screened women, in contrast to women screened with visual inspection or cytology⁶⁵.

Triaging screen-positive women

HPV infection is common but usually transient. Reaching high sensitivity for detection of underlying high-grade CIN requires inclusion of all high-risk types in the assays, which inevitably reduces specificity because it includes weaker carcinogenic HPV genotypes²⁵. Therefore, when HPV-based screening for cervical cancer is considered, the challenge will be to identify appropriate triage algorithms that limit the burden of hrHPV positive women needing follow-up. Cytology triage is one possibility^{66,67}. Biomarkers which are widely expressed in transforming infections could also fulfil this role^{68,69}. Biomarkers can also be used to triage low-grade or borderline cytology^{60,70}, when cytology is used for primary screening.

Overexpression of p16

A recent meta-analysis (including mainly phase 1 studies) summarised the correlation between p16^{INK4a} (abbreviated as p16) over-expression and the severity of squamous cytological lesions, and demonstrated a high variation in the proportion of p16 positives (ranging between 10% and 100% in ASCUS [atypical squamous cells of undetermined significance] and between 24% and 86% in LSIL [low-grade squamous intraepithelial lesions]), underlining lack of standardisation in immuno-staining, interpretation and reporting⁷¹. Nevertheless, in experienced hands and using clearly defined criteria, p16 immuno-staining has shown excellent results with sensitivities for CIN2+ similar to hrHPV testing⁶⁰, remarkably lower positivity rates (27% in ASCUS, 24% in LSIL) and consequently substantially higher specificities (84% and 81%, in respectively in ASCUS and LSIL) (one phase 2 study)⁷².

Currently, we must acknowledge the lack of good triage studies comparing p16 with currently used alternative strategies to triage equivocal cytological results. Concerning triage of hrHPV positive women, we note only one recent Italian study where hrHPV testing followed by p16-enhanced cytology showed a higher sensitivity for high-grade CIN and similar referral rate to colposcopy compared to primary screening by non-stained conventional cytology⁷³. Pepe did not include triage studies in the framework of ranking evidence for efficacy of screening (Table 3). We propose to consider triage studies as providing evidence of level 2, if designed as a diagnostic study with concurrent gold standard assessment. Randomisation of two or more triage options including longitudinal outcome assessment (via screening and cancer registries, or via systematic gold standard assessment 2–3 years after triage testing) should be classified at a superior level (2+ level).

The question whether sufficient evidence exists to recommend p16-immunostaining as an alternative primary cervical cancer screening method must be answered negatively (many phase 1 studies⁷¹, a small number of pending phase 2 studies [C. Bergeron, personal communication], and one trial targeting p16-triage of HPV positive women [phase 2+]⁷³). Yet these promising results warrant further evaluation by for more powerful and well-designed studies (of higher phases).

In order to explore the potential to use p16 over-expression as a progression marker in triage, we propose to set up an international workshop to standardise issues of sample processing and to define clear criteria for categorising levels of positivity⁶⁰. In table 4, we propose a comprehensive set of studies, which are needed to demonstrate performance of p16 testing in screening.

Which requirements must be fulfilled for new tests similar to clinically validated existing ones?

This question intrigues not only the developers of new assays but also the public and health policy makers who wish to avoid dependency from one manufacture. It is agreed that lower-level evidence can be accepted for systems similar to those for which already sufficient evidence of efficacy is available.

Alternative cytology systems

Liquid-based cytology and/or automated cytology could be accepted as an alternative for conventional cytology if at least equal sensitivity and/or specificity, or preferentially, superior sensitivity and equal specificity or, equal sensitivity and superior specificity, using CIN2+ as outcome, can be demonstrated in a screening population. This can be achieved through a cross-sectional study with double testing (conventional and new assay) and blind interpretation of both assays and blind verification of subjects with cytological abnormality according to standard follow-up algorithms. A preferred alternative is the randomised trial, where colposcopists and histologists are blinded to the type of screen test. Example are the RCT currently being conducted in the Netherlands, comparing liquid and conventional cytology⁷⁴ and that conducted in Italy⁷⁵. In case of comparable accuracy, other elements, such as the proportion of unsatisfactory preparations, reading time, possibility of ancillary testing and costs should be considered, which can be done through a decision analysis.

hrHPV DNA testing assays

Accepting that screening using HC2 or GP5/6+ PCR significantly reduces the prevalence of CIN3+^{14,64g}, experts recently proposed that a new high-risk HPV test should reach a minimum relative sensitivity of at least 0.90 and a relative specificity of at least 0.98, using HC2 as comparator test and CIN2+ as threshold for disease. Moreover the new test should be highly reproducible (agreement >87%, minimum 500 samples)⁷⁶.

