ELSEVIER

Journal of Clinical Epidemiology ■ (2013) ■

Journal of Clinical Epidemiology

REVIEW ARTICLE

A framework provided an outline toward the proper evaluation of potential screening strategies

Wim J. Adriaensen[a,*], Cathy Mathei[a], Frank J. Buntinx[a,b], Marc Arbyn[c]

[a]Centre of General Practice, Department of Public health and Primary Care, Katholieke Universiteit Leuven, Kapucijnenvoer 33, Blok J, 3000 Leuven, Belgium

[b]Department of General Practice, Maastricht University, 6200 MD Maastricht, The Netherlands

[c]Unit of Cancer Epidemiology, Scientific Institute of Public Health, Juliette Wytsmanstreet 14, 1050 Brussels, Belgium

Accepted 17 September 2012; Published online xxxx

Abstract

Objectives: Screening tests are often introduced into clinical practice without proper evaluation, despite the increasing awareness that screening is a double-edged sword that can lead to either net benefits or harms. Our objective was to develop a comprehensive framework for the evaluation of new screening strategies.

Study Design and Setting: Elaborating on the existing concepts proposed by experts, a stepwise framework is proposed to evaluate whether a potential screening test can be introduced as a screening strategy into clinical practice. The principle of screening strategy evaluation is illustrated for cervical cancer, which is a template for screening because of the existence of an easily detectable and treatable precursor lesion.

Results: The evaluation procedure consists of six consecutive steps. In steps 1−4, the technical accuracy, place of the test in the screening pathway, diagnostic accuracy, and longitudinal sensitivity and specificity of the screening test are assessed. In steps 5 and 6, the impact of the screening strategy on the patient and population levels, respectively, is evaluated. The framework incorporates a harm and benefit trade-off and cost-effectiveness analysis.

Conclusion: Our framework provides an outline toward the proper evaluation of potential screening strategies before considering implementation.

Keywords: Cancer screening; Screening evaluation; Cervical cancer; Screening strategy; Harm benefit; Framework

1. Introduction

Almost 40 years ago, Wilson and Jungner [1], for the World Health Organization, formulated a number of criteria (called ''principles''), which a screening strategy should meet. One of the criteria was that there should be a suitable screening test or examination detecting latent or early phases of the target disease. Unfortunately, still no comprehensive guideline exists concerning the assessment of screening strategies. Moreover, the specific context of screening applied to large groups of apparently healthy persons among whom the disease usually is rare, makes the evaluation of screening strategies a difficult, delicate, and costly exercise.

In this article, we propose a comprehensive framework for the evaluation of new screening strategies, using cervical cancer screening as a case example. When dealing with terms such as screening tests, strategies, or programs, clear definitions should be made. Evaluation of a potential screening strategy, involving a new screening test, comprises the test, patient, and population level. Generally, it includes the determination of age ranges and screening intervals and assessment of its cost-effectiveness. Although the effectiveness of a screening program depends on the properties of the screening test itself, other factors including natural history of the disease, screening organization, level of participation of the target population, compliance with follow-up and efficacy of workup, and treatment of the screen-detected lesion also determine the success [2−5]. The evaluation of screening programs in all their aspects, however, lies beyond the scope of this article, which focuses on the evaluation of new screening strategies.

Our reasoning starts from the assumption that before a possible screening strategy is considered, a clear decision has been made on the exact aim and the general target of the intervention. The aim should be formulated as the net benefit for the screenees and in terms of avoiding

Conflict of interest statement: the authors report no biomedical financial interests or potential conflicts of interest.

* Corresponding author. Tel.: +32-1633-2730; fax: +32(0)16 337 480.
E-mail address: wim.adriaensen@med.kuleuven.be (W.J. Adriaensen).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2012.09.018

What is new?

Key findings

- The development of a comprehensive framework building further on the existing concepts, based on a stepwise evaluation process including a harm and benefit trade-off and cost-effectiveness analysis of the screening tests before introduction into clinical practice as a screening strategy.

What is the implication and what should change now?

- New screening tests should go through a proper evaluation process before considering implementation as a screening strategy, to avoid doing more harm than benefit.

worsening public health. The broad target population can be, depending on the target condition, either an age and sex subgroup of the general population or a high-risk subgroup, for example, people working or living in specific conditions or exposed to risk factors [6]. These general ideas will guide the researcher to precise screening intervals and target populations chosen for the individual observational studies and trials.

1.1. Methodological considerations that are specific for the evaluation of a screening strategy

When people actively present with a health problem that requires treatment, they accept that the diagnostic process or treatment carries some risk of inflicting harm. When the same processes are applied to healthy people, the acceptable level of risk is much lower. Additionally, motivation for screening often is encouraged by invitations and often includes some degree of social pressure.

