HISTORY OF MEDICINE

History of medical screening: from concepts to action

A Morabia, F F Zhang

Postgrad Med J 2004;**80**:463–469. doi: 10.1136/pgmj.2004.018226

The objective of medical screening is to identify disease in its preclinical, and therefore hopefully still curable, phase. This may have been an old quest in medicine but it became historically possible when at least four conditions were met: the availability of simple, valid and acceptable forms of tests, the discovery of effective treatments, the establishment of a theory of screening, and the wide access to health care. Five selected examples that illustrate the history of medical screening are reviewed: screening for psychiatric disorders in the United States army as it is one of the oldest screening programmes; screening for syphilis as it used one of the earliest screening tests; screening for diabetes as one of the first modern forms of mass screening; screening for cervical cancer using the Pap test as one of the greatest successes of screening; and screening for breast cancer by mammography as this offers a good opportunity to discuss the development of modern evaluation of screening programmes. The evaluation of the impact of screening on human health slowly progressed, from obvious changes in the vital statistics such as the decline in incidence of syphilis, to less obvious changes such as the decline in mortality of cancer of the uterus, to finally more subtle changes, such as the impact of mammographic screening on breast cancer mortality. Methods of evaluation had therefore to adapt, evolving from simple surveys to case-control studies and randomised trials. The history of screening is short, but very rich and mostly still to be written.

See end of article for authors' affiliations

Correspondence to: Professor Alfredo Morabia, Division of Clinical Epidemiology, Geneva University Hospitals, 25 rue Micheli-du-Crest, 1211 Geneva 14, Switzerland; Alfredo.Morabia@hcuge.ch

Submitted
17 December 2003
Accepted 7 February 2004

Screening consists in identifying the presence of a disease while it is still in its preclinical stage. Rose and Barker, in their classic series of papers, indicated that in order to determine whether screening is beneficial, doctors had to answer three questions: ''Does earlier treatment improve the prognosis?'', ''How valid and repeatable is the screening test?'', ''What is the yield of the screening service?''.[1] These questions are very general, but are still valid. They also put stringent time limits to the history of screening as they imply that a full medical screening requires the simultaneous availability of some treatments (question 1), some screening tests as well as the concepts to assess their validity (for example, sensitivity and specificity) and their repeatability (question 2), and of some diagnostic tools to identify the true cases among those screened, that is, the yield (question 3). It was only during the 20th century that these three conditions became satisfied.

The history of the concepts and methods used for mass and clinical screening has not been written yet. It is quite an endeavour given the rapid diffusion of screening programmes, especially after 1945. This paper does therefore not pretend to be exhaustive and tell the whole story of screening. We decided to illustrate the history of medical screening by describing five diseases which show the evolution of ideas in this area of medicine. The examples are the screening of psychiatric disorders, syphilis, diabetes, cervical cancer, and breast cancer.

There are specific ''historical'' reasons for selecting these examples among a myriad of others. Screening for psychiatric disorders in the United States army is one of the oldest screening programmes we have found. The Wasserman test for syphilis is one of the earliest screening tests available. Its sensitivity and specificity were known at the beginning of the 20th century. The urine and blood glucose tests for diabetes have been intensively used in mass screening since the 1940s and this is one of the first examples of a modern form of screening. The Pap test for cervical cancer is one of the rare screening programmes that have achieved an almost exhaustive coverage in many female populations of the world. Finally, mammographic screening for breast cancer offers a good opportunity to discuss the evolution of a randomised controlled trial to assess the efficacy of screening and the classic biases (for example, lead time, length) related to cancer screening.

EARLY DEFINITION OF SCREENING

In 1951 the United States Commission of Chronic Illness defined screening as ''the presumptive identification of unrecognised disease or defect by the application of tests, examinations, or other procedures which can be applied rapidly. Screening tests sort out apparently well persons who probably have a disease from those who probably do not. A screening test is not intended to be diagnostic. Persons with positive or suspicious findings must be referred to their physicians for diagnosis and necessary treatment''.[2] The commission indicated that the most commonly used screening tests were blood glucose determination, a serological test for syphilis, radiography for chest pathology, and cytology for cancer detection. Indeed, in this

Abbreviations: HIP, Health Insurance Plan; NSA, Neuropsychiatric Screening Adjunct (test); OGTT, oral glucose tolerance test; RPR, rapid plasma regain; VDRL, Venereal Disease Research Laboratory (test)

paper we will review historical examples of the implementations of three of these screening procedures.