The future of molecular progression markers

Research for other new markers, based on molecular processes associated with carcinogenesis, should undergo all phases of evaluation. Possible applications of p16 immuno-cytochemistry, mRNA testing and HPV genotyping to secondary cervical cancer prevention are passing through the hierarchical path of generating evidence, unfortunately not always following the logic framework outlined in table 3. Triage of women with LSIL is a particularly pertinent research field for molecular biomarkers since neither hrHPV testing nor repeated cytology appear to be sufficiently discriminatory to find underlying or incipient relevant disease⁷⁷.

The expected reduction in background risk of several cancers brought about by future HPV vaccination will be an additional dimension that must be integrated in search of screening methods with an acceptably high predictive value^{78,79}. In fact, screen and follow-up strategies with high positive predictive value are also needed in well-screened populations, where over time, prevalent, large CIN3 with significant invasive potential will be preferentially detected and eliminated, leaving fewer CIN3 that have lower invasive potential. It is the intention of the authors to try assisting the research community by offering advice on future straight forward study designs. The environment of the Cochrane Review Collaboration, involving cooperation with methodologists in diagnostic research, clinicians and clinical epidemiologists could offer a fruitful forum to realise the ambition of assessing current and future evidence for cervical cancer prevention strategies.

^gLevel of evidence (see Table 1): outcome: reduction of CIN3+ (level 3); study type: RCT (level 1).

Acknowledgments

Financial support was received from: (1) The Belgian Foundation Against Cancer, Brussels, Belgium; (2) the Gynaecological Cancer Cochrane Review Collaboration (Bath, United Kingdom); (3) the European Commission (Directorate of SANCO, Luxembourg, Grand-Duchy of Luxembourg) through the ECCG (European Cooperation on development and implementation of Cancer screening and prevention Guidelines, IARC, Lyon, France) and the European Research EURO-COURSE (Optimisation of the Use of Registries for Scientific Excellence in research) Network, funded by the 7th Framework programme through the Comprehensive Cancer Centre South (Eindhoven, The Netherlands); (4) IWT (Institute for the Promotion of Innovation by Science and Technology in Flanders (through the Unit of Health Economics and Modelling Infectious Diseases, Vaccine & Infectious Disease Institute, University of Antwerp; project number 060081).