1.1.1. Cervical cancer screening

Screening can effectively prevent cervical cancer. The International Agency for Research on Cancer estimated that well-organized cytologic screening for cervical cancer precursors every 3−5 years between ages of 35 and 64 years can reduce the incidence of cervical cancer by 80% or more among the women screened [7]. Nevertheless, cervical cancer was worldwide the third most common cancer in women and the fourth most common cause of cancer death and even the most common cause in many developing countries in 2008 [8]. It occurs at a relatively young age when women are actively involved in their careers or caring for their families, resulting in proportionally more life-years lost compared with most cancers [8]. The rationale of cervical cancer cytologic screening is to identify and treat high-grade cervical intraepithelial neoplasia (CIN) or precancerous lesion and prevent its progression to invasive cancer. The mean time of initial dysplasia to invasion is at least 10 years, and the probability of detection increases as the preclinical phase progresses [9,10]. Removing these precursor lesions is effective in avoiding progression to invasive malignancy. Although screening for cervical cancer is well established, there were until recently no randomized clinical trials to demonstrate its effectiveness. The observational evidence, however, showing a reduced incidence of and mortality from cervical cancer is widely accepted [11,12]. The recognition of a strong causal relationship between the persistent high-risk human papillomavirus (HPV) infection of the genital tract and occurrence of cervical cancer has resulted in the development of several HPV detection systems providing new preventive strategies that could potentially result in an even greater reduction in incidence and mortality than cytology.

1.1.2. Outcome of screening

The main purpose of screening is to reduce the disease-specific mortality. Therefore, the primary indicator of effect is the observed disease-specific mortality compared with the expected mortality in the absence of screening, best expressed in terms of absolute risk difference or its reciprocal, the number needed to screen. In addition, several alternative end points can be used as a proxy. Table 1 shows a list of indicators used to establish effectiveness of cervical cancer screening ranked by decreasing level of evidence [13].

Studying cervical cancer mortality is particularly difficult because the certified cause of death often does not indicate the exact anatomical origin but rather is indicated as death from uterine cancer. An alternative end point can be all-cause mortality as has been advocated for breast cancer [14], but a significant effect on all-cause mortality is rarely demonstrable with screening. In cervical cancer screening, in which precursors are detected and treated, reduction in cervical cancer incidence, is a convincing end point, but reaching this outcome requires hundreds of thousands of women to monitor over many years. CIN3 as a direct precursor of invasive cancer is an acceptable proxy outcome of effectiveness [15]. The increased detection of CIN2+ or CIN3+ is clinically not so relevant as they rarely progress to cancer [16], leading to overtreatment. Consequently, outcomes 6 and 7 in Table 1 should not be targeted by a screening strategy.

1.1.3. Screening for low-prevalent diseases

Contradictory to most diagnostic studies, in screening, the prevalence of disease, especially for cancer, is typically low. This has an impact on the predictive values. Sensitivity is an indicator of the proportion of detected and missed prevalent predisease and determines the effectiveness. A very high specificity is needed to minimize the number of false-positive test cases. However, a high specificity can still be associated with high absolute numbers of false-positive test results (and thus anxiety, costs, and additional procedures in a lot of people) in case of low prevalence, for

Table 1. Possible outcome measures for benefit in cervical cancer screening [12]

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

Abbreviation: CIN, cervical intraepithelial neoplasia.

instance in a population that is well covered by HPV vaccination. A sensitivity and specificity of 90% can relate to around 78%, 92%, and 97% of false-positive results if the outcome prevalence is 3%, 1%, and 0.3%, respectively, proportions we are dealing with in screening situations.

1.1.4. Benefits and harms of screening

What should a screening intervention in healthy people achieve? With regard to cervical cancer, people who profit from screening are those who (1) would have died of the cancer but are cured, owing to earlier detection; (2) would have been successfully treated for their cancer anyhow, but whose quality of life is improved owing to earlier detection (down staging) and less mutilating treatment; and (3) do not have a cancer or cancer precursor and are reassured by the negative results of a screening test that correctly shows that they do not have the disease.

However, screening can also be harmful. People who might be harmed by screening are those who (1) die from a screen-detected cancer and whose clinical course was not improved by treatment; (2) have cancer that normally would have showed up clinically at a later point of life, but whose mortality and morbidity do not differ compared with without early detection; (3) have screen-detected nonprogressive cancer precursor, resulting in overdiagnosis and unnecessary treatment; (4) have cancer or progressive precursor but have a false-negative screening test result leading to a false feeling of security and delayed effective diagnoses and possibly delayed diagnosis and treatment; and (5) have a false-positive result, which results in anxiety or unnecessary further investigation and treatment. A screening strategy may cause both benefits and harms, resulting in the need to trade-off into a net benefit or harm [2,4].

1.2. Important biases in studies that evaluate the effects of screening

For screening, the randomized controlled trial (RCT) with mortality (or incidence of overt disease when the screening targets preclinical disease) as the outcome and intention-to-treat analysis is the only study design that allows unbiased comparison of outcomes in screened and unscreened groups.