EXAMPLES

Example 1: screening for psychiatric disorders in the army

The earliest intervention we have found that in retrospect qualifies as a "screening" programme is the attempt to exclude subjects with psychological disorders from young men eligible to join the United States army. The Division of Psychology within the Medical Department of the United States Army had started in 1917 to administer mental tests to officers, drafted and enlisted men (p x–xi).[3] The purpose of psychological tests was "to help to eliminate from the Army at the earliest possible moment those recruits whose defective intelligence would make them a menace to the military organisation" (p 185).[3] The tests were used to examine "large numbers rapidly" and were expected to have "a high degree of validity as a measure of intelligence", to be objective, free "of personal judgment concerning correct answers" and "scoring could be done rapidly and with the least chance of error" (pp 2–3).[3] Subjects who scored positive on the test were referred to detailed individual psychological examination (p 10).[3] "Between April 27 and November 30, 1918, 7,749 men (0.5 per cent) were reported for discharge by psychological examiners because of mental inferiority" (p 21).[3]

During World War II, the Research Branch for the Surgeon General developed a standardised "pencil and paper" test to eliminate individuals identified as having psychiatric disorders from military service. The test was later named the Neuropsychiatric Screening Adjunct (NSA) test, and became a routine application at all induction stations of the United States Army.[4] Men with suspected results at the screening were referred to an observation ward for further diagnosis before a decision was made to recommend discharge from the Army (Star, p 491).[4 5]

The Army also developed a theory for the conception and interpretation of screening tests. In 1944 Hunt *et al* presented "an attempt at a theoretical evaluation of the screen test in military selection", in which they proposed the "selection index": TP/(TP+FN+FP) where TP = true positive, FN = false negative and FP = false positive.[5] The selection index differs from the current definition of the yield of a screening test in that it eliminates the true negatives from the denominator.

Around 1950, Star proposed to classify screening tests as being either *criterion-exclusive* or *criterion-inclusive* (see fig 1).[4] The goal of a *criterion-exclusive* screening was "to select a group all of whom have the desired characteristic without necessarily including within that group all who have that trait".[4] We would call it today a test with high specificity but not necessarily high sensitivity. A *criterion-inclusive* screening "strived to include within the selected group all persons who have the desired characteristic even though not all persons in the group will have the characteristic".[4] This corresponds to a test with high sensitivity but not necessarily high specificity. Indeed, the NSA test was referred to as "*criterion-inclusive*" because it intended to pick up at least as many of the "psychoneurotics" as the psychiatric interview identified. Actually, the scoring of the NSA test was "so designed that about a third of American young men, on the average, would manifest signs indicative of the need for very careful psychiatric examination…the cutting point was set high enough to include among the one third practically all of the psychiatric cases, as well as some who, upon careful examination, should be accepted by the Army" (p 475).[6] Using the distribution of psychosomatic complaints scores in the NSA test, fig 2 is probably one of the first demonstrations that the choice of a specific dichotomous cut point generates varying fractions of false positive and false negative results.

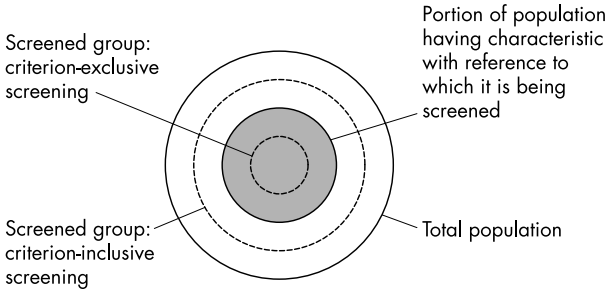

Figure 1 Difference between criterion-inclusive screening and criterion-exclusive screening. The dashed circle of the criterion-inclusive screening separates three groups of people: true negatives (most external white zone), false positives (internal white zone) and true positives (dotted central area). There are no false negatives. Therefore, a criterion-inclusive screening has perfect sensitivity and imperfect specificity. The dashed circle of the criterion-exclusive test separates three groups: true positive (the core dotted area), false negatives (the dotted area outside of the core) and true negatives (all the white areas). There are no false positives. Therefore, a criterion-exclusive test has perfect specificity and imperfect sensitivity (source: Star[4]).

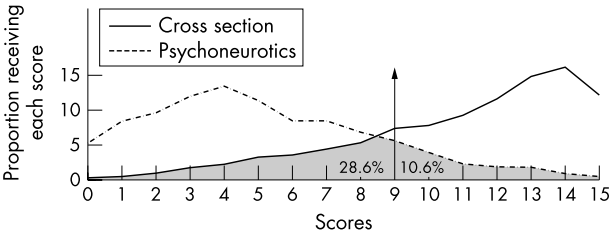

Figure 2 Demonstration of false positive and false negative using a dichotomous cut point. The arrow indicates the cut point of the NSA test; the solid line represents the distribution of psychosomatic complaints scores among the normals while the dash line depicts the distribution among the psychoneurotics. The shaded area below the solid line corresponds to the proportion of false positive (28.6%) and the shaded area below the dashed line denotes the proportion of false negative (10.6%) (source: Star[4]).

In 1950, Stouffer expressed both concepts of sensitivity and specificity without naming them as such: "For each man we have data on his score on the Neuropsychiatric Screening Adjunct (NSA) … and on the disposition made by the examining physicians" (p 474).[6] Results suggested among all psychiatric "rejects", 69.5% (that is, sensitivity) receive critical scores of NSA; among non-psychiatric "rejects", 21.8% (that is, 1−specificity) received critical scores (pp 475–476).[6]

Documents show that the test-retest reliability of the NSA test was also computed and that psychosomatic complaints showed that 90% of 400 subjects were consistently classified in the retest (p 503).[4]

It is of note that the Army had planned an evaluation of "the theoretical suppositions behind the screen test; and the economics involved, the practical factors of cost, necessary facilities, ease of administration, etc" (p 37).[5] However, the NSA test was officially adopted for use in October 1944, a few months before the end of the war. Its need diminished rapidly after the war and its impact could never be really evaluated.