References

1. Solomon D, Davey D, Kurman R, Moriarty A, O'Connor D, Prey M, Raab S, Sherman ME, Wilbur D, Wright TC, Young N. The 2001 Bethesda System: terminology for reporting results of cervical cytology. *JAMA* 2002;287:2114–9. [PubMed: 11966386]
2. Herbert A, Bergeron C, Wiener H, Schenck U, Klinkhamer PJ, Arbyn M. European guidelines for quality assurance in cervical cancer screening: recommendations for cervical cytology terminology. *Cytopathology* 2007;18:213–9. [PubMed: 17635161]
3. Wright TC Jr, Massad LS, Dunton CJ, Spitzer M, Wilkinson EJ, Solomon D. 2006 Consensus Guidelines for the Management of Women With Abnormal Cervical Screening Tests. *J Low Genit Tract Dis* 2007;11:201–22. [PubMed: 17917566]
4. Morrison, AS. Screening in Chronic Disease. Vol. 2. Oxford University Press, Inc; 1992. p. 1-254.
5. Arbyn M, Kyrgiou M, Simoons C, Raifu AO, Koliopoulos G, Martin-Hirsch P, Prendiville W, Paraskevaides E. Peri-natal mortality and other severe adverse pregnancy outcomes associated with treatment of cervical intraepithelial neoplasia: a meta-analysis. *BMJ* 2008;337:a1284, 1–11. [PubMed: 18801868]
6. Arbyn, M.; Dillner, J.; Schenck, U.; Nieminen, P.; Weiderpass, E.; Da Silva, D.; Jordan, J.; Ronco, G.; McGoogan, E.; Patnick, J.; Sparen, P.; Herbert, A.; Bergeron, C. European Commission. Chapter 3: Methods for Screening and Diagnosis. In: Arbyn, M.; Anttila, A.; Jordan, J.; Ronco, G.; Schenck, U.; Segnan, N.; Wiener, H.; Daniel, J.; von Karsa, L., editors. *European Guidelines for Quality Assurance in Cervical Cancer Screening*. Luxembourg: Office for Official Publications of the European Communities; 2008. p. 69-152.
7. Hakama M, Chamberlain J, Day NE, Miller AB, Prorok PC. Evaluation of screening programmes for gynaecological cancer. *Br J Cancer* 1985;52:669–73. [PubMed: 4063143]
8. van Oortmarssen GJ, Habbema JD. Epidemiological evidence for age-dependent regression of pre-invasive cervical cancer. *Br J Cancer* 1991;64:559–65. [PubMed: 1911199]
9. van Oortmarssen GJ, Habbema JDF, van Ballegooijen M. Predicting mortality from cervical cancer after negative smear test results. *BMJ* 1992;305:449–51. [PubMed: 1392957]
10. Day N, Moss S, Berrino F, Choi NW, Clarke EA, Döbrösy L, Geirsson G, Habbema DF, Hakama M, Hougen A, Johannesson G, Langmark F, Macgregor JE, Magnus K, Malker B, Jensen OM, Nelson NA, Parkin DM, Pettersson F, Poll P, Prorok PC, Raymond L, van Oortmarssen GJ. Screening for squamous cervical cancer: duration of low risk after negative results of cervical cytology and its implication for screening policies. *BMJ* 1986;293:659–64. [PubMed: 3092971]
11. Day NE. Screening for cancer of the cervix. *J Epidemiol Community Health* 1989;43:103–6. [PubMed: 2687425]
12. Nanda K, McCrory DC, Myers ER, Bastian LA, Hasselblad V, Hickey JD, Matchar DB. Accuracy of the Papanicolaou Test in Screening for and Follow-up of Cervical Cytologic Abnormalities: A Systematic Review. *Ann Intern Med* 2000;132:810–9. [PubMed: 10819705]
13. 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. [PubMed: 18165406]
14. Arbyn M, Cuzick J. International agreement to join forces in synthesizing evidence on new methods for cervical cancer prevention. *Cancer Lett* 2009;278:1–2. [PubMed: 18930588]