All the observational studies are prone to biases, which can be summarized into two broad types of bias: selection and information biases. There are, however, some biases particularly important for studies evaluating screening strategies. In observational studies comparing screened and unscreened people, those whose disease was diagnosed through screening can appear to survive longer than those who presented with symptoms, even if there is no benefit from screening. This is caused by clinical symptoms presenting later in the natural history compared with abnormal screening results, which is called ‘‘lead time bias.’’ In addition, for cancer screening, tumors that are detected as a result of screening are more likely to be indolent, slow growing, or less aggressive than tumors in nonscreened patients who present with symptoms or in the interval between two scheduled rounds of screening (interval cancers). This phenomenon is referred to as ‘‘length bias’’ and results in the false conclusion that patients die less or later if their cancer is detected by screening. *Overdiagnosis* is an extreme more general case of length bias: it refers to the detection of nonprogressive (pre)disease, which would never have caused overt disease. This results in unnecessary treatment and, at the very least, cause anxiety and possible adverse effects [17,18]. Whether a screen-detected predisease is an overdiagnosis cannot be determined in the individual case.

In the case of HPV-based cervical cancer screening, for example, the extent of overdiagnosis can be estimated from randomized trials, in which an initially elevated incidence of cancer or precursor in the HPV arm during the first screening rounds persists during subsequent years or screening rounds [6,19].

In a diagnostic study, all subjects, test positives and test negatives, should receive the reference test or gold standard to assess the accuracy of the test [20]. However, in screening, because of ethical or cost considerations, especially when the ascertainment of true disease status requires invasive testing, either none or a small proportion of patients whose test results are negative, may receive the reference test. This results in an inflated estimate of the sensitivity and an underestimated estimate of the specificity. However, this verification bias does not influence the relative sensitivity (=detection rate of confirmed disease among subjects with positive screen test A vs. screen test B) nor the relative positive predictive value (PPV) in studies comparing the effect of two or more screening tests [13,21].

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, in which the objective is to evaluate the cross-sectional accuracy of a screening test, verification should be performed independently. This can be difficult when screening test and gold standard are based on the same principle, for instance in case of visual inspection with acetic acid screening (visual inspection of the cervix after application of acetic acid) validated using colposcopy. It is usually assumed that colposcopy followed by histologic examination of material obtained from suspected areas provides a valid ascertainment of the true disease status, making it

the gold standard. Recent prospective studies, however, suggested that up to 50% of prevalent precancers might be missed during colposcopy [22,23], which also has a high interobserver variability. Random biopsies from normal-appearing regions and follow-up can be used to compensate partially for the lack of sensitivity of colposcopy.

2. A framework for the evaluation of screening strategies

Introducing a new screening strategy in clinical practice requires the evaluation of its characteristics as an added value to the existing procedures. Guidelines regarding appropriate study designs to address questions on benefits of screening for disease precursors in which the target disease is not yet present and in which the management is restricted to screen positives are urgently needed [13]. We propose a stepwise framework for the evaluation of screening strategies building further on the work by Van den Bruel et al. [24] and Haynes and You [25] for the evaluation of diagnostic tests, possible diagnostic trial designs described by Lijmer and Bossuyt [26], and detailed concepts proposed by Arbyn et al. [13], Pepe et al. [27], and Harris et al. [3] (Fig. 1).

At each step, questions are raised which are progressively more relevant for the screening strategy, providing