In conclusion, this first example shows that the United States Army already had a clear concept in 1917 of what a screening programme should be and do. During World War II it developed a standard "paper and pencil" test along with all the theory needed to establish the threshold for positivity ("criterion-inclusive") and interpret the results.

Example 2: syphilis

The conditions for screening syphilis progressed very fast at the turn of the 20th century. Schmudinn and Hoffman isolated the bacteria, *Treponema pallidum*, in 1905. In 1906, Wassermann developed the first non-treponemal test using antigen extracted from the liver of newborns who died of congenital syphilis. In 1907 Ehrlich discovered the first effective treatment, arsenical compound salvarsan (or number 606), but it was associated with substantial and sometimes fatal risks of arsenical poisoning resulting in aplastic anaemia, exfoliative dermatitis, haemorrhagic encephalitis, and purpura haemorrhagica (p 15).[7] About 5.6% of United States Army enlistees had evidence of some type of venereal infection (p 36),[7] and most cases among the rank and file were probably never diagnosed. During World War I, the serological test was available but the therapy was not available or impracticable on a large scale. This may explain why the Army did not screen for syphilis at that time.

Screening for syphilis progressed between the world wars as some states passed laws requiring premarital (for both husbands and wives-to-be) or prenatal blood tests in pregnant women. There were pre-employment examinations, blood donor screening, and routine serological testing on hospital admission.[7]

In June 1944 penicillin became available to the United States Public Health Service for use in rapid treatment centres (p 16).[7] Because of production and refinement difficulties it was used primarily by the military, where it contributed to eradicate almost all venereal infections (p 38).[7] Immediately after World War II, penicillin was also made available for syphilis control in the civilian population (p 39).[7] Galenic preparations evolved to single daily shots over seven to 10 days and, in 1953, to oral benzathine penicillin G. The conditions were now present for syphilis detection by mass screening followed by treatment among virtually the whole population. In about 10 years, after World War II, health departments throughout the United States diagnosed and treated over two and a quarter million persons with syphilis (pp 38–39).[7]

Two types of tests became available for syphilis screening. The first category, such as Wasserman's, were *non*-treponemal tests that comprised the Venereal Disease Research Laboratory (VDRL) tests, the rapid plasma regain (RPR) teardrop card tests, the RPR circle card test, the toluidine red unheated serum test, and the automatic regain screen test. These tests were widely available, inexpensive, convenient to perform on large numbers of specimens, and were technically ideal for mass screening. The problem was that their sensitivity was low, ranging from 78% to 86% for primary syphilis but their specificity was satisfactory (>97%).[8]

The second category, *treponemal* antibody tests, are technically more difficult and costly to perform. The earlier ones, such as the *T pallidum* immobilisation test, had a poor validity.[8] Since 1957, the fluorescent treponemal antibody test and its improved versions provided highly specific tests, which remain the standard treponemal tests for syphilis today.

Treponemal tests appear to have similar sensitivity (from 76% to 84%), for primary stage infection, and specificity (>97%) as their non-treponemal counterparts.[8] It seems therefore that the reason for picking non-treponemal tests for screening and treponemal tests for confirmation has not much to do with their respective validity but with considerations of cost and availability.

Thus, simultaneous availability of a rapid test and rapid treatment at the end of World War II opened new perspectives for syphilis screening. Of the 15 million men entering the United States Armed Services, 750 000 had positive serological tests for syphilis and/or clinical signs or symptoms. Over 300 000 of the infected men were treated rapidly by civilian health departments, rendered non-infectious, and inducted into the Armed Service.[7]

The so-called ''blitzes'', an unfortunate import of German Nazi military jargon, were a new form of screening widely used in the 1950s and 1960s. ''The blitz procedure in syphilis control is an intensive attack on the disease in which efforts are directed at rapid examination and treatment of all named contacts of interviewed patients with syphilis''.[9] For example, in August 1965, the Department of Public Health of Alabama conducted five blitzes on syphilis, which took from two to nine days each and screened 739 contacts of 196 initial patients with primary, secondary, or early latent syphilis using an immediate RPR card test and VDRL serological test for syphilis.[9] Persons with positive tests or negative contact but exposed to a person with infectious syphilis within the previous four months received 2.4 million units of benzathine penicillin G or an alternative antibiotic.[9]

There was no specific evaluation of the screening programmes but the impact of the campaigns on vital statistics was so striking that this may have been deemed sufficient. The number of syphilis patients dropped dramatically and reduced the cost effectiveness of mass testing, which was then eliminated and replaced by selective testing of suspected high incidence subgroups of the population (p 39).[7] By the mid-1950s reported cases of syphilis had declined so sharply that the disease was considered another conquered problem. Since 1960s, the routine serological testing programmes, including premarital screening tests and hospital required preadmission testing, were discontinued in many states. However, since the 1980s, a significant increase in incidence of syphilis has occurred, in particular among subgroups of people at high risk of HIV infections.[10]

Example 3: diabetes

The notion of normal blood glucose homoeostasis and its grave disturbance had been appreciated since the 19th century. Around 1900, the medical departments of life insurance companies in New York apparently performed urine glucose tests on 71 729 persons. Prevalence of glycosuria in men was 2.8%, and a urine glucose concentration of 1% or greater was presented in 0.9% of men.[11] There are indications that screening for diabetes was performed in specific groups before 1940s, including tests aiming to reject those with diabetes to the military service during World War I.[12–15] Unfortunately, we haven't found data allowing us to compute the validity of these early tests.