15. Zhu X, Lv J, Yu L, Zhu X, Wu J, Zou S, Jiang S. Proteomic identification of differentially-expressed proteins in squamous cervical cancer. *Gynecol Oncol* 2009;112:248–56. [PubMed: 19007971]
16. Wentzensen N, Sherman ME, Schiffman M, Wang SS. Utility of methylation markers in cervical cancer early detection: Appraisal of the state-of-the-science. *Gynecol Oncol* 2009;112:293–9. [PubMed: 19054549]
17. Siddiqi AM, Li H, Faruque F, Williams W, Lai K, Hughson M, Bigler S, Beach J, Johnson W. Use of hyperspectral imaging to distinguish normal, precancerous, and cancerous cells. *Cancer* 2008;114:13–21. [PubMed: 18213691]
18. Cardenas-Turanzas M, Freeberg JA, Benedet JL, Atkinson EN, Cox DD, Richards-Kortum R, MacAulay C, Follen M, Cantor SB. The clinical effectiveness of optical spectroscopy for the in vivo diagnosis of cervical intraepithelial neoplasia: where are we? *Gynecol Oncol* 2007;107:S138–S146. [PubMed: 17908588]
19. Davies P, Arbyn M, Dillner J, Kitchener HC, Ronco G, Hakama M. A report on the current status of European research on the use of human papillomavirus testing for primary cervical cancer screening. *Int J Cancer* 2006;118:791–6. [PubMed: 16287075]
20. Pagliusi SR, Teresa AM. Efficacy and other milestones for human papillomavirus vaccine introduction. *Vaccine* 2004;23:569–78. [PubMed: 15630792]
21. Wilson, JMG.; Jungner, G. Principles and practice of screening for disease. Geneva: World Health Organisation; 1968. Public Health Papers 34
22. Ostor AG. Natural history of cervical intraepithelial neoplasia: a critical review. *Int J Gynecol Pathol* 1993;12:186–92. [PubMed: 8463044]
23. Holowaty P, Miller AB, Rohan T, To T. Natural History of Dysplasia of the Uterine Cervix. *J Natl Cancer Inst* 1999;91:252–8. [PubMed: 10037103]
24. Stoler MH, Schiffman MA. Interobserver reproducibility of cervical cytologic and histologic interpretations. *JAMA* 2001;285:1500–5. [PubMed: 11255427]
25. Schiffman MA, Herrero R, Desalle R, Hildesheim A, Wacholder S, Rodriguez AC, Bratti MC, Sherman ME, Morales J, Guillen D, Alfaro M, Hutchinson M, Wright TC, Solomon D, Chen Z, Schussler J, Castle PE, Burk RD. The carcinogenicity of human papillomavirus types reflects viral evolution. *Virology* 2005;337:76–84. [PubMed: 15914222]
26. Sherman ME, Schiffman MA, Cox JT. Effects of age and human papilloma viral load on colposcopy triage: data from the randomised atypical squamous cells of undetermined significance/low-grade intraepithelial lesion triage study (ALTS). *J Natl Cancer Inst* 2002;94:102–7. [PubMed: 11792748]
27. Sherman ME, Wang SS, Tarone R, Rich L, Schiffman MA. Histopathologic extent of cervical intraepithelial neoplasia 3 lesions in the atypical squamous cells of undetermined significance low-grade squamous intraepithelial lesion triage study: implications for subject safety and lead-time bias. *Cancer Epidemiol Biomarkers Prev* 2003;12:372–9. [PubMed: 12692113]
28. Carreon JD, Sherman ME, Guillen D, Solomon D, Herrero R, Jeronimo J, Wacholder S, Rodriguez AC, Morales J, Hutchinson M, Burk RD, Schiffman M. CIN2 is a much less reproducible and less valid diagnosis than CIN3: results from a histological review of population-based cervical samples. *Int J Gynecol Pathol* 2007;26:441–6. [PubMed: 17885496]
29. Castle PE, Schiffman M, Wheeler CM, Solomon D. Evidence for Frequent Regression of Cervical Intraepithelial Neoplasia-Grade 2. *Obstet Gynecol* 2009;113:18–25. [PubMed: 19104355]
30. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41–4. [PubMed: 12511463]
31. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:1–13. [PubMed: 12515580]
32. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;39:207–15. [PubMed: 6871349]
33. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol* 1992;45:581–6. [PubMed: 1607897]
34. Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient Study Designs to Assess the Accuracy of Screening Tests. *Am J Epidemiol* 1994;140:759–69. [PubMed: 7942777]

35. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford: Oxford University Press; 2003. p. 318
36. Ratnam S, Franco EL, Ferenczy A. Human papillomavirus testing for primary screening of cervical cancer precursors. *Cancer Epidemiol Biomarkers Prev* 2000;9:945–51. [PubMed: 11008913]
37. Schatzkin A, Connor RJ, Taylor PR, Bunnag B. Comparing new and old screening tests when a reference procedure cannot be performed on all screenees. Example of automated cytometry for early detection of cervical cancer. *Am J Epidemiol* 1987;125:672–8. [PubMed: 3826045]
38. Chock C, Irwig I, Berry G, Glasziou P. Comparing dichotomous screening tests when individuals negative on both tests are not verified. *J Clin Epidemiol* 1997;50:1211–7. [PubMed: 9393377]
39. Gaffikin L, McGrath J, Arbyn M, Blumenthal P. Avoiding verification bias in screening test evaluation in resource poor settings; a case study from Zimbabwe. *Clin Trials* 2008;5:496–503. [PubMed: 18827042]
40. Pretorius RG, Zhang X, Belinson JL, Zhang WH, Ren SD, Bao YP, Qiao YL. Distribution of cervical intraepithelial neoplasia 2, 3 and cancer on the uterine cervix. *J Low Genit Tract Dis* 2006;10:45–50. [PubMed: 16378031]
41. Gaffikin L, McGrath JA, Arbyn M, Blumenthal PD. Accuracy of visual inspection with acetic acid as a cervical cancer test validated using Latent Class Analysis. *BMC Med Res Methodol* 2007;7:1–10. [PubMed: 17217545]
42. Arbyn M, Sankaranarayanan R, Muwonge R, Keita N, Dolo A, Gombe Mbalawa C, Nouhou H, Sankande B, Wesley R, Somanathan T, Sharma A, Shastri S, Basu P. Pooled analysis of the accuracy of five cervical cancer screening tests assessed in eleven studies in Africa and India. *Int J Cancer* 2008;123:153–60. [PubMed: 18404671]
43. Gage JC, Hanson VW, Abbey K, Dippery S, Gardner S, Kubota J, Schiffman M, Solomon D, Jeronimo J. Number of cervical biopsies and sensitivity of colposcopy. *Obstet Gynecol* 2006;108:264–72. [PubMed: 16880294]
44. Pretorius RG, Kim RJ, Belinson JL, Elson P, Qiao YL. Inflation of sensitivity of cervical cancer screening tests secondary to correlated error in colposcopy. *J Low Genit Tract Dis* 2006;10:5–9. [PubMed: 16378026]
45. Jeronimo J, Schiffman M. Colposcopy at a crossroads. *Am J Obstet Gynecol* 2006;195:349–53. [PubMed: 16677597]
46. Jeronimo J, Massad LS, Castle PE, Wacholder S, Schiffman M. Interobserver agreement in the evaluation of digitized cervical images. *Obstet Gynecol* 2007;110:833–40. [PubMed: 17906017]
47. Massad LS, Jeronimo J, Schiffman M. Interobserver agreement in the assessment of components of colposcopic grading. *Obstet Gynecol* 2008;111:1279–84. [PubMed: 18515509]
48. Pretorius RG, Zhang WH, Belinson JL, Huang MN, Wu LY, Zhang X, Qiao YL. Colposcopically directed biopsy, random cervical biopsy, and endocervical curettage in the diagnosis of cervical intraepithelial neoplasia II or worse. *Am J Obstet Gynecol* 2004;191:430–4. [PubMed: 15343217]
49. Wentzensen N, Schiffman M, Dunn T, Zuna R, Walker J, Allen R, Zhang R, Sherman M, Wacholder S, Jeronimo J, Gold M, Wang S. A study of HPV genotype distribution, cytology, and histopathology among 1700 women referred to colposcopy in Oklahoma: implications for disease classification. *Int J Cancer* 2008;1–24.
50. Anttila, A.; Ronco, G.; Lynge, E.; Fender, M.; Arbyn, M.; Baldauf, JJ.; Patnick, J.; Mc Googan, E.; Hakama, M.; Miller, A. European Commission. Chapter 2: Epidemiological Guidelines for Quality Assurance in Cervical Cancer Screening. In: Arbyn, M.; Anttila, A.; Jordan, J.; Ronco, G.; Schenck, U.; Segnan, N.; Wiener, H.; Daniel, J.; von Karsa, L., editors. *European Guidelines for Quality Assurance in Cervical Cancer Screening*. Luxembourg: Office for Official Publications of the European Communities; 2008. p. 11–52.
51. Sasieni P, Adams J, Cuzick J. Benefit of cervical screening at different ages: evidence from the UK audit of screening histories. *Br J Cancer* 2003;89:88–93. [PubMed: 12838306]
52. Andrae B, Kemetli L, Sparen P, Silfverdal L, Strander B, Ryd W, Dillner J, Törnberg S. Screening-Preventable Cervical Cancer Risks: Evidence From a Nationwide Audit in Sweden. *J Natl Cancer Inst* 2008;100:622–9. [PubMed: 18445828]

53. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001;285:1987–91. [PubMed: 11308435]
54. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 1999;354:1896–900. [PubMed: 10584742]
55. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;8:239–51. [PubMed: 16698768]
56. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* 2001;93:1054–61. [PubMed: 11459866]
57. Arbyn, M.; Andersson, K.; Bergeron, C.; Bogers, JP.; von Knebel-Doeberitz, M.; Dillner, J. *Methods in Biobanking*. Tutowa (New Jersey, USA): The Humana Press Inc; 2009. Chapter 16: Cervical Cytology Biobanks as a Resource for Molecular Epidemiology. in-press
58. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst* 2008;100:1432–8. [PubMed: 18840817]
59. ASCUS-LSIL Triage Study Group. A randomized trial on the management of low-grade squamous intraepithelial lesion cytology interpretations. *Am J Obstet Gynecol* 2003;188:1393–400. [PubMed: 12824968]
60. Arbyn M, Martin-Hirsch P, Buntinx F, Van Ranst M, Paraskevaidis E, Dillner J. Triage of women with equivocal or low-grade cervical cytology results. A meta-analysis of the HPV test positivity rate. *J Cell Mol Med* 2009;13:648–59. [PubMed: 19166485]
61. Arbyn M, Sasieni P, Meijer CJ, Clavel C, Koliopoulos G, Dillner J. Chapter 9: Clinical applications of HPV testing: a summary of meta-analyses. *Vaccine* 2006;24(Suppl 3):S3-78–89. [PubMed: 16950021]
62. Naucler P, Ryd W, Tornberg S, Strand A, Wadell G, Elfgren K, Radberg T, Strander B, Forslund O, Hansson BG, Rylander E, Dillner J. Human papillomavirus and Papanicolaou tests to screen for cervical cancer. *N Engl J Med* 2007;357:1589–97. [PubMed: 17942872]
63. Bulkman N, Berkhof J, Rozendaal L, van Kemenade F, Boeke A, Bulk S, Voorhorst F, Verheijen R, van Groningen K, Boon M, Ruitinga W, van Ballegooijen M, Snijders P, Meijer C. 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;370:796–802.
64. Ronco, G.; Segnan, N.; Gillio-Tos, A.; Rizzolo, R.; Confortini, M.; Carozzi, F. Detection rate of high grade CIN 3 years after normal cytology and after normal HPV testing: preliminary follow up results from phase 1 of the NTCC randomised study. Beijing. Proceedings 24th International Papillomavirus Conference; 3–9 November, 2007; 2007.
65. Sankaranarayanan R, Nene BM, Shastri SS, Jayant K, Muwonge R, Budukh AM, Hingmire S, Malvi SG, Thorat R, Kothari A, Chinoy R, Kelkar R, Kane S, Desai S, Keskar VR, Rajeshwarkar R, Panse N, Dinshaw KA. HPV screening for cervical cancer in rural India. *N Engl J Med* 2009;360:1385–94. [PubMed: 19339719]
66. Cuzick J, Szaewski A, Cubie H, Hulman G, Kitchener HC, Luesley D, McGoogan E, Menon U, Terry G, Edwards R, Brooks C, Desai M, Gie C, Ho L, Jacobs I, Pickles C, Sasieni P. Management of women who test positive for high-risk types of human papillomavirus: the HART study. *Lancet* 2003;362:1871–6. [PubMed: 14667741]
67. Naucler P, Ryd W, Tornberg S, Strand A, Wadell G, Elfgren K, Radberg T, Strander B, Forslund O, Hansson BG, Hagmar B, Johansson B, Rylander E, Dillner J. Efficacy of HPV DNA testing with cytology triage and/or repeat HPV DNA testing in primary cervical cancer screening. *J Natl Cancer Inst* 2009;88–98. [PubMed: 19141778]
68. Cuschieri K, Wentzensen N. Human Papillomavirus mRNA and p16 Detection as Biomarkers for the Improved Diagnosis of Cervical Neoplasia. *Cancer Epidemiol Biomarkers Prev* 2008;17:2536–45. [PubMed: 18842994]
69. Lie AK, Kristensen G. Human papillomavirus E6/E7 mRNA testing as a predictive marker for cervical carcinoma. *Expert Rev Mol Diagn* 2008;8:405–15. [PubMed: 18598223]