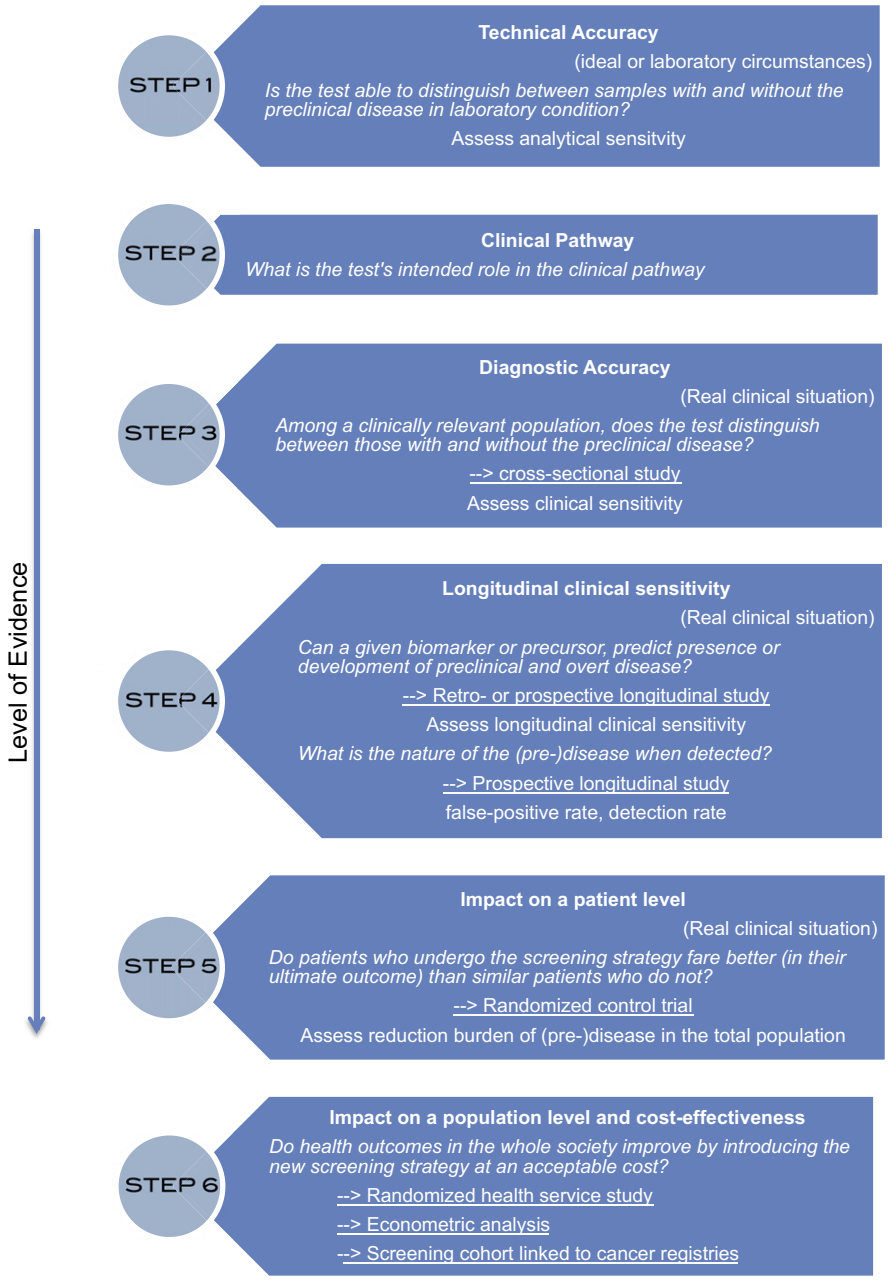

Fig. 1. Evaluation of a new screening strategy.

a higher level of evidence, but requiring also a more stringent study design and resources. Not all proposed study designs will be necessary, only the best possible design that is appropriate for the intended place of the test in the screening pathway. It is important to stress that no further research is planned when results at a certain step are unsatisfactory, and the strategy is only definitely evaluated if all the steps of the checklist have been addressed, taking all relevant aspects of health into consideration [2−5,28].

2.1. Step 1: Technical or analytical accuracy

> Is the test able to distinguish between samples with and without the preclinical disease in laboratory conditions?

Every screening test under evaluation should be assessed with respect to its technical accuracy. The *technical accuracy* of a test refers to its ability to produce usable information under research conditions. First, a correlation between the level of expression of a biomarker and severity of the predisease should be established. Then, the *analytical sensitivity* of a test is its ability to detect a well-defined preclinical condition or specified quantity of a biomarker. The analytical specificity refers to the ability of an assay to detect only that specific biomarker or preclinical condition and not closely related biomarkers or preclinical conditions. The *reproducibility* is the ability to obtain the same test results on repeated testing or observations. Reproducibility is influenced by analytical variability and observer interpretation. Analytical variability is the result of imprecision (random error) and inaccuracy (systematic error or bias).

2.2. Step 2: Place in clinical pathway

> What is the test's intended role in the clinical pathway?

New screening tests will eventually become a part of a clinical pathway. Knowing whether a new test is intended to replace an existing test, act as a triage for subjects being positive for the existing screen test (aiming to increase the specificity of the existing screening strategy) or act as an add-on screening test (aiming to increase the sensitivity of the screening strategy), is crucial [29]. In the add-on option, we can distinguish two suboptions: all subjects receive the two tests, or the new test is offered only to those who are negative with existing test. This potential place will determine the characteristics the new test needs to have and guide the researcher in the choice of the study designs that will (not) be needed to evaluate its efficacy. However, besides the results from these targeted studies and trials, the strategy's costs, participation rates, and so forth can ultimately result into a different optimal place of the screening strategy in the clinical pathway (steps 4−6).

HPV DNA testing has been considered as a potentially useful screening test solely (replacing cytology), as a triage test, or in combination with a papanicolaou (PAP) smear (add-on option) to detect cervical cancer precursors. As a candidate to replace PAP smear as the primary screening test, HPV DNA detection should have either better diagnostic sensitivity or less harm (higher specificity) or both. Studies have shown that HPV DNA testing with validated assays such as the hybrid capture 2 test and general HPV primers GP5+/6+ polymerase chain reaction is indeed substantially more sensitive in identifying CIN2, CIN3, or cancer than cytology at cutoff of atypical squamous cells of undetermined significance (ASCUS) or low-grade squamous intraepithelial lesion (LSIL), but it is less specific [30].

