DOI: 10.1093/annonc/mdf658

Cancer screening for medical oncologists: definitions and aims

M. Hakama

University of Tampere School of Public Health, Tampere, Finland

Definition of screening

Screening is the identification of preclinical disease by a relatively simple test. It is usually regarded as public health policy in that it is applied to populations. It is not always possible to distinguish those with symptoms and those without, as many screening policies are based, for example, on invitation of the individuals in the total target population. In fact, the aim is to identify disease not recognized by the health services and the term preclinical refers rather to such unrecognized disease than to clinical detectability or recognition. Also, in practice the distinction between screening and case finding is sometimes difficult to make. For example, detection of preclinical cancer of cervix uteri may occur either within normal clinical practice or at mass screening.

The objective of screening for cancer is to reduce mortality and to improve the quality of life. Screening is usually employed to detect serious chronic diseases such as cancer, for which reduction in mortality is the primary objective. In fact, many basic aspects of screening in general can be illustrated by the experience with cancer.

There are several prerequisites for successful screening. The most well-known list was published by Wilson and Jungner [1].

- The condition sought should be an important health problem.
- There should be an accepted treatment for patients with the disease.
- Facilities for treatment and diagnosis should be available.
- There should be a recognizable latent or early symptomatic stage.
- There should be a suitable test or examination.
- The test should be acceptable to the population.
- The natural history of the condition should be adequately understood.
- There should be an agreed policy on whom to treat as patients.
- The cost should be economically balanced in relation to possible expenditure on medical care as a whole.
- Case-finding should be a continuing process and not a once and for all project.

The observation that early diagnosis reduces mortality through improved prognosis is the impetus for screening. This is not always true, however, which makes the understanding of screening and its evaluation particularly important.

Validity of screening

Early indicators of the effect, called process measures, give a first but not sufficient indication of the potential effectiveness. Describing the different aspects of the process of screening is difficult and there is no single axis on which various comparisons could be made. Validity of screening is a summary measure on the success of the screening process and it is given by two indicators: sensitivity and specificity. Sensitivity is an indicator of the extent of preclinical disease that is identified by the screening process and specificity describes to what extent healthy individuals are identified in the population subjected to screening.

Validity of the test

Validity is usually limited to the screening test only. Sensitivity is estimated as the proportion of persons with a positive test among those with the disease. Specificity is the proportion of persons with a negative test among those free from the disease (Table 1). Therefore sensitivity is the basic measure of the success of screening and it indicates the yield. Specificity is a basic measure of the disadvantages of screening: poor specificity of the test results in higher financial costs and in adverse effects due to false positive tests.

Sensitivity and specificity are inversely related for the same test e.g. when sensitivity is improved specificity goes down. It is to be noted that the components of validity have different and incomparable implications. Sensitivity is mainly related to the objective and specificity with the harm, and these

Table 1. Validity of screening

Screening	Disease	
	Present	Absent
Positive	*a*	*b*
Negative	*c*	*d*

Sensitivity = $a/(a + c)$.
Specificity = $d/(b + d)$.

aspects cannot be regarded as being of equal importance. Nor does there exist any objective combination of sensitivity and specificity. Selecting a particular combination of validity always involves value judgements.

This does not mean that any combination of specificity and sensitivity would be acceptable for any disease to be screened. There are some general aspects to be considered.

The relationship between specificity and sensitivity depends on the test and on the disease to be screened. Pap smear testing is aimed at a high sensitivity and some false positives are regarded as acceptable. This is because the yield is regarded as being of primary importance and confirmation of diagnosis is regarded (this is not necessarily true) as relatively reliable, non-invasive and inexpensive. Another extreme is provided by a rather hypothetical screening for selected mental problems. Mental diseases are commonly valued such that the society should be tolerant and allow some mental abnormalities to be regarded within normal limits. A false positive diagnosis of mental disease with its severe consequences is regarded as a more severe form of error than a false negative diagnosis.

The balance between specificity and sensitivity clearly depends on the values, both humanitarian and financial, attached to false negative and false positive screening tests.