Before the discovery of insulin in 1923, diabetic patients were prescribed ''several days of starvation following a diet of undernourishment''.[16] Insulin injections rapidly changed the history of treatment of diabetes, becoming, along with dietary control, its major therapy. When there was a rapid increase in deaths from diabetes in the 1940s, conditions were ripe to widely carry out mass screening.[17–19]

The first large scale community diabetes screening was probably done in Oxford, Massachusetts, in 1946–47[18] by the United States Public Health Service. Its aims were to ''(a) determine the prevalence of diabetes in a typical American community; (b) evaluate the techniques and methods or large scale diabetes diagnosis; (c) instill in the members of a community a realisation of the need for periodic examinations for diabetes, and (d) discover every cases of diabetes, so that, through prompt treatment by the family physician, further progression and complications may be avoided'' (pp 210–211).[18] Of the 4983 inhabitants in Oxford, 70.6% received both the urine and the blood glucose testing. Urine and blood were obtained about an hour after the midday or evening meal. The urine glucose was determined by a standard Benedict's qualitative test, followed up when positive by a

quantitative Benedict's test. Blood glucose was determined according to the method of Folin and Wu.[18] Prevalence of diabetes was 1.7% in this population. We could not compute sensitivity and specificity of the tests because no data are available for the percentage of diabetes among those who tested negative. In order to evaluate the methods of large scale diabetes diagnosis, family physicians were forwarded the screening results with an offer to perform any further tests they might wish for diagnosis purposes.

The validity of *urinary* glucose as a standard screening test was always known to be low because of its poor sensitivity, which is estimated to be 16.7% for fasting values and 72.7% for two hour post-load values.[20] Hence, a urine test was used only where blood testing was either not available or particularly expensive.

On the contrary, *blood* glucose determination including random blood glucose test and fasting whole blood (or plasma) glucose measurement demonstrated superior sensitivity and specificity. It was often used together with the urinary glucose test to improve the sensitivity of mass screening.[18 21–23] Low cost and automated methods of assessing blood glucose greatly enhanced mass screening in the 1950s (p 42).[12]

Later, the oral glucose tolerance test (OGTT) was developed to take into account the time period preceding carbohydrate intake because of the argument that a normal fasting blood glucose does not exclude diabetes and the length of the tolerance curve is more important than its peak.[21 22] However, because of the inconvenience associated with multiple venepunctures, the OGTT was used only when fasting glucose testing was inadequate to categorise the subjects.[18 19]

Despite several tests developed for screening asymptomatic diabetes, the clinical diagnosis still depended on thorough clinical evaluations. Individuals with borderline or positive findings at screening had to be notified and advised to seek a further medical check-up to obtain a definite diagnosis. This complexity of the screening procedure, compared with that of syphilis, for example, may explain why the first monograph systematically reviewing the interpretation of screening tests in the United States appeared in 1961 and was about diabetes screening.[24] In its foreword, Thorner and Remein wrote "The experience of the chronic disease programme of the Public Health Service in aiding in the conduct of seminars concerned with screening programmes, chiefly for diabetes, indicated the need for a compendium of information on screening tests. One session of these seminars is always devoted to test attributes and especially to sensitivity and specificity. Though the literature includes many papers

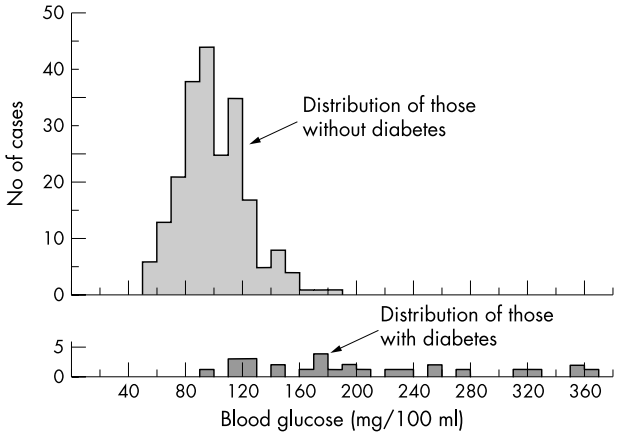

Figure 3 Distribution of blood glucose levels by screening test in patients with diabetes and patients without diabetes; to convert mg/100 ml to mmol/l multiply by 0.05551 (source: Blumberg[25]).