A related study question is whether a new marker can improve an existing triage procedure. Until recently, women with minor cytologic abnormalities were followed by repeat cytology. A recent meta-analysis concluded that follow-up of women with equivocal cytologic lesions by HPV testing with the hybrid capture 2 assay is more sensitive and equally specific in finding high-grade CIN compared with repeat cytology [31]. HPV testing could therefore be a better triage strategy than repeat cytology to select those women with minor cytologic lesions in a PAP smear who require referral for diagnosis and treatment.

2.3. Step 3: Diagnostic accuracy—cross-sectional study

> In a clinically relevant population, does the test distinguish between those with and without the preclinical disease?

The accuracy of a screening test is the test's ability to correctly detect or exclude a precursor or early stage of disease in a clinically relevant population. It should be noted that an increased analytical sensitivity does not necessarily imply an increased detection of preclinical disease in a relevant population (clinical sensitivity) (Table 2). Cross-sectional accuracy of a screening test is most comprehensively evaluated in a clinical setting in which all women irrespective of the screening test result—or at least all test positives and a random sample of all test negatives—are submitted to verification by a valid reference standard without prior knowledge of the screen result. The use of a random sample of test negatives requires weighting of the results during analysis. It should be noticed that correction for verification bias by additional verification of test-negative cases can yield erroneous results (sometimes even more biased than the original verification bias) if subjects are not selected at random [32]. The clinical sensitivity and specificity are used as outcomes if the biomarker is measured on a binary scale, or receiver operating characteristic curves are set up in case of a test with a continuous or ordinal scale.

For example, the rather low sensitivity of PAP smear and the high proportion of unsatisfactory samples prompted the development of new technologies such as new sampling devices and liquid-based cytology (LBC). A candidate-screening test for the replacement of conventional PAP

Table 2. Different screening strategy sensitivities

Sensitivities	Definition	Comment	Cervical cancer example (two HPV tests)
Analytical	Ability to detect a specified quantity of a biomarker or defined preclinical condition in laboratory conditions	A test is analytically sensitive when a minimal amount of expression yields a positive signal	HPV test 2 is positive even when the viral load is very low, whereas HPV test 1 is positive only beyond a higher viral load cutoff.
Clinical (cross-sectional)	Ability to pick up subjects with relevant disease	Patients with disease or predisease	HPV test 2 is more often positive than test 1, but it does not detect more CIN2+ than test 1.
Clinical (longitudinal)	Sensitivity to pick up current and future clinically relevant disease	The capacity to detect early disease (present or incipient)	Certain HPV-positive subjects do not yet have disease (currently false positives) but can develop the disease in subsequent years (future true positives) or regress (future false positives).

Abbreviations: HPV, human papillomavirus; CIN, cervical intraepithelial neoplasia.

smear should at least have equal diagnostic accuracy. Only a few studies are available with complete assessment using a gold standard, which permits evaluation without verification bias [33–35]. Available studies did not reveal a statistically significant difference between conventional and LBC in clinical sensitivity or specificity for detecting CIN2.

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. When, for example, two screening tests are applied to the same study subjects and all subjects positive for one or both tests are verified with an acceptable reference standard, unbiased estimation of each test's positive predictive value, and the relative sensitivity and detection rate of true positives is possible [13,21,36]. Thus, although the true absolute sensitivity cannot be determined, test performance can be ranked in an unbiased fashion.

So, measurement of sensitivity relative to a reference standard offers a good framework to compare alternative screening tests. However, an RCT may be needed to evaluate the performance of the new screening strategy at the patient and population levels (steps 4–6), even if a new test fares no better compared with an existing screening test conducted with the same reference standard. For example, although no statistical difference could be showed between LBC and the conventional assay, the proportion of unsatisfactory samples is lower in LBC, and the interpretation requires less time. The cost of an individual LBC test is considerably higher, but it allows ancillary testing such as HPV detection.

Note that it is important to clearly define the relevant threshold of disease, for instance CIN2 or worse CIN2+ or CIN3+. Whether the detection of more CIN2+ corresponds with progressive disease rather than with regressive disease cannot be assessed from cross-sectional studies.

2.4. Step 4: Longitudinal sensitivity

> Can a given biomarker or precursor, predict presence or development of preclinical and overt disease?

2.4.1. Retrospective design

After the cross-sectional evaluation in subjects with known outcome, a retrospective longitudinal study can be set up if a relevant biobank is installed for some time, to provide evidence regarding the capacity of the biomarker to detect present or future disease or its precursor (longitudinal sensitivity), if earlier detection is wanted (Table 2). Also, the impact of covariates should be explored to better select possible appropriate subpopulations and determine a screening interval if repeated screening is of interest.

A biobank-based case–control study on archived cervical smears from women with cervical cancer selected from a cancer registry and matched controls can enable the detection of HPV DNA in the most recent and earlier samples [28]. The sample selection should be representative of the samples that would be collected from a target population. Such study is only feasible if such a biobank is available and accessible and if the biomarker is not prone to degradation over time or the samples are not collected to a prespecified time schedule [13].

