Understanding screening: requirements for a successful programme

Julian PL Davis, Iain K Crombie, Huw TO Davies

Screening has generally been successful in identifying those at risk from disease. This success has led to the belief that screening in the general population is always a good thing. However, there are pitfalls which must be avoided if screening programmes are to achieve what is intended for them.

INTRODUCTION

Early physicians would probably have looked with suspicion upon one of their number claiming to be able to identify individuals who have a disease before they showed symptoms; to them it may well have smacked of witchcraft. To us, the advent of modern biochemical and histological techniques has now made this a commonplace fact of clinical life, and a multitude of screening programmes are now in place across the world.

There have been many successful screening programmes, such as that for cervical cancer in the UK. This was initiated in the 1960s, and followed up by the introduction of the national recall system in 1988. This programme has been credited with a reduction in the incidence of cervical cancer of 35% over the 15-year period up to 1995, and with a significant fall in mortality (Quinn et al, 1999).

Screening is usually seen as a good thing, because it uses simple tests to identify disease in a proportion of the population before symptoms present. Treatment can then relieve suffering and save lives. As a consequence of this, there is a tendency to call for screening programmes to be set up as soon as a test becomes available. This has been an area of heated debate for many years in both popular and medical journals.

For example, the issue of screening for prostate cancer has been a point of contention. Although some articles in the popular press advocate it (Stuttaford, 1995; Smith, 1995), the weight of medical opinion is that its use is not yet justified (Selley et al, 1997; Woolf, 1997; Abbasi, 1998). As well as the political drive, there is also a greater direct public interest. Information is now more freely available via the internet, and patients (and potential patients) will discover for themselves the existence and availability of research and of specific tests. The result of this is that the pressure for screening will grow. A positive test can cause great mental distress, and it is important that the requirements for the establishment of a successful and effective screening programme are better and more widely understood.

This article aims to give an introduction to the epidemiological basis of screening, to give a brief overview of the ways in which it works (or does not work), and to highlight some of the problems faced in setting up such programmes.

Dr Julian PL Davis is Research Fellow and **Professor Iain K Crombie** is Professor of Epidemiology and Public Health, Department of Epidemiology and Public Health, University of Dundee, Ninewells Hospital and Medical School, Dundee, and **Dr Huw TO Davies** is Reader in Health Care Policy and Management, Department of Management, University of St Andrews, St Andrews, Fife KY16 9AL

Correspondence to: Dr HTO Davies

WHY SCREEN, AND WHEN IS IT SUITABLE TO DO IT?

Screening may be defined as the use of testing to identify patients, who have not presented with symptoms, who are at sufficient risk of developing a disorder that it warrants further investigation or treatment. The main aim of screening is to prevent disease, or to minimize its consequences. This may be primary prevention, such as the screening for risk factors for a condition (e.g. hypertension in cardiovascular disease), or secondary prevention, such as the early detection of cancer.

CRITERIA FOR SUCCESSFUL SCREENING

There are a number of criteria which must be satisfied before a screening programme should be launched upon an unsuspecting public. The criteria were set out in depth for the World Health Organization 30 years ago (Wilson and Jungner, 1968). Included below is a simplified version of the criteria which are generally accepted attributes of a successful screening test (modified from Mausner and Kramer, 1985):

1. The condition must be sufficiently common in the target population for the number of cases identified and treated as a result of screening to be significant. A screening programme is complex to establish and maintain, and benefits in terms of lives saved must be sufficient to justify this investment.
2. The condition must have a well-defined latent period, a relatively slow rate of progression from onset, and a well understood natural history. Also, since screening is done on a cyclical basis, the repeat interval for the test should be shorter than the latency of the condition.
3. There must be a suitable test available for the condition, and the test itself must satisfy certain validity criteria (see below).

4. There must be an effective (and cost-effective) treatment for the condition — you cannot tell a patient they have a high risk of contracting something if you can do nothing about it. Equally, if a treatment exists, but is not universally available, those screened must be eligible for treatment if it is shown that they are at risk.
5. The early detection of the condition must substantially improve the prognosis compared to waiting until symptoms present. The margin is clearly a subjective one, but there must be consensus beforehand on what level of improved prognosis is acceptable. There should also be agreement that the proposed screening programme will deliver that improvement.
6. The screening test itself should not be harmful, frightening or overly intrusive. Even if a test exists, the take-up (see yield below) will be poor if the test is painful, undignified or unpleasant. For example sigmoidoscopy screening for rectal cancer is possible, but many patients find the test uncomfortable, and prefer faecal occult blood test screening. This is somewhat less sensitive, but less invasive (Scholefield, 2000). The patient must perceive that taking the test now is much less unpleasant than the threatened symptoms would be several years hence.
7. The costs must be justified. The costs of screening are considerable, and in today's financial climate, cost-effectiveness is essential. The test itself is often inexpensive, but follow-up testing for positives, treatment if appropriate and all the incidental costs and resources should also be taken into account. Equally, decisions about the basis of recall and patient eligibility can be made with reference to cost-effectiveness (Boer et al, 1998). Screening should result in lower costs than treating once symptoms present.

