Could or should, ought or nought?

Ethical, methodological and technical considerations of cancer screening

Jan F. Nygård and Hege Wang

The Cancer Registry of Norway

NORSK SAMMENDRAG

Denne artikkelen identifiserer tre aspekter som bør vurderes før man innfører et masseundersøkelsesprogram mot kreft. Disse aspektene er av etisk, metodologisk og teknisk art. Den etiske balansegangen mellom positive og negative effekter er helt og holdent av subjektiv karakter, og det er ingen grunn til å anta at medisinske eksperter noensinne vil komme til enighet på dette punktet. Med hensyn til de metodologiske og tekniske aspekter skulle, og burde, det være mulig å bli enige om fremgangsmåter for å etablere medisinsk kunnskap ved korrekt applisering av anerkjente vitenskapelige prinsipper. De medisinske kriteriene som ligger til grunn før man iverksetter en masseundersøkelse er det generell enighet om. Studier som vurderer masseundersøkelsesprogram bør ta hensyn til de disse kriteriene, og estimere effekten i to trinn. Først må det undersøkes i hvilken grad sykdommen blir oppdaget i et tidligere stadium fordi man tester for usymptomatisk sykdom, som dermed fører til en stadiemigrasjon. Deretter må det undersøkes om denne stadiemigrasjonen medfører en redusert dødelighet av sykdommen. Dette medfører at vurdering av effekt av masseundersøkelser fortrinnsvis gjøres med deskriptive studier. Studier som vurderer effekt ved å sammenligne ulike grupper (med og uten intervensjon), som f.eks. en randomisert kontrollert studie, er mindre egnet til å vurdere effekt av masseundersøkelser.

INTRODUCTION

> *Those who suffer from hidden cancer it is better not to treat. For should they be treated, they will rapidly die, but should they remain untreated they may continue to live for a long time.*
>
> *Hippocrates*

In the statement above Hippocrates compares a group that is treated with a group that is left untreated, and follows both groups of persons until they die. His deduction seems valid, but is his evidence of a kind that justifies the conclusion? Some 2500 years later, different cancer screening programs have been implemented to detect and treat "hidden", i.e. unsymptomatic cancer. It seems that what screening constitutes, the prerequisites and the goals are neither controversial nor subject to much dispute. However, implementation of screening programs remains controversial and the effects are often disputed.

A closer look reveals that at least two different questions are often asked: "should we do it?" and "could we do it?" The former question opens a range of ethical arguments, the latter a blend of methodological and technical arguments. The ethical ("should we do it?") disagreements concern both how to assess the value of the benefit of screening, and the question of transformation of facts to norms: what measures, if any, should be implemented when the size of effect of screening is known. The methodological question ("could we do it?") concerns the validity of certain types of evidence (experimental or observational) and its transformation into knowledge. The technical question ("could we do it?") concerns the feasibility of achieving the wanted effect in different settings.

The arguments and contra-arguments concerning the effects of screening, or whether a screening program should be implemented, invariably interweave all ethical, methodological and technical aspects.

In an attempt to untangle and clarify the seemingly different positions in this debate, this article examines three different aspects of cancer screening. These aspects are:

- the role of prevention in medicine (ethical considerations).
- the design of studies to show benefit of preventive measures (methodological considerations).
- limitations of the screening test and/or the screening program (technical considerations).

DEFINITIONS AND CRITERIA

The following definitions and criteria are used in the article:

Screening

The pursuit of diagnosis in the absence of any known manifestation of the illness, with a view to early intervention, is termed *screening* for the illness (1).

Screening Program

A screening program is the application of screening to a well-defined population. The program is divided into three distinct steps: 1) The screening-test, a test given to a single individual without symptoms. Based on the outcome of this test the person will be classified as either probably healthy or diseased. The probably healthy persons are left untampered, and if the protocol specifies it, they will be invited to another screening-test after a predetermined interval. 2) Diagnostic work-up of the screen positives. This could mean a repeated test, but usually several and more extensive and expensive tests. Based on the results of the diagnostic work-up some will be considered as having the disease in question and given a diagnosis. 3) Treatment or observation of diagnosed patients.

Preventive medicine

Preventive medicine is often said to operate at three different levels, customarily called primary, secondary and tertiary (2). Primary prevention aims to reduce or eliminate causal factors. Secondary prevention emphasises early detection and treatment of the disease. Tertiary prevention improves treatment or postpones death from the disease. Screening for a disease is, clearly, secondary prevention.