Validity of the program

Sensitivity and specificity are measures of the success of the screening test in meeting its purpose. Performance of the test is not, however, a sufficient condition for success of a screening. Screening is a public health policy and it is the validity of the total program which is a more relevant aspect of the process of screening.

A screening program based on a valid test may still fail in its objective i.e. in reducing the mortality in the screening population.

The success of screening depends on the performance of the test, on the ability of the program as a whole to identify the disease at preclinical phase in the target population and on the ability of the program to improve the average prognosis of the disease in the target population. The first success measure was termed test validity, the second can be termed program validity [2] and the third could be termed as outcome validity. However, the more traditional term for the outcome validity is effectiveness.

A high level of program validity assumes in addition to a valid test that

- the individuals in the target population are identified and will attend the program;
- there are adequate field facilities to take the test in addition to an adequate screening center for assessing the screening test;
- there are adequate facilities for diagnostic confirmation and
- there is a carefully designed and agreed referral system.

Table 2. Predictive values of a screening test

Screening	Disease	
	Present	Absent
Positive	*a*	*b*
Negative	*c*	*d*

Positive predictive value = $a/(a + b)$.
Negative predictive value = $d/(c + d)$.

Table 3. Effect of preclinical prevalence on the predictive value (assuming specificity = sensitivity = 95%)

Prevalence (%)	Predictive value (%)	
	Positive	Negative
10	68	99.4
1	16	99.9
0.1	2	99.99

Program validity indicates to what extent the disease stemming from the target population will be identified before it will surface clinically or be otherwise recognized by a clinical facility.

Predictive values of a screening test

It is important to know the implications, especially for the screenee, of the result of a screening test. These can be described by the predictive values of a test (Table 2). A positive predictive value is the proportion of those with the preclinical disease among those with a positive test. A negative predictive value is the proportion of those who are healthy among those with a negative test. The predictive values depend on the validity of the test and on the prevalence of the preclinical disease (Table 3). High predictive values assume a valid test, but, particularly for a rare disease, a positive screening test does not necessarily imply the presence of the disease. In contrast, a negative test gives a very high likelihood of absence of the disease if the prevalence is low. Many of those who attend a screening program are in fact seeking reassurance that they do not have the disease. In practice, this is the greatest service of the screening test for rare diseases, rather than the indication of disease in the case of a positive test.

Estimation of process indicators is subjected to such deficiencies as lead time, length bias and overdiagnosis bias. There is substantial variation in the methods. Some comparability and ability to generalize is provided by sensitivity estimated by the interval cancer method, where difference between interval cancer incidence between two screening rounds and incidence in the absence of screening indicates sensitivity.

Table 4. Mayo Lung Project [3]

	Screened arm	Control arm
Incidence (per 1000 pyrs)	5.3	4.1
Resectability (%)	49	31
Survival at 5 years (%)	35	15
Lung cancer mortality (per 1000 pyrs)	2.9	2.8
Total mortality (per 1000 pyrs)	23.5	23.1

Pyrs, person-years.

Table 5. Health Insurance Plan (HIP) study: deaths from breast cancer and person years lost (PYLL) during 18 years of follow-up among patients with breast cancers diagnosed within the first 10 years of the study [5]

Randomized group	Incident cases	Deaths	PYLL
Study	623	260	2260
Control	632	305	2815

Evaluation of screening

The effect of any intervention can be evaluated by detection rates, sensitivity and specificity, referred to as process measures, which are necessary for effective screening. However, they are not sufficient; detection of lesions classified as malignant does not guarantee an effect on mortality and a valid test may result in an ineffective program. The evaluation should, with a few exceptions, be related to the objective of screening. As the main objective of screening for cancer is to reduce the mortality from cancer, the evaluation of screening takes place by means of a potential decrease in mortality from the cancer subjected to screening.

Screening for lung cancer provides a well-known demonstration of the insufficiency of the process measures. For example, the Mayo Lung Project [3] included approximately 10 000 men who were randomly allocated to the screening arm, with 4 monthly sputum cytology and radiology, or to the control arm. The process measures, or intermediate indicators, showed a favorable effect of screening, whereas there was no mortality difference between the two arms (Table 4).

The confusing results have been regarded as evidence against screening for lung cancer. However, there is still discussion [4] on the issue which demonstrates an incomplete understanding of the theory of screening.