2.4.2. Prospective design

If a retrospective study is not feasible, we should follow patients over time with a well-conducted longitudinal observational or randomized study to have proof for a substantial likelihood of progression (longitudinal sensitivity) and knowledge about the nature of progressive precursors. A prospective study reduces biases inherent to retrospective studies because the outcome is unknown at the time of exposure documentation. Results from such a study can also clarify the level of overdiagnosis, although time can be an issue when death of cancer is the adopted end point. For cervical cancer, however, CIN3+ is the end point adopted in the currently running European trials, which permits the estimation of overdiagnosis with a reasonable study size and duration [37].

2.5. Step 5: Impact on the patient level

> Do patients who undergo this screening strategy fare better (in their ultimate health outcomes) than similar patients who do not?

The final aim of screening is to prevent the burden of overt disease with minimal harm caused and not simply to detect preclinical disease. For cervical cancer screening, harms may include overdiagnosis and unnecessary treatment, long-term anxiety because of labeling, test-related anxiety, discomfort, time investment, and minor or major adverse reactions (pain, hematoma, anaphylaxis, etc.) [2−4,28]. By balancing the benefits against the harms, a net benefit on patient health can be derived. To have direct evidence on effectiveness or at least evidence on overdiagnosis of regressive predisease with minimal bias, different groups need to be screened, managed, and treated according to different strategies and followed over time to observe the possible occurrence of disease. Only an RCT or a well-designed cohort study comparing large populations can provide conclusive evidence on the net benefit and actual impact of screening because observational studies are prone to selection bias and to confounding by other known and unknown variables that can explain differences between populations.

Thus, both steps 4 and 5 can be covered by an RCT, although sometimes a retro- or prospective observational study can be preferred before or instead of an RCT. An RCT may not always be necessary or feasible because of ethical considerations or large sample sizes needed. PAP smear screening for instance was never evaluated in a randomized trail because of this reasons. Nevertheless, the epidemiologic evidence indicates a very substantial beneficial effect of cytologic screening on cervical cancer incidence and mortality. Also, the effects of randomized trials may not be generalizable as the high-quality setting of trials run by dedicated scientists may be different from normal clinical care. This can be avoided by applying the new screening strategies within the routine screening activity. An acceptable methodological approach is the initial introduction of the new policy in a random selection of regions or birth cohorts using the other regions as nonintervention groups (randomized health care design). In such case, randomization is not performed at the level of the individual but on the level of region or birth cohort (clustered randomized clinical trial). With this strategy, very large research funds are not required and results apply to a real health care network and not merely to the research setting. This approach has been successfully applied in Finland to evaluate the new cervical cancer screening strategies [38].

In population-based randomized clinical trials in five European Union (EU) member states (Finland, Italy, Sweden, The Netherlands, and the United Kingdom), cytology screening is currently being compared with primary HPV screening or combined cytology/HPV screening. In all arms of these trials, the incidence of CIN2 and CIN3 in screen negatives can be assessed based on detection rates in 3−5 years after initial screening. The outcomes of these European trials, observed in the second screening round, supplemented by mathematical modeling, and taking into account costs, psychosocial aspects, and women's preferences, are believed to be pivotal for defining the final future screening policy in the EU [39].

2.6. Step 6: Impact on the population level and cost-effectiveness

> Do health outcomes in the whole society improve by introducing the new screening strategy at an acceptable cost?

Often, the ''real-world dimension'' differs from that of a ''trial dimension,'' so cost-effectiveness analyses, econometrical analyses, and models of the different modalities of application or triage/management can provide us with a more comprehensive and broader picture. Our final goal is the reduction of the burden of cancer on the total population. Any type of burden should be outweighed against benefit to improve patient outcome. This step goes beyond individual harms and benefits but also addresses the acceptability for the society in terms of costs. Table 3 presents an overview of the cost components attributed to screening.

A decrease in either specificity of a test or prevalence of disease can increase the costs rapidly. Loss in specificity of a screen test can be limited by raising the screening interval, increasing the age at onset of screening, and raising the cutoff for test positivity. However, all such changes will be assessed against the possible losses in sensitivity.

Cost-effectiveness of a candidate-screening strategy should preferentially be assessed by means of an econometrical analysis based on observed results of a well-organized RCT. Reliable mathematical models can be used to estimate the final outcome per unit of cost. They should rely on accurate and complete estimates of performance including measures of imprecision and should only be made after a net benefit on patient health has been established. Mathematical models can also be used to explore the impact of multiple variables, such as changes in target population, screening frequency, compliance of the population, and management options that cannot all be included in randomized trials (sensitivity analyses). In a pioneering series of studies by van Ballegooijen [40−42], mathematical modeling has been used to assess which program designs are likely to be most cost-effective and identify critical areas of uncertainty in which research is particularly needed [39]. Major conclusions from these studies were that the longitudinal performance of the HPV screening strategy was critical for achieving cost-effectiveness because the high cost and lower specificity of HPV screening compared

Table 3. Cost components [12]

• Cost price of the screening strategy (including overscreening), fees, and administrative and logistical costs
• Cost for follow-up and treatment of false positives
• Cost for follow-up and treatment of true positives
• Human costs (time spent, anxiety and discomfort, and side effects)
• Consequences of delay in the detection of cancer
• Need for repeat tests

with cytology-based screening needs to be compensated for by a longer screening interval to become cost-effective.