Criteria for implementation of a screening program

WHO has established medical criteria that should be fulfilled before screening may commence (3). These criteria have later been modified (4-6) and can be summed up as follows:

I. The disease should be common and cause serious morbidity or death.
II. The natural history, from inception of the disease to a stage where death or serious morbidity no longer can be prevented, should be known. The disease should have a long asymptomatic, preclinical stage, which can be detected and diagnosed.
III. The test must classify the person as likely healthy or likely diseased, with a reasonable degree of accuracy in the asymptomatic stage, i.e. have high sensitivity and specificity.
IV. An effective treatment for the disease at an early stage should be available.

Screening requires organisational skills in addition to medical knowledge and whether these criteria are met often depends on technical considerations. The organisational prerequisites for a successful screening program dictate that (4):

a. the target population, and the individuals within it, are identified
b. measures to guarantee high coverage are available, such as a letter of invitation
c. there are adequate field facilities for collecting the screen material and adequate laboratory facilities to examine it
d. adequate facilities exist for diagnosis and for appropriate treatment of confirmed disease, and for the follow-up of treated individuals
e. there is a carefully designed and agreed upon referral system, with organised quality control of all procedures and treatments

ETHICAL CONSIDERATIONS

The supreme ethical consideration in cancer screening is to decide if one *should* implement preventive measures. To some it is a subjective and normative endeavour, often *a priori* given, e.g. any available intervention should be tried. To others it is the deliberation of evidence; the effect of the intervention should justify its implementation.

Any wise debate of the ethical considerations necessitates illuminations of both the positive and negative effects of intervention. It is rare for the effects to be solely positive, and weighing the beneficial effects against the adverse therefore becomes necessary. Should the decision makers chose to intervene, it follows that they accept that some persons will experience adverse effects, so that other may benefit. The decision-makers have thus agreed upon an "acceptable" level of adverse effect, making a subjective quantitative ethical decision. Even on an individual level one could experience both negative and positive effects, e.g. a woman could be cured of breast cancer, but lose her breast.

A deeply held human feeling is that it may be better to live a day at a time rather than to be anxious about distant problems that may never materialise. People are generally motivated only by the prospects of a benefit that is visible, early, and likely. Health benefits rarely meet these criteria (7), and clinical medicine accordingly emphasises diagnosis and treatment. Prevention demands knowledge of etiology and is of no immediate clinical value. Prevention is thus left to the more esoteric discipline of epidemiology (or more correctly, the subject matter of public health). Clinicians often act as though their professional responsibility is limited to the sick and the nearly sick (those at imminent risk) (3). The mantra is that treatment should be given to *one* patient who should benefit from that specific treatment. In this view, measures should always be evaluated at the individual level and be effective at the individual level. Preventive measures would rarely meet these expectations. The difference between public health benefit and individual benefit of a preventive measure gives rise to the preventive paradox: a preventive measure may bring large benefits to the community, but offer little to most participating individuals (7). Clinical medicine requires a classification of individuals as either diseased (patients) *or*

healthy (persons). Preventive medicine deviates from clinical medicine by introducing a third category: people without symptoms, but with a disease. Many of the disagreements of ethical concerns stem from the formation of this third category. The prerequisites of secondary prevention demand a concept of disease as a continuum of severity, or at least an ordinal categorisation (medical criteria II). This is in contrast with the dichotomous view of clinical medicine.

Those with a positive inclination to screening express the view that screening is a purposeful extension of the process by which people perceive symptoms of their own illnesses and then consult physicians for diagnosis and treatment. Screening simply calls attention to the likelihood of the disease before symptoms appears (8). They would argue that medical care already includes considerable screening: blood pressure, intraocular pressure and urine sugar are often measured in asymptomatic persons. Moreover, popular pressure often demands extensions of screening for a diversity of conditions. The latest example concerns the use of depleted uranium in Kosovo and a possible association with leukaemia, which has led to a demand for a health check (screening) of all military personnel. The concept of benefit of early intervention is also appreciated outside medicine and is often done, e.g. car maintenance.

When a test exists, the consequences of either prohibiting or providing its use might prove significant. It can be argued that public knowledge (or belief) that a test exists but is not made available will create unnecessary anxiety. When the general public became aware of the effects of mammography screening, the use of spontaneous mammography examinations increased twenty fold in Norway (9). An organised program was implemented several years later (10).