CHARACTERISTICS OF TESTS

There is no point in initiating a screening programme based upon a test which is not able reliably to detect the condition in which you are interested. There are a number of inherent characteristics which must be assessed when considering whether a test is suitable for becoming the basis of a screening programme. These are reliability, validity and yield.

Reliability

Any screening programme can only be as reliable as the test upon which it is based. A test's reliability depends upon two factors, the test's own repeatability (method variation), and the repeatability with which the results are interpreted (observer variation). The repeatability of the test itself can increasingly be assured by the growing use of technology in the testing process. Observer variation is more difficult to control, since it can not only result from differences between observers, but also from different performances of the same observer at different times. Careful methodology, and intensive training and revalidation can help to reduce these sources of error.

Validity

The validity of a test is a measure of its ability to perform its intended function, that is to detect those in a population who have a given condition and those who do not. There are two components to validity — sensitivity and specificity. These are inherent characteristics which are vital to the successful use of a test in a screening programme.

Sensitivity: The sensitivity of a test is its accuracy in detecting those individuals in a population who have a condition. In mathematical terms, this is the number who test positive expressed as a percentage of all those who actually are positive, that is to say who have the condition (*Figure 1*). Thus a test which has a sensitivity of 100% will detect all individuals in a given population who have the condition. It will give no false negatives.

Specificity: The specificity of a test is its ability to identify those individuals who do *not* have a condition. This is expressed as those who test negative as a percentage of those who are negative (*Figure 1*). Thus a test with 100% specificity will identify all those who do not have a condition. Such a test will give no false positives.

The sensitivity/specificity trade-off: The majority of tests do not have 100% sensitivity and specificity. In other words they are generally better either at identifying sufferers or non-sufferers, but not both. As with much in life, the result is a trade-off, in which the decision must be made as to where to draw the line between two extremes.

A simple example will serve to illustrate this point. The hypothetical data in *Table 1* show a sample of patients who are to be part of a screening exercise for kidney disease. The test used is the serum creatinine level.

The trade-off is illustrated by *Table 2*, in which the sensitivity and specificity are calculated by using the numbers from *Table 1* and the formula from *Figure 1*. As the cut-off point is moved, so the balance between specificity and sensitivity changes. If the choice is made to classify all patients with a blood creatinine of >100 mg/100 ml as having kidney disease, the sensitivity is excellent, that is to say the test will pick up all patients

	Disease	No disease
Positive test	a	b
Negative test	c	d
Total	a+c	b+d

Sensitivity = a/(a+c) x 100

Specificity = d/(b+d) x 100

Positive predictive value = a/(a+b) x 100

Negative predictive value = d/(c+d) x 100

Figure 1. Method for calculating sensitivity, specificity, positive and negative predictive value.

who have kidney disease. However, the specificity is poor, and this means that the test will also pick up a great many patients with normal renal function — false positives.

At the other extreme, if the cut-off is set at 180 mg/100 ml, the specificity increases, reducing the number of false positives dramatically, but the sensitivity suffers, leaving 10 patients who do have kidney disease undetected — false negatives.

A cut-off of 140 mg/100 ml gives a compromise in which both sensitivity and specificity are over 80%. Clearly this cut-off point can be moved to suit the primary purpose of the screening test, whether this is identification of those at risk, or reassurance of those who are not.

The effect of prevalence: The sensitivity and specificity are not in themselves sufficient to predict whether a given test will perform well in identifying a particular individual with a given condition. The prevalence of the condition in the population is also an important factor.

Again, in order to explain this, we will use a fictitious example. Suppose we have developed a test that has a sensitivity of 95%, and a specificity of 90%, for detecting a mythical condition. We are interested in a sample of 1500 patients who may have this condition. However, if we select our samples from different populations, the results will be very different.

Testing in the outpatient clinic: In the first scenario, the sample is drawn from those attending a hospital outpatient clinic. Here, patients are more likely to be suffering from disease, since they will have been referred there on the basis of symptoms. For this example, assume the prevalence of our condition in this population is 20%. *Table 3* illustrates how the numbers work out.

Before testing, we know that each individual has a 20% chance of having the disease and an 80% chance of not having it (we can tell this from the prevalence). After testing, the positive and negative predictive values update these figures according to whether the test was positive or negative. If the test is positive, the chance of actually having the disease is now seen to be 70%; if the test is negative the chance of not having the disease is now 98.6%. So whatever the test result we now know much more than we did before.