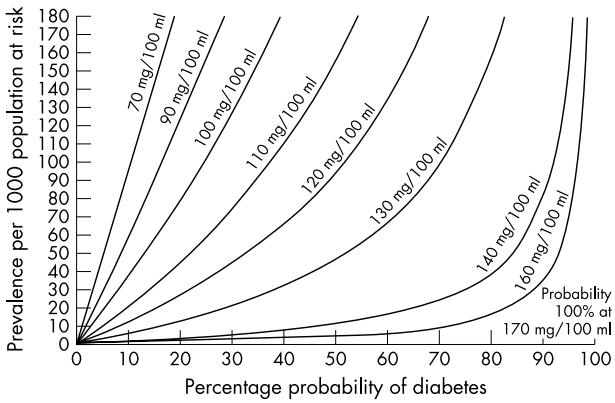

Figure 4 Diabetes prevalence probability curves. The horizontal axis reflects the probability of diabetes among positive screenees; the vertical axis indicates the prevalence of diabetes per thousand population at risk. The different curves correspond to different cut points of the blood glucose test, varying from 70 mg to 160 mg per 100 ml of venous blood two hours after a meal. The sensitivity and the specificity of the tests are 64.3% and 96.9% respectively. This graph shows that the probability of diabetes among positive screenees (positive predictive value) increases as the prevalence of diabetes increases, and the probability also goes up when the cut point is set higher assuming the same prevalence and a constant sensitivity and specificity; to convert mg/100 ml to mmol/l multiply by 0.05551 (source: Kessel[23]).

dealing with various aspects of screening tests, up to now there has been no single source to which seminar students could be referred" (p iv).[24]

This monograph could still be used today to present the theory of screening as the concepts described and their definitions have not much evolved. Sensitivity is "the ability of a test to give a positive finding when the person tested truly has the disease under study". Specificity is "the ability of the test to give a negative finding when the person tested is free of the disease under study". Two by two tables and formula indicated how to calculate the sensitivity and specificity—for example, "Sensitivity is calculated by a/(a+c)×100", "Specificity is calculated by d/(b+d)×100". The monograph explained how to select a cut point for positivity when the test is a continuous variable, and how to combine screening tests in parallel or in serial (pp 2–7).[24] In a striking intuition of what will be called later "spectrum bias", it also described how sensitivity and specificity could vary with "attributes" of the population, such as age.

The now classic example and graph (fig 3) by Blumberg illustrates the basic problem in selecting an optimum cut point for positivity. Data came from "a group of 218 hospital outpatients [who] were all given a screening blood-sugar test for diabetes. Each was then given a more elaborate set of diagnosis tests for diabetes. This diagnostic testing was done to validate or calibrate the screening procedure".

Blumberg wrote that "it can be seen that there is no blood-sugar screening level which will ensure that separation of all those with the disease (true positives) from all those without the disease (true negatives). Instead, screening at any blood-sugar level would result in either false positives (indicating that people have the disease when they do not) or false negatives (indicating that people do not have the disease when they do)".[25]

Nonetheless, in the field, some of these concepts were still fresh and sometimes still mixed up. For example, in the report of the Brookline study, specificity was defined as "What proportion of those singled out for further testing are truly positive", which is really the positive predictive value. Based on table 4 of the original paper the positive predictive

value for the urine glucose test is 35.7% and for the blood glucose test is 49.2%.[19]

The relation of prevalence to predictive values of diabetes screening tests was appreciated in the early 1960s. Kessel stressed that sensitivity and specificity ''tell us nothing about the probability of any screenee or group of screenees having diabetes''.[23] He proposed instead graphs and formulas which describe exactly how the positive predictive value is calculated today and how it changes with increasing pre-test prevalence, given fixed sensitivity and specificity (fig 4).

We found the first reference to the term ''predictive value'' in 1966, when Vecchio defines it as ''the likelihood that a subject yielding a positive test actually has the disease'' or ''the likelihood that a subject with a negative test does not have the disease''.[26]

Example 4: cervical cancer

Cervical cancer was probably the most frequent malignancy in Western Europe in the middle of the 19th century. Its natural history makes it detectable in its preclinical phase and increases the chances of cure and of mortality reduction. Before World War II there had only been pilot programmes of cancer detection.[27] After World War II, the development of a cervical cancer cytological test (mainly the Pap smear) created a public health paradox previously unknown in the field of cancer: incidence increased and mortality decreased, mainly because many of the newly detected tumours were in situ and had therefore an excellent prognosis after surgery.[28]

Papanicolaou and Traut first reported the usefulness of Papanicolaou (''Pap'') smear for detecting neoplastic cervical cells in 1943. After that, it was widely used as the screening tool in the late 1940s to early 1950s. The Pap test is not perfect in that the technique contains a cytological interpretation of a smear of cells taken from the cervix that is subject to error at a number of levels.[29] In the 1960s, the sensitivity of the Pap smear to detect carcinoma of the cervix at different stages was considered to vary between 89% and almost 100%.[30 31] Cases detected as positive at screening will be confirmed as a matter of routine by standard clinical diagnostic methods including biopsy and histological diagnosis.

An early and maybe the earliest cancer detection centre was established in 1937 by Dr Elise L'Esperance in New York City and offered comprehensive examinations to asymptomatic adults for the purpose of early cancer diagnosis.[32] The women received a cervical Papanicolaou smear with a confirming surgical biopsy on finding a suspicious lesion. If a precancerous lesion was detected, patients were referred to a surgeon for excision of the lesion or were carefully observed in the clinic. In 1937, there were only 71 applicants; in 1946, the number went up to 1356 and there were 3016 return visits. The centre at Memorial Hospital was established three years later, and there were 150 new patients registered; in 1946, the new patients admitted went up to 5713 and the return visits amounted to 8740. By the 1950s, over 250 similar cancer detection centres had been established in the United States.[32 33] Treatments of preinvasive lesions evolved from early radical methods such as hysterectomy and radiotherapy to cone biopsies, which could entirely remove some in situ lesions by a simpler operation. The advent of colposcopy made local ablative therapy possible, including diathermy, cryosurgery, and laser therapy.[34] Treatment for these preinvasive lesions has proved to be very successful in that more than 99% of the cases will not progress into invasive cancer after early treatment.