Ongoing screening and level of medical awareness in the community may be related, as people are likely to be more sensitive to their own symptoms in areas where screening is done and information is widespread (8). Preventive effects might be achieved by early detection of symptoms simply through public education, as was the case with cervical cancer in the first half of the twentieth century in Sweden (11). However, people can only be anxious about diseases of which they have some knowledge. Following this line of reasoning, the best way of preventing anxiety would be to keep medical knowledge secret and to consider providing information as unethical.

Another concern of those opposed to screening is persons who receive a negative screening result, who may subsequently feel in some way protected against the disease in question. However, there is no preventive effect of having a test. The test provides a retrospective reassurance, but cannot protect against the development of detectable disease in the future. Repeating the test at regular intervals is therefore often recommended. Screening has a particular responsibility to convey information correctly. There is a monumental difference in saying, "we did not find a tumour", and "you do not have a tumour". One should bear in mind that medicine is about probabilities; a ninety-nine percent chance of not having a disease *does* mean that there is a one percent of having it.

There is always the risk of error or mistake when something is done. There are two main possibilities of making an error when applying a test (dichotomous), a false positive or a false negative result. Public knowledge of uncertain test results might lead to lower compliance, which again reduces the effect further. A person with a false positive test is given a result indicating that they, falsely, have the disease in question. These false results are created by the screening and would not have occurred in its absence. The magnitude of this error is determined by the level of quality control in the second step in the screening programme, the diagnostic work-up. The false positive error leads to unnecessary tests, examinations and treatments, all with their own corresponding risks, and to the anxiety inherent in receiving a positive result. These errors must then be considered and weighed against the benefit given to those with a true positive test and the reassurance to those with a true negative test (medical prerequisite III).

A person with a false negative test is given a result indicating that they falsely do not have the disease. A false negative test leads to a later diagnosis than if the test result had been correct. There is reason to be concerned if these diagnoses occur later than they would have without a screening program, i.e. would have been clinically detected. Another concern is that a high rate of false negative tests will lead to a diminished effect of the screening program. In addition, persons experiencing false negative tests are probably less likely to comply with the program in the future.

Another type of error is that of overdiagnosis, with its subsequent overtreatment. In cancer screening this occurs when a lesion is detected and labelled as clinically important, when it has an extremely long latent phase – so long that the patient will die from another cause before the cancer becomes clinically evident. However, to determine if this particular person *would* have benefited from early detection or not, is only possible after his or her death, which is somewhat late. Overdiagnosis is related to the prerequisites of screening (medical criteria II), and is a consequence of insufficient knowledge of the natural history of the disease and of its prognostic factors. Fast-growing tumours are difficult to intercept before they reach a fatal state, in contrast to slow-growing tumours.

METHODOLOGICAL CONSIDERATIONS

Assessment of prevention constitutes of two distinct steps: the first is to establish a cause-effect relationship, and the second is to evaluate the efficacy of the

particular preventive activity. The rationale for establishing causality is that a change in the determinant state, say from unscreened to screened, gives a change in health state, say reduced mortality rate. This establishes the basis for intervention, both preventive and therapeutic. The cause-effect relationship should be abstract, so that it could be transferred to another place and time (repeatability) i.e. it should be based on scientific principles for causal assessment. The evaluation of efficacy involves extra-scientific elements that are place and time specific (technical and economical considerations), and also involves considerations of values (ethical) (12).

Study design

There are two ways of providing evidence of a causal relationship in medical research. The two modalities are determinant centred (prospective) studies and outcome centred (retrospective) studies. Results from both prospective and retrospective studies can be used in preventive medicine by identifying the casual determinants (primary prevention) or influencing the natural course of the disease to stop progression (secondary and tertiary prevention).

Prospective studies

The definitional characteristic of a prospective study is that it is determinant centred, i.e. the determinants are firstly assessed. The investigators then have to wait for an outcome that might or might not occur. Prospective studies may be purely descriptive (e.g. describe the natural course of a disease) or experimental (for assessment of the effect of interventions). When the interest is in the effect of an intervention, there is a need to compare. By proper use of selection and randomisation, two study groups that probably would have equal incidences of the disease or mortality in the absence of the assigned intervention are created. One of the groups receives the new treatment (intervention), while the other is given a placebo treatment (control). The two groups are then followed prospectively and outcomes in both groups recorded. The measure of effect would then be the difference in the rates of a given outcome between the two groups. If randomisation is used for assigning persons to treatment and control, respectively, the term for this study is Randomised Controlled Trial (RCT). This type of study has been referred to as the golden standard of medical research, due to the success of experimental approaches in other scientific fields.