70. 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. [PubMed: 14970277]
71. Tsoumpou I, Arbyn M, Kyrgiou M, Wentzensen N, Koliopoulos G, Martin-Hirsch P, Malamou-Mitsi V, Paraskevaidis E. p16INK4a immunostaining in cytological and histological specimens from the uterine cervix: a systematic review and meta-analysis. *Cancer Treat Rev* 2009;35:210–20. [PubMed: 19261387]
72. Wentzensen N, Bergeron C, Cas F, Vinokurova S, von Knebel DM. Triage of women with ASCUS and LSIL cytology: use of qualitative assessment of p16INK4a positive cells to identify patients with high-grade cervical intraepithelial neoplasia. *Cancer* 2007;111:58–66. [PubMed: 17186505]
73. Carozzi F, Confortini M, Palma PD, Del Mistro A, Gillio-Tos A, De Marco L, Giorgi-Rossi P, Pontenani G, Rosso S, Sani C, Sintoni C, Segnan N, Zorzi M, Cuzick J, Rizzolo R, Ronco G. Use of p16-INK4A overexpression to increase the specificity of human papillomavirus testing: a nested substudy of the NTCC randomised controlled trial. *Lancet Oncol*. 2008
74. Siebers AG, Klinkhamer P, Arbyn M, Raifu AO, Masuger LFAG, Bulten J. Cytological detection of cervical abnormalities using a liquid-based compared with conventional cytology: a randomized controlled trial. *Obstet Gynecol* 2008;112:1327–34. [PubMed: 19037043]
75. Ronco G, Cuzick J, Pierotti P, Cariaggi MP, Dalla PP, Naldoni C, Ghiringhello B, Giorgi-Rossi P, Minucci D, Parisio F, Pojer A, Schiboni ML, Sintoni C, Zorzi M, Segnan N, Confortini M. Accuracy of liquid based versus conventional cytology: overall results of new technologies for cervical cancer screening: randomised controlled trial. *BMJ* 2007;335:28. [PubMed: 17517761]
76. Meijer CJLM, Castle PE, Hesselink AT, Franco EL, Ronco G, Arbyn M, Bosch FX, Cuzick J, Dillner J, Heideman DA, Snijders PJ. Guidelines for human papillomavirus DNA test requirements for primary cervical cancer screening in women 30 years and older. *Int J Cancer* 2009;124:516–20. [PubMed: 18973271]
77. Arbyn M, Paraskevaidis E, Martin-Hirsch P, Prendiville W, Dillner J. Clinical utility of HPV DNA detection: triage of minor cervical lesions, follow-up of women treated for high-grade CIN. An update of pooled evidence. *Gynecol Oncol* 2005;99 (Suppl 3):7–11.
78. Franco EL, Cuzick J. Cervical cancer screening following prophylactic human papillomavirus vaccination. *Vaccine* 2008;26 (Suppl 1):A16–A23. [PubMed: 18642468]
79. Ronco G, Rossi PG. New paradigms in cervical cancer prevention: opportunities and risks. *BMC Womens Health* 2008;8:23. [PubMed: 19091066]

Table 1

Ranking of indicators by level of decreasing evidence for effectiveness of cervical cancer screening methods according to the studied outcome and the used study design (adapted from ⁶).

Outcome:

- 1 Reduction of mortality from cervical cancer, (quality-adjusted) life-years gained.
- 2 Reduction of morbidity due to cervical cancer: incidence of cancer (Ib+).
- 3 Reduction of incidence of cancer (including micro-invasive cancer).
- 4 Reduction of incidence of CIN3 or worse disease (CIN3+).
- 5 Increased detection rate of CIN3+ or CIN2+.
- 6 Increased test positivity with increased, similar or hardly reduced positive predictive value.

Study design^h:

- 1 Randomised clinical trial, randomised population based trial.
- 2 Cohort studies (possibly with imbedded case-control studies).
- 3 Case-control studies.
- 4 Trend studies, ecological studies on routinely collected data.

^h Only controlled studies are considered, this means studies where two or more screening methods are compared.