There is a wealth of evidence indicating that Pap smears are associated with a significant decrease in the mortality of cervical cancer. Of historical relevance is the use of case-control studies to evaluate the efficiency of this form of screening (p 105).[35] In a study by Clarke and Anderson conducted between 1 October 1973 and 30 September 1976, 212 cases were compared to 1060 controls free of cervical cancer.[36] Cases were women aged 20 to 69 hospitalised with newly diagnosed invasive carcinoma of the cervix. Five controls were matched to each of the cases by age, neighbourhood, and type of dwelling. The subjects were interviewed at home regarding their history of Pap smears, and other factors.[35 36] The relative risk of invasive cervical cancer was estimated as 3.3 in women who had not, compared with women who had, been screened, as 68% of cases and 44% of controls had *no* screening Pap smear during the five years before the year of diagnosis.

However, the potential biases associated with screening evaluation based on case-control studies were soon recognised. For example, cases may have received more Pap smears than controls due to the presence of gynaecological symptoms, resulting in more Pap smears being associated with invasive tumour than with subclinical forms. The overall bias might therefore spuriously reduce the observed benefit of the screening.[35] Also, the efficiency of screening will be underestimated when newly diagnosed (and therefore recently screened) cases are compared with controls because more cases are likely to have undergone screening than controls.

Example 5: breast cancer

Mass screening of breast cancer virtually started when mammography became available as a screening test in the 1960s. Egan and others suggested that mammography was able to detect impalpable breast tumour and distinguish malignant tumour from benign breast disease.[37 38] However, in contrast to all our previous examples, it remained uncertain whether early detection of breast cancer could reduce the cancer mortality rate. The innovation here was to propose a randomised trial to evaluate the potential of mammography screening and clinical examination of the breast for lowering mortality of female breast cancer.[39]

A preliminary study, based on 15 hospitals, showed that ''Egan's technique of mammography was highly reproducible among many radiologists following the standard training''.[40 41] In December 1963, the Health Insurance Plan (HIP) of Greater New York launched a randomised clinical trial which lasted for almost 25 years and was ended in 1986.[35] Here is how Shapiro, its conceptor describes the trial: ''The HIP is a comprehensive, prepaid medical insurance plan. Care is provided by a number of participating group practices. There were 62 000 women aged 40 to 64 years (members of HIP for at least one year) that were identified for study [between December 1963 and June 1966]. These women were stratified by age, family size, and the type of employment on which HIP membership was based. Then, the women were assigned alternatively, based on identification number, to the 'study' (screened) group or the 'control' (usual care) group [the total number of women in each group was about 31 000]. Since the assignment of the identification number was not related to any personal characteristics, the method of allocation to screened and control groups was effectively random. … Each woman in the study group was invited to have a screening examination for breast cancer, and each woman who had one (without cancer being found) was asked to appear for three annual follow up examinations. Members of the control group were not screened, but were eligible for all HIP benefits that included, if desired, general physical examinations…….''.[40]

The screening examination consisted of film mammography (cephalocaudal and lateral radiographs of each breast), and of an independently conducted physical examination of the breasts done by a physician, usually a surgeon.

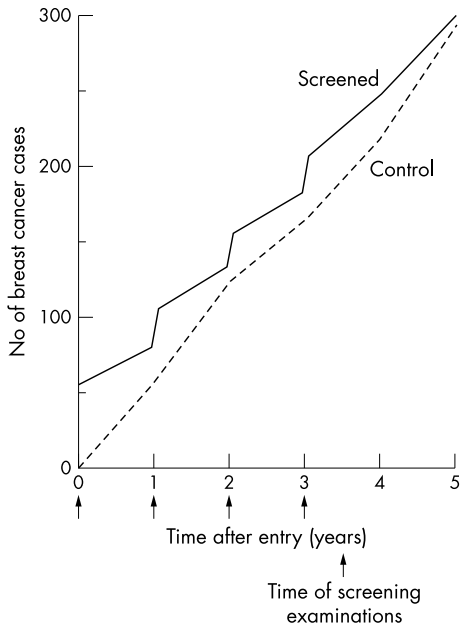

Figure 5 Cumulative numbers of cases of breast cancer diagnosed by time in the HIP study. The figure was created by Morrison[35] based on the tables reported in the paper by Shapiro *et al*.[42]

For each positive finding, a biopsy was usually conducted and eventually followed by treatment (mostly, radical mastectomy).