However, RCTs have not been proven to be the holy grail of medical research. The main limitation of the RCT design, when used in preventive medicine, is that the appropriateness of a RCT is reciprocal to the time from randomisation to outcome. To evaluate the effect of a cancer screening trial, several years have to pass before the study is finalised. This is in stark contrast to ordinary clinical trials. A RCT over several years will be marred by non-adherence in the intervention group and contamination of the control group. This will result in a bias favouring no effect of screening. Invalid inferences have unfortunately ensued from this, e.g. dismissing a probable effect of lung cancer screening (13).

Retrospective studies

The effect of screening can also be estimated by the use of retrospective studies. The definitional characteristic of a retrospective study is that the starting point is the outcome, the occurrence of the disease of interest. Then the investigators look backwards in time, and take into consideration causal relationship characteristics as latency, to establish determinants that could have caused this particular outcome.

Such a design would provide a direct estimate of the impact of screening activities in much the same way that a randomised study does (8). Selection bias might arise with this study design, but with proper data collection (longitudinal data on both outcome and determinants), known confounders could be documented and statistically controlled. However, unknown and unobserved effect modifiers might be a source of error, but this should be considered in light of medical prerequisite II.

Non-comparative/descriptive studies

Another option for assessing the effect of cancer screening is to provide evidence of stage migration. The criterion of knowledge of natural history (medical prerequisite II), with focus on growth of the lesion, is free from any demand of comparability with a control group. To provide evidence of tumour growth, the comparison should be done within the same tumour at different times. A new screening program can be evaluated by determining the distribution of disease stage (in baseline and repeat screenings, respectively), and comparing this distribution with the distribution without screening or with an earlier screening program (13). An improvement indicates a real stage migration, and not just a change in the stage distribution as a consequence of an increase in detection of non-malignant, slowly growing tumours. When growth has been documented, overdiagnosis is no longer a credible explanation for higher empirical rates of pre-cancerous cancers after the introduction of screening. The proper effect of any given screening program should be assessed by the extent to which the screening test advances diagnosis with respect to disease stage. It is this change in stage distribution that will lead to reduced cancer mortality rates.

TECHNICAL CONSIDERATIONS

Screening is, in its simplest form, the application of a test on a symptom-free person, but the term screening is also used for the organised testing of an entire

population. It is necessary to assess separately how well the screening test detects the disease (test validity) and how well the screening program reduces the incidence/mortality of the disease (program validity), irrespective of whether implementing a new program or improving an existing one.

Test validity

Medicine is not about maximising accuracy (the proportion of correct test results), as accuracy is highly dependent upon prevalence. Focus on accuracy would give a false sense of correctness when dealing with diseases with low prevalence, as is the case with most cancers. A test for precursors of cervical cancers would be 90% accurate if all tests were said to be normal without actually analysing the test (as the percentage of normal smears is approximately 90%). Medical prerequisite I demands that the disease be common, but without quantification this is not particularly useful.

The proper test characteristics to be considered are sensitivity and specificity (medical prerequisite III). These are *a priori* characteristics and should already have been estimated in a previous trial where the screening test was simultaneously taken with the "golden standard" test. Sensitivity and specificity are interdependent. Increasing sensitivity by adjusting the cut-off level for a positive test would decrease specificity, and vice versa. When a test is to be used for screening, it should be tuned for high specificity. The lower the prevalence the more impact of specificity. A low specificity results in bringing in a crowd of persons for diagnostic work-up (false positives), with major implications for the cost-effectiveness of the program. A low specificity could also scare people from repeated screening-tests, which in the long run would reduce the coverage. The backside of high specificity is lower sensitivity. The lower the sensitivity, the lesser the proportion of persons that will enjoy the advantages of early detection. However, even a test with 50% sensitivity would substantially reduce the mortality of interest, especially if repeated regularly, which is the case with cervical cancer (14). Sensitivity could be considered a measure of benefit, and specificity a measure of cost.

Program validity

Screening programs should not be considered in terms of sensitivity or specificity, but of false positives and negatives. These are *a posteriori* terms to be assessed during the run of the program. Sensitivity and specificity are theoretical issues, while false positive and negative test results are highly empirical and occur in identifiable persons. If the program is to succeed, the rates of false negatives and false positives must be minimised.