Prevalence of disease, availability of resources in terms of staff and technology, expectations and acceptability by the population, and financial costs of an intervention largely differ between countries and regions. Therefore, the results of population-based evaluations in one region can therefore never be translated as such to a different region, but studies (depending on which aspects are different) will have to be repeated [2].

3. Conclusion

Screening strategies are often introduced into clinical practice without proper evaluation. We have proposed a stepwise framework to evaluate new screening strategies, in which an increasing level of evidence is gathered but require progressively more stringent and expensive studies. Regional, national, or international authorities should impose requirements for testing and implementing new screening strategies, as currently has been done in a number of countries. Together with other recently published criteria, our framework may be helpful in defining such requirements, which should lead toward a net benefit in general health of the individual at an acceptable cost for society.

Acknowledgments

All authors contributed substantially to conception and design, revised the article critically for important intellectual content, and gave final approval of the version to be published.

Special thanks should be given to the the Belgian Foundation Against Cancer (Brussels, Belgium) for providing the financial means.

M.A. received financial support from the Belgian Foundation Against Cancer (Brussels, Belgium), the International Agency for Research on Cancer (IARC, Lyon, France), the 7th Framework Programme of DG Research of the European Commission through the PREHDICT project (grant no. 242061, coordinated by the Vrije Universiteit Amsterdam, the Netherlands), and the Gynaecological Cancer Cochrane Review Collaboration (Bath, UK).

References

[1] Wilson J, Jungner G. Principles and practice of screening for disease. WHO public health pap. 1968;34. Available at: http://www.who.int/bulletin/volumes/86/4/07-050112bp.pdf.

[2] Dans LF, Silvestre MA, Dans AL. Trade-off between benefit and harm is crucial in health screening recommendations. Part I: general principles. J Clin Epidemiol 2011;64:231−9.

[3] Harris R, Sawaya GF, Moyer VA, Calonge N. Reconsidering the criteria for evaluating proposed screening programs: reflections from 4 current and former members of the U.S. Preventive Services Task Force. Epidemiol Rev 2011;33:20−35.

[4] Silvestre MA, Dans LF, Dans AL. Trade-off between benefit and harm is crucial in health screening recommendations. Part II: evidence summaries. J Clin Epidemiol 2011;64:240−9.

[5] Health Council of the Netherlands. Population screening for cervical cancer. The Hague: Health Council of the Netherlands; 2011.

[6] Ronco G, Giorgi-Rossi P, Carozzi F, Confortini M, Dalla Palma P, Del Mistro A, et al. Efficacy of human papillomavirus testing for the detection of invasive cervical cancers and cervical intraepithelial neoplasia: a randomised controlled trial. Lancet Oncol 2010;11:249−57.

[7] Cervix Cancer Screening/IARC working group on the evaluation of cancer-preventive strategies. Lyon: IARCPress; 2004.

[8] Arbyn M, Castellsaqué X, de Sanjosé S, Bruni L, Saraiya M, Bray F, et al. Worldwide burden of cervical cancer in 2008. Ann Oncol 2011;22:2675−86.

[9] Van Oortmarssen GJ, Habbema JD. Epidemiological evidence for age-dependent regression of pre-invasive cervical cancer. Br J Cancer 1991;64:559−65.

[10] Hakama M. Implications of screening on the biology of cervical cancer. Nowotwory 1986;36:1−5.

[11] Zwahlen M, Low N, Borisch B, Egger M, Künzli N, Obrist R, et al. Population-based screening—the difficulty of how to do more good than harm and how to achieve it. Swiss Med Wkly 2010;140:1−11.

[12] Koliopoulos G, Arbyn M, Martin-Hirsch P, Kyrgiou M, Prendiville W, Paraskevaidis E. Diagnostic accuracy of human papillomavirus testing in primary cervical screening: a systematic review and meta-analysis of non-randomized studies. Gynecol Oncol 2007;104:232−46.

[13] Arbyn M, Ronco G, Cuzick J, Wentzensen N, Castle PE. How to evaluate emerging technologies in cervical cancer screening? Int J Cancer 2009;125:2489−96.

[14] Gøtzsche P, Nielsen M. Screening for breast cancer with mammography. Cochrane Database Syst Rev 2011;19:CD001877.

[15] Arbyn M, Anttila A, Jordan J, Ronco G, Schenck U, Segnan N, et al. European guidelines for quality sssurance in cervical cancer screening. Second edition—summary document. Ann Oncol 2010;21:448−58.