Based on earlier data from the HIP study in 1960s, physical examination was more sensitive than mammography (67% *v* 55%). A decade later, due to the improvement in mammography technique, the Breast Cancer Detection Demonstration Project study suggested mammography was almost twice as sensitive as physical examination (94% *v* 54%). Shapiro *et al* estimated that the sensitivity and specificity of initial screening examinations (mammography + physical examination) were 80% and 93% respectively (pp 53–54).[40]

The primary outcome to evaluate the efficacy of the screening programme was the change in mortality rates due to breast cancer in the screened group and the control group. During the entire 18 year follow up period, the initial diagnosis of breast cancer and deaths from all causes in the study and control group were identified from various sources to ensure the complete and accurate ascertainment. There were no cases lost to follow up in the HIP study. However, the investigators realised that a misinterpretation of the findings could stem from two sorts of biases: lead time bias and length bias.

Lead time and length biases

Because the benefit of breast cancer screening was not as obvious as those for cervical cancer, the issue of biases, which would spuriously boost the effect of screening, became major issues. Two of them, ''lead time'' bias and ''length'' bias are now classic examples.

First, lead time bias: screening could advance the diagnosis of cancer without necessarily prolonging the overall duration of the woman's life. Practically, screened cases should be less advanced. Indeed, 71% of the HIP screen-detected cases were in the localised stage at the time of diagnosis, but only 46% were localised in the unscreened ''control'' group. And cancer detection rate should be higher in the screened group during the first years of the programme but not later. The graph (fig 5) illustrating the latter phenomenon in the HIP breast cancer study is now a classic. The number of diagnosed cases in the screened group clearly exceeded the number in the control group up to the fifth year of the programme, but the difference diminished after five years of follow up.

Second, length bias: cases with very *short* preclinical phases have little chance of being detected before they become clinical, but cases with *long* preclinical phases are very likely to be detected by the screening programme. The first description of this concept of length time bias we found was by Dunn.[43] The term itself may have been invented by Zelen in 1976 and defined as ''the tendency for screening to identify cases with a relatively long preclinical phase''.[35 44]

The HIP study found about a 30% reduction in mortality from breast cancer during the first 10 years of follow up in the group of women aged 40 to 64 years at entry. By the end of 18 years from entry, the reduction was close to 25%.[40] However, the history of breast cancer screening and of its controversies goes on.

CONCLUSION

Medical screening has existed for about 60 years, and has a very rich history. The preclinical identification of disease has been a major component of modern medicine and public health. It has contributed to some of its major successes, examples of which have been discussed in this paper.

We observed that, after World War II, the convergence of cheap and non-invasive tests, handy galenic forms of treatment, a theory of screening, to which we add the access to care, made the development of mass screening campaigns possible. The evaluation of the impact of screening on human health slowly progressed, from obvious changes in the vital statistics such as the decline in incidence of syphilis, to less obvious changes such as the decline in mortality of cancer of the uterus to finally more subtle changes, such as the impact of mammographic screening to breast cancer. Methods of evaluation had therefore to adapt, evolving from simple surveys to case-control study and randomised trials.

Our history of screening was essentially based on examples from the United States and Canada. The abundance of material made our search easier. We did not add ''in North America'' to the title of this review because we have no reason to suspect that the history of screening has been different elsewhere. We may be proved wrong on this aspect by future research. We have also left out some major candidates—old ones, such as screening tuberculosis using miniradiophotography, new ones, such as screening lung cancer using computed tomography, or controversial ones, such as mass screening for HIV. We hope that others will be tempted to explore the history of screening for other diseases and also for other countries and other continents to contribute to achieve a more exhaustive picture.

Authors' affiliations

A Morabia, F F Zhang, Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA
A Morabia, Division of Clinical Epidemiology, Geneva University Hospitals, Geneva, Switzerland

REFERENCES

1 **Rose G**, Barker D. Epidemiology for the uninitiated. Screening. *BMJ* 1978;**ii**:1417–18.
2 **Commission on Chronic Illness**. *Chronic illness in the United States. Vol I. Prevention of chronic illness*. Cambridge, MA: Harvard University Press, 1957;**1**:45.
3 **Yoakum CS**, Yerkes RM. *Army mental tests*. New York: Henry Holt And Company, 1920.
4 **Star S**. The screening of psychoneurotics in the army: technical development of tests (chapter 13). In: Guttman L, Suchman E, Lazarsfeld P, *et al*, eds. *Studies in social psychology in World War II*. Vol IV. Princeton, NJ: Princeton University Press, 1950.
5 **Hunt W**, Wittson C, Harris H. The screen test in military screening. *Psychol Rev* 1944;**51**:37–64.