The organisational prerequisites must be fulfilled, i.e. the program must yield a high coverage and a high quality of follow-up of treatment. The program will always be less effective than the test. Usually the effectiveness of the program is termed efficacy, and is a function (the product) of the effectiveness of the test and the compliance of the program.

Should the cause of a cancer be environmental or unknown, repetitive measures such as several screenings are often necessary, as one must assume that tested subjects are at continuous risk. In addition, slowly growing cancers need repetitive testing, as the cancers might not be detectable at the first screening.

IN CONCLUSION

To summarise the three aspects of cancer screening, the asymmetry between rejection and acceptance leads one to reject implementing a screening program if only one aspect is unacceptable and accept only if all aspects are met. In Norway, organised screening programs against cervical and breast cancer have been implemented, while studies on the effect of colorectal cancer screening are carried out, and organised screening against prostate cancer has been rejected.

The ethical balancing of adverse effects against the intended is inherently a subjective matter, and there is no reason to believe that there will be an end to the debate on this issue. However, both the methodological and technological issues under debate could, and should, be resolved by rigorous application of knowledge and reason.

Under medical prerequisite II, studies should assess the extent to which a given regimen of screening advances the stage of diagnosis. Subsequently, the concern is the extent to which that stage migration prevents death (or untimely death) from the disease by treatment (1). This knowledge is, however, readily available from cancer registry statistics (15). From this it follows that screening is a topic of the time of diagnosis in reference to tumour growth, rather than of intervention. Stage migration is a descriptive issue. Comparative studies, like prospective (RCT) and retrospective studies, are therefore not suitable to assess the effects of a screening program.

The problem of inference, the transition from evidence to knowledge, seems as prevalent now as 2500 years ago. Hippocrates might have been replaced by evidence-based medicine, but similar dogmatic and unsound reasoning is still echoed in scientific articles in highly ranked journals (16).

REFERENCES

1. Miettinen OS. Screening for lung cancer. *Radiol Clin North Am* 2000; **38**: 479-86.
2. Last JM. *A Dictionary of Epidemiology*. New York: Oxford University Press, 1995.
3. Wilson JMG, Jugner G. Principles and practice of screening for disease. Public Health Papers 34. Genève: WHO, 1968.
4. Hakama M. Screening for cervical cancer: Experience of the Nordic countries. In: Franco E, Monsonego J, eds. *New Developments in Cervical Cancer Screening and Prevention*. Oxford: Blackwell Sciences, 1997.
5. Parkin DM. The epidemiological basis for evaluating screening policies. In: Franco E, Monsonego J, eds. *New Developments in Cervical Cancer Screening and Prevention*. Oxford: Blackwell Sciences, 1997.
6. Koss LG. Performance of cytology in screening for precursor lesions and early cancer of the uterine cervix. In: Franco E, Monsonego J, eds. *New Developments in Cervical Cancer Screening and Prevention*. Oxford: Blackwell Sciences, 1997.
7. Rose G. *The Strategy of Preventive Medicine*. Oxford: Oxford University Press, 1992.
8. Morrison AS. Screening. In: Rothman KJ, Greenland S, eds. *Modern Epidemiology*. Philadelphia: Lippincott-Raven, 1998.
9. Widmark A, Olsen JB. Mammography in Norway. Report of the Norwegian Radiation Protection Authority, 1995.
10. Wang H, Kårensen R, Hervik A, Thoresen SØ. Mammography screening in Norway. Results from the first screening round in four counties and cost-effectiveness of a modelled nationwide Screening. *Cancer Causes Control* 2001; **12**: 39-45.
11. Ponten J, Adami HO, Bergstrom R, et al. Strategies for global control of cervical cancer. *Int J Cancer* 1995; **60**: 1-26.
12. Vineis P. Evidence-based primary prevention? *Scand J Work Environ Health* 2000; **26**: 443-8.
13. Miettinen OS. Screening for lung cancer. Do we need randomized trials? *Cancer* 2000; **89**: 2449-52.
14. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. *Am J Epidemiol* 1995; **141**: 680-9.
15. Kreftregisteret. Cancer in Norway 1997. Oslo: The Cancer Registry of Norway, 2000.
16. Gøtzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000; **355**: 129-34.