Randomized preventive trials

The randomized preventive trial is the choice of preference for evaluation of the effectiveness of screening. Screening for breast cancer and screening for colorectal cancer were shown to be effective by a randomized experiment. Screening for cervical cancer was never subjected to a randomized trial as the test precedes the time when the scientific rules for evaluation were developed.

The Health Insurance Plan (HIP) study on the screening for breast cancer [5] is probably one of the best known screening experiments. It started in the early 1960s with approximately 60 000 women randomly allocated to the experimental (screening) group and the control group. The test consisted of two-view mammography, clinical examination and interview. Both groups were followed-up for deaths and the 18 year results have been reported.

Within a 10 year follow-up, 623 cases of breast cancer were diagnosed in the experimental group and 632 among the controls. During the 18 years of follow-up, 260 women (41%) among the 623 experimental cases and 305 (48%) of the 632 controls died from breast cancer. The difference equates to about 500 person-years in favor of the study group (Table 5). This is convincing evidence for the overall effectiveness of screening for breast cancer.

Only two-thirds of those allocated to the screening group actually participated. It is likely that if attenders only were considered in the analysis, the contrast in the mortality experience would have been greater between the two groups. This would not have been an acceptable analysis, however. The original groups randomized for screening and control are comparable, whereas the attenders are comparable with neither the controls nor the non-attenders (Table 6). The non-attenders had much higher death rates than the attenders. To consider only those attending would break the design. The comparisons would be analogous to those based on a non-experimental cohort study with the same problems and limitations in interpretation.

The HIP study was designed for a research setting. Demonstration projects and non-experimental designs have been proposed for evaluating a screening program that is to be applied as public health policy. However, there are ways of achieving randomized implementation of a screening program at the population level. The purpose of this is to avoid the general limitations of non-experimental design. An example of this

Table 6. Health Insurance Plan (HIP) study: mortality rate (per 10 000 woman-years) from causes other than breast cancer by screening status [5]

Cause of death	Mortality rate per 10 000 (woman-years)			
	Screened	Refused	Total	Control
All causes except breast cancer	56.8	92.3	68.4	68.7
All cancers except breast cancer	18.9	24.4	20.7	20.3

approach is the implementation of mammographic screening as a public health policy in Finland. [6]

The screening program as a public health policy can be planned in an experimental setting when the test is first applied and evidence on the success of the effectiveness of the service is inconclusive. The evaluation assumes a control population that was not offered the screening. When the program is introduced, resources are not available to cover the total target population immediately and at the same time the limited resources available are used to apply the screening test to a randomly allocated sample of the population, not a self-selected or haphazardly selected fraction of the population. As long as the program only covers a small proportion of the population, it is ethically acceptable to carry out a randomized trial because the service is withheld from nobody and the trial gives an *a priori* equal chance to those in the target population of benefiting from the program and of avoiding any adverse effects of the program. For those who will be subjected to the public health services in the future, it will provide the most reliable basis for accepting new activities within the services.

References

1. Wilson J, Jungner G. Principles and practice of screening for disease. World Health Organization Public Health Paper 34. Geneva: WHO 1968.
2. Hakama M. Screening. In Holland WW, Detels R, Knox EG (eds): Oxford Textbook of Public Health, Vol. 3. Oxford: Oxford University Press 1991; 93–106.
3. Fontana RS. Screening for lung cancer. In Miller AB (ed.): Screening for Cancer. Orlando, FL: Academic Press 1985; 377.
4. Strauss GM. The Mayo lung cohort: A regression analysis focusing on lung cancer incidence and mortality. J Clin Oncol 2002; 20: 1973–1983.
5. Shapiro S, Venet V, Strax P, Venet L. Current results of the breast cancer screening randomized trial: The Health Insurance Plan of Greater New York Study. In Day NE, Miller AB (eds): Screening for Breast Cancer. Toronto: Hans Huber 1988; 3.
6. Hakama M, Pukkala E, Heikkilä M, Kallio M. Effectiveness of the public health policy for breast cancer screening in Finland: population-based cohort study. Br Med J 1997; 314: 864–867.