[16] Holowaty P, Miller AB, Rohan T, To T. Natural history of dysplasia of the uterine cervix. J Natl Cancer Inst 1999;91:252−8.

[17] Kyrgiou M, Koliopoulos G, Martin-Hirsch P, Arbyn M, Prendiville W, Paraskevaidis E. Obstetric outcomes after conservative treatment for intraepithelial or early invasive cervical lesions: systematic review and meta-analysis. Lancet 2006;367:489−98.

[18] Arbyn M, Kyrgiou M, Simoens C, Raifu AO, Koliopoulos G, Martin-Hirsch P, et al. Perinatal mortality and other severe adverse pregnancy outcomes associated with treatment of cervical intraepithelial neoplasia: meta-analysis. BMJ 2008;337(sep18 1):a1284-[a].

[19] Arbyn M, Ronco G, Meijer CJLM, Naucle P. Trials comparing cytology with human papillomavirus screening. Lancet Oncol 2009;10:935−6.

[20] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy—the STARD initiative. BMJ 2003;326:41−4.

[21] 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.

[22] Pretorius RG, Zhang WH, Belinson JL, Huang MN, Wu LY, Zhang X, et al. 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.

[23] Jeronimo J, Schiffman M. Colposcopy at a crossroads. Am J Obstet Gynecol 2006;195:349−53.

[24] Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and

diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J Clin Epidemiol 2007;60:1116−22.

[25] Haynes RB, You JJ. The architecture of diagnostic research. In: Knottnerus JA, Buntinx F, editors. The evidence base of clinical diagnosis: theory and methods of diagnostic research. 2th ed. West Sussex, UK: John Wiley & Sons Ltd.; BMJ Books; 2009. pp. 20−41.

[26] Lijmer JG, Bossuyt PM. Diagnostic testing and prognosis: the randomized controlled trail in test evaluation research. In: JA Knottnerus FB, editor. The evidence base of clinical diagnosis: theory and methods of diagnostic research. 2th ed. BMJ Books; 2009:63−83.

[27] 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.

[28] Arbyn M, Van Veen E, Andersson K, Bogers J, Boulet G, Bergeron C, et al. Cervical cytology biobanking in Europe. Int J Biol Markers 2010;25:117−25.

[29] Schünemann H, Oxman A, Brozek J, Glasziou P, Jaeschke R, Vist G, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106−10.

[30] 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.

[31] 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.

[32] Gaffikin L, McGrath J, Arbyn M, Blumenthal PD. Avoiding verification bias in screening test evaluation in resource poor settings: a case study from Zimbabwe. Clin Trials 2008;5:496−503.

[33] Buntinx F, Brouwers M. Relation between sampling device and detection of abnormality in cervical smears: a meta-analysis of randomised and quasi-randomised studies. BMJ 1996;313:1285−90.

[34] Martin-Hirsch P, Lilford R, Jarvis G, Kitchener HC. Efficacy of cervical-smear collection devices: a systematic review and meta-analysis. The Lancet 1999;354:1763−70.

[35] Arbyn M, Bergeron C, Klinkhamer P, Martin-Hirsch P, Siebers A, Bulten J. Liquid compared with conventional cervical cytology: a systematic review and meta-analysis. Obstet Gynecol 2008;111:167−77.

[36] Siebers A, Klinkhamer P, Grefte J, Massuger L, Vedder J, Beijers-Broos A, et al. Comparison of liquid-based cytology with conventional cytology for detection of cervical cancer precursors: a randomized controlled trial. JAMA 2009;302:1757−64.

[37] Davies P, Arbyn M, Dillner J, Kitchener HC, Meijer CJ, Ronco G, et al. 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. [Review].

[38] Anttila A, Hakama M, Kotaniemi-Talonen L, Nieminen P. Alternative technologies in cervical cancer screening: a randomised evaluation trial. BMC Public Health 2006;6:252.

[39] Arbyn M, Anttila A, Jordan J, Ronco G, Schenck U, Segnan N, et al. European guidelines for quality assurance in cervical cancer screening. Second edition - Summary Document. Ann Oncol 2010;21:448−58.

[40] van Ballegooijen M, van den Akker-van Marle ME, Patnick J, Lynge E, Arbyn M, Anttila A, et al. Overview of important cervical cancer screening process values in European Union (EU) countries, and tentative predictions of the corresponding effectiveness and cost-effectiveness. Eur J Cancer 2000;36:2177−88.

[41] van Ballegooijen M, van den Akker-van Marie ME, Warmerdam PG, Meijer CJLM, Walboomers JMM, Habbema JD. Present evidence on the value of HPV testing for cervical cancer screening: a model-based exploration of the (cost-)effectiveness. Br J Cancer 1997; 76:651−7.

[42] van den Akker-van Marle ME, van Ballegooijen M, Habbema JD. Low risk of cervical cancer during a long period after negative screening in the Netherlands. Br J Cancer 2003;88:1054−7.