Table 2

Overview of cost components of a screening programme

1	Cost price of the screen-test (investment and recurrent costs); fees of health professionals (time for preparation, interpretation of the screen test, documentation, training); logistical costs (transport, processing, storage); administrative costs (invitation, registration and analysis of data).
2	Specificity of the screen test: cost of follow-up and treatment of women with false-positive results or having non-progressive screen-detected lesions (over-diagnosis).
3	Sensitivity of the screen test (longitudinal): cost for follow-up and treatment of true positives; this cost may be off-set by cost savings in avoided treatment of advanced disease.
4	Human costs: time spent by women to be screened, anxiety and discomfort for follow-up and/or treatment of women with true and false-positive results, increased risk of adverse obstetrical outcomes in treated women; consequences of delay in detection of cancer in false-negative women.
5	Specificity of quality control, triage and diagnostic follow-up procedures, contributing to increased positive predictive value and savings by avoiding treatment of false-positive women.
6	Quality of screen test procedures; satisfactory rate influencing the need for repeat tests.

Table 3**Phases in the evaluation of a biomarker for future use in cancer screening**

-
- Phase 1—"preclinical exploratory studies": assessment of markers in biosamples of cancer patients and healthy individuals or in a series of biopsies of selected subjects with no dysplastic lesions, mild, moderate and severe dysplastic lesions or in a series of cervical cell samples reported as negative, equivocal, low- or high-grade intraepithelial lesions.
 - Phase 2—"clinical assay development for clinical disease and assessment in non-invasive samples" (for instance Pap smears) in selected subjects with known outcome. Purpose: estimation of sensitivity and specificity in relation to test-cutoffs; ROC curve analysis. A typical example is a diagnostic study, in a colposcopy clinic, where the new test is applied and all women are verified with colposcopy and biopsies.
 - Phase 3—"retrospective longitudinal repository studies": for instance biobank-based case-control studies with cases selected (at random) from the cancer registry and controls selected according to appropriate matching variables from the population registry; assessment of biomarkers in samples years or decades before diagnosis of cancer in archival biosamples. Biosample degradation can be a major shortcoming. Its impact can be restricted by high-quality biobanking, and at least partially adjusted for by quality monitoring.
 - Phase 4—"prospective screening studies" involving baseline assessment of healthy subjects for presence of biomarkers and follow-up over time. The results of baseline assessment can be concealed or not.
 - Phase 5—"prospective intervention study" which preferentially should be a population-based randomised trial where screen-positive subjects are followed or managed when indicated. Aimed outcome: reduction in cause-specific mortality and/or incidence of invasive disease (beyond a certain stage) are the major outcomes.
-

Table 4

Studies needed to establish evidence to use p16-overexpression as a screening test for cervical cancer.

Phase	Study design
1.	Assessment of the p16 positivity rate in selected series of cervical cytology samples ASCUS, LSIL, HSIL and in biopsies without CIN, CIN1, CIN2, CIN3, AIS, squamous cancer, adenocarcinoma. A significant positive trend is established ⁷¹ . The test needs further standardisation. Systematic review to do on reproducibility of histological interpretation with versus without p16 staining.
2.	a) Diagnostic study in a colposcopy clinic, where biopsies are taken from all women referred for diagnostic work-up with p16 immunostaining on the cell samples that triggered referral. Outcome: absolute accuracy. b) Triage studies, including women with ASCUSL or LSIL (setting of cytology based screening) or hrHPV-positive women (setting of HPV-based screening), with p16 immuno-staining versus other triage tests, followed by colposcopy and biopsy of all women [outcomes: absolute sensitivity, specificity] or on those with a positive triage test [outcomes: PPV, relative PPV, detection rate of CIN2+, relative sensitivity]. Allocation of p16 versus other triage tests can be randomised. Outcome assessment can be done via cancer (screening) registries [outcome absolute risk of CIN3+ in p16+ and p16-].
2b.	Biobank-based case-control study, with as cases women with LSIL who subsequently progressed to CIN2+ and, as controls, women whose lesions regressed. If p16 on the index LSIL samples, is consistently over-expressed in cases and hardly recognisable in controls, the test can be used to predict prognosis.
3.	Biobank-based case-control study, including retrieval of archived Pap smears from women with cervical cancer selected from the cancer registry (cases) and from cancer-free women (age-matched controls). Biomarker assessment on retrieved Pap smears.
4.	Baseline immunostaining of screening samples and follow-up through cancer screening registries
5.	RCT comparing cytology- or HPV-based screening with p16-based screening.