6 **Stouffer SA**. Two case studies in prediction: introductory comments (chapter 12). In: Guttman L, Suchman E, Lazarsfeld P, *et al*, eds. *Studies in social psychology in world war. Measurement and prediction*. Vol IV. Princeton, NJ: Princeton University Press, 1950.
7 **Brown WJ**. *Syphilis and other venereal disease*. Edited by the American Public Health Association. *Vital and health statistics monographs*. Cambridge, MA: Harvard University Press, 1970.
8 **Larsen S**, Steiner B, Rudolph A. Laboratory diagnosis and interpretation of tests for syphilis. *Clin Microbiol Rev* 1995;**8**:1–21.
9 **Smith W**. Blitz on syphilis in Alabama. *Public Health Rep* 1966;**81**:835–41.
10 **Johnson PC**, Farnie MA. Testing for syphilis. *Dermatol Clin* 1994;**12**:9–17.
11 **Barringer TJ**. The incidence of glycosuria and diabetes in New York City between 1902 and 1907. *Arch Intern Med* 1909;**3**:295–8.
12 **West K**. *Epidemiology of diabetes and its vascular lesions*. New York: Elsevier, 1978:42.
13 **John H**. Observations on diabetes mellitus in the US Army. *Proc Am Diabetes A* 1943;**3**:69–77.
14 **Joslin E**. Diabetes and military service. *JAMA* 1943;**121**:198–200.
15 **Emerson H**, Larimore LD. Diabetes mellitus: a contribution to its epidemiology based chiefly on mortality statistics. *Arch Intern Med* 1924;**34**:585–630.
16 **Striker C**. *Famous faces in diabetes*. Boston: G K Hall & Co, 1961.
17 **Marks H**. Statistics of diabetes. *Diabetes* 1946;**235**:289–94.
18 **Wilkerson H**, Krall L. Diabetes in a New England town. A study of 3516 persons in Oxford, Mass. *JAMA* 1947;**135**:209–216.
19 **Harting D**, Glenn B. A comparison of blood-sugar and urine-sugar determinations for the detection of diabetes. *N Engl J Med* 1951;**245**:48–54.
20 **Forrest RD**, Jackson CA, Yudkin JS. The glycohaemoglobin assay as a screening test for diabetes mellitus: the Islington diabetes survey. *Diabetic Medicine* 1987;**4**:254–9.
21 **Blotner H**, Marble A. Diabetes control: detection, public education and community aspects. *N Engl J Med* 1951;**245**:567–75.
22 **Chesrow E**, Bleyer J. Results of diabetes detection drives. *Geriatrics* 1956; March:119–26.
23 **Kessel E**. Diabetes detection: an improved approach. *J Chronic Dis* 1961;**15**:1109–21.
24 **Thorner R**, Remein Q. *Principles and procedures in the evaluation of screening for disease*. United States Department of Health. Public Health Monograph, 1961:67.
25 **Blumberg M**. Evaluating health screening procedures. *Operations Research* 1957;**5**:351–60.
26 **Vecchio T**. Predictive value of a single diagnostic test in unselected populations. *N Engl J Med* 1966;**274**:1171–3.
27 **Day E**. Cancer screening and detection: medical aspects. *J Chronic Dis* 1963;**16**:397–405.
28 **Kessler I**. Cervical cancer epidemiology in historical perspective. *J Reprod Med* 1974;**12**:173–85.
29 **Coppleson L**, Brown G. Estimation of thte screening error-rate from the observed detection rates in repeated cervical cytology. *Am J Obstet Gynecol* 1974;**119**:953–8.
30 **Patten SFJ**. *Diagnostic cytology of the uterine cervix*. In: Wied GL, Haam EV, Koss LG, *et al*, eds. *Manographs in clinical cytology*. Baltimore: Williams & Wilkins, 1969.
31 **Friedell GH**, Hertig AT, Younge PA. *Carcinoma in situ of the uterine cervix*. Springfield: Charles C Thomas, 1960.
32 **American Cancer Society**. *Proceedings of the conference on cancer detection*. Portsmouth, NH: ACS, 1949.
33 **L'Esperance E**. The Strang cancer prevention clinics. *J Am Med Wom Assoc* 1948;**3**:131–46.
34 **Chamberlain J**. Reasons that some screening programmes fail to control cervical cancer. In: Hakama M, Miller A, Day E, eds. *Screening for cancer of the uterine cervix*. New York: Oxford University Press, 1986.
35 **Morrison A**. *Screening in chronic disease*. In: Lilienfeld A, ed. *Monographs in epidemiology and biostatistics*. Vol 7. New York: Oxford University Press, 1985.
36 **Clarke EA**, Anderson TW. Does screening by ''Pap'' smears help prevent cervical cancer? *Lancet* 1979;**ii**:1–4.
37 **Egan R**. Mammography as an aid to diagnosis of breast carcinoma. *JAMA* 1962;**182**:1075–90.
38 **Gershon-Cohen J**, Harmel M, Berger S. Detection of breast cancer by periodic x-ray examination. *JAMA* 1961;**176**:1114–16.
39 **Shapiro S**, Strax P, Vennet L. Periodic breast cancer screening. *Arch Environ Health* 1967;**15**:547–53.
40 **Shapiro S**, Vennet W, Strax P, *et al*. *Periodic screening for breast cancer: the Health Insurance Plan Project and its sequelae, 1963–1986*. Baltimore: John Hopkins University Press, 1988.
41 **Clark R**, Copeland M, Egan R. Reproducibility of the techinique of mamography (Egan) for cacner of the breast. *Amercian Journal of Surgeons* 1965;**109**:127.
42 **Shapiro S**, Goldberg JD, Hutchison GB. Lead time in breast cancer detection and implications for periodicity of screening. *Am J Epidemiol* 1974;**100**:357–66.
43 **Dunn J Jr**. Screening for cancer. *J Chronic Dis* 1955;**2**:450–60.
44 **Zelen M**. Theory of early detection of breast cancer in the general population. In: Heuson JC, Mattheim WH, Rozencweig M, eds. *Breast cancer: trends in research and treatment*. New York: Raven Press, 1976:287–300.