Testing in the community: In the second scenario, however, the sample is drawn instead from the community. Here, the prevalence of our hypothetical condition is only 2%. *Table 4* shows what effect this has on the positive and negative predictive values. Although a negative test is now almost definitive (99.9% sure that there really is no disease), a positive test only gives a 16.5% chance of there really being a disease. Because of the lower initial prevalence in the community compared to the clinic, the test performs much less well — and erroneously identifies many well people as sick.

The message from this is that the prevalence of a condition has a powerful effect on the ability of a screening test to identify which patients are at risk. If a screening programme is to be instigated in a particular setting, it is important to know beforehand that the prevalence of the condition in that population is sufficiently high to allow the test to identify correctly a high proportion of sufferers.

TABLE 1.
The sensitivity/specificity trade off

	Serum creatinine (mg/100 ml)					Total
	<100	100–139	140–179	180–199	≥200	
Patients with normal kidney function	41	152	32	2	0	227
Patients with kidney disease	0	4	6	9	15	34

TABLE 2.
Moving the cut-off point

Serum creatinine cut off (mg/100 ml)	≥100	≥140	≥180	≥200
Test sensitivity	100%	88.2%	70.6%	44.1%
Test specificity	18.1%	85.0%	99.1%	100%

TABLE 3.
Fictitious data for illustration of the prevalence problem. Scenario 1: in the outpatient clinic

	Disease (sensitivity 95%)	No disease (specificity 90%)	Total
Positive test	285	120	405
Negative test	15	1080	1095
Total	300	1200	1500

Positive predictive value = 70.4%; negative predictive value = 98.6%

TABLE 4.
Fictitious data for illustration of the prevalence problem. Scenario 2: in the community

	Disease (sensitivity 95%)	No disease (specificity 90%)	Total
Positive test	29	147	176
Negative test	1	1323	1324
Total	30	1470	1500

Positive predictive value = 16.5%; negative predictive value = 99.9%

Yield

The number of previously undiagnosed cases of disease which are picked up by a screening programme is the yield. This is obviously affected by the test itself, and by the prevalence of the condition. However, it is also affected by the frequency with which the test is applied (the recall period) and by the proportion of the population who take up the opportunity to be screened (Jepson et al, 2000).

Screening programmes need to be 'sold' to the population to encourage take-up, and tests which are invasive or painful are less likely to be popular. A recent development is the concept of the 'number needed to screen' (Rembold, 1998) which formalizes the means of deciding how many people should be screened for how long in order to prevent one death.

THE FUTURE

The number of screening programmes will continue to increase. Technological advance will provide new and more accurate tests for conditions of all kinds, and as intimated earlier, this will drive demand for screening. It is important that the potential pitfalls of screening are not overlooked, and that the ability to detect disease is not allowed to outstrip the ability to treat it.

If screening programmes are established before there is consensus within the medical community on their effectiveness, the public will receive conflicting messages, and uptake will suffer. Above all, if the prime tenet of screening — don't screen unless you can treat — is violated, the credibility of screening as a tool in medicine will be severely undermined. If this happens, it will be a long and hard job to re-establish public confidence. **HM**

Abbasi K (1998) To screen or not to screen? *Br Med J* **316:** 484

Boer R, de Koning H, Threlfall A (1998) Cost effectiveness of shortening screening interval or extending age range of NHS breast screening programme: computer simulation study. *Br Med J* **317:** 376–9

Jepson R, Clegg A, Forbes C (2000) The determinants of screening uptake and interventions for increasing uptake: a systematic review. *Health Technol Assess* **4**(14): 103–5

Mausner J, Kramer S (1985) *Epidemiology, An Introductory Text.* WB Saunders, Philadelphia

Quinn M, Babb P, Jones J (1999) Effect of screening on incidence of and mortality from cancer of the cervix in England: evaluation based on routinely collected statistics. *Br Med J* **318:** 904–8

Rembold C (1998) Number needed to screen: development of a statistic for disease screening. *Br Med J* **317:** 307–12

Scholefield J (2000) Screening (ABC of colorectal cancer). *Br Med J* **321:** 1004–6

Selley S, Donovan J, Faulkner A (1997) Diagnosis, management and screening of early localised prostate cancer. *Health Technol Assess* **1**(2): 75–85

Smith R (1995) Conflict of interest in The Times. *Br Med J* **310:** 1417

Stuttaford T (1995) Plain guide to the health checks every man and woman needs each year. *The Times* **May 16:** 15

Wilson J, Jungner F (1968) *Public Health Papers No 34: Principles and Practice of Screening for Disease.* World Health Organization, Geneva

Woolf S (1997) Should we screen for prostate cancer? *Br Med J* **314:** 989

KEY POINTS

- Screening has been very successful at reducing the consequences of disease.
- Screening is often assumed to be universally good — this is not always true.
- There are a number of criteria which must be satisfied before a screening programme becomes viable.
- The test itself must have certain attributes which define its effectiveness.
- The prevalence of the condition is an important factor in the success of screening.
- Screening will become more common, and care must be taken not to allow quality to suffer at the expense of quantity.