Choice of screening tests

PETER R. SIMPSON AND JOCELYN CHAMBERLAIN
From the Department of Community Medicine, University College Hospital Medical School, London

H. S. E. GRAVELLE
From the Department of Economics, Queen Mary College, London

SUMMARY In this paper, we consider one of the decisions that have to be made about a screening programme: which type of test to use. Our study shows that knowledge of the sensitivities, specificities, and costs of alternative tests is an inadequate basis for the choice of test. The monetary values of the different possible results of the test must also be estimated, or judgements made about the likely magnitude of these values. If judgements have to be made, they should be explicit, because different individuals are likely to judge differently, and their opinions will critically affect the choice of test.

If screening for early detection of a disease has been shown to be effective and is to be introduced as a service, three important questions need to be answered: (i) Who should be offered screening? (ii) How frequently should it be offered? and (iii) Which screening test should be used?

These three questions are interdependent, but for the sake of simplicity, this paper considers only the third. The basic literature on evaluation of screening (for example, McKeown, 1968) lists the evidence required about any screening test, including its safety, acceptability, sensitivity, specificity, cost, improvement in prognosis for cases detected by it, and how often it needs to be repeated. Unfortunately this knowledge alone does not enable a choice to be made between alternative tests. It must be supplemented by information on the monetary value of the benefits of the different alternatives. Assuming that alternative tests are being considered which are equally acceptable and safe, which need to be repeated at the same frequency, and which carry an equally good prognosis, the choice between them will depend on their sensitivities, specificities, and costs, and on the value assigned to the benefits of the different possible outcomes.

The choice of test

When a group of people are screened, the test divides them into four categories (true positive, false positive, true negative, and false negative) according to whether or not they have the disease and whether or not the test result is positive or negative (Thorner and Remein, 1961). This is illustrated in Table 1, which also shows the derivations of the definitions of sensitivity and specificity.

<table>
<thead>
<tr>
<th>Screening verdict</th>
<th>True diagnosis</th>
<th>No disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True positive</td>
<td>False positive</td>
<td>a + b</td>
</tr>
<tr>
<td>Negative</td>
<td>False negative</td>
<td>True negative</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>a + b + c + d</td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{a}{a + c} \)

Specificity = \( \frac{b}{b + d} \)

There will be different consequences for the person screened, and for the rest of the community, according to the category in which he or she is placed. For example, a true positive will receive earlier treatment and, if this is effective, improved health. The rest of the community will be affected by this result in a number of ways. Health service costs may possibly be reduced if the earlier treatment a patient receives is less expensive than treatment of his condition at an advanced stage; and these costs may possibly be increased if his improved prognosis means that he will survive to a greater age to make future demands on health services. The improved health of an individual may also cause changes in government tax revenue (as the result of a longer or fuller working life) and in government expenditure (for example, old age pensions).
Choice of screening tests

Finally, other members of the community may be altruistically concerned about the individual and they may benefit from his improved health. Similarly, values may be placed by society on the other three outcomes of the test.

Let $B_1$, $B_2$, $B_3$, and $B_4$ denote the values assigned to the true positive, false positive, false negative, and true negative outcomes respectively. It is reasonable to assume that in screening for most diseases, $B_1 > B_4 > B_2 > B_3$, with $B_2$ and $B_3$ being negative.

If $P$ is the prevalence of the disease in the people screened, $\pi_1$ the sensitivity (the proportion of people with the disease who have a positive test), and $\pi_2$ the specificity (the proportion of people without the disease who have a negative test), then the proportion of people in the four groups will be:

- True positive $P\pi_1$
- False positive $(1-P)(1-\pi_2)$
- False negative $P(1-\pi_1)$
- True negative $(1-P)\pi_2$

The expected net value of the test is equal to the sum of the probabilities of each of the four outcomes multiplied by the value of each outcome, less the cost of the test. If $C$ is the cost of the test, the expected net value will be:

$$B_1\pi_1 + B_2(1-P)(1-\pi_2) + B_3P(1-\pi_1) + B_4(1-P)\pi_2 - C$$

This expression will vary with the population screened and the type of test used. The population screened will determine the prevalence ($P$) and also the value of the four outcomes ($B_1$, $B_2$, etc.). For example, future health service costs are likely to be affected by the age of the people screened. In introducing screening for a given population, therefore, health authorities should have information not only on the sensitivities, specificities, and costs of alternative tests, but also on the values of the four outcomes of screening.

Valuation of outcomes is very difficult, and it is likely to be based on subjective judgements rather than objective measurements. Nevertheless, an attempt should be made (Gravelle, 1976); or, at the very least, the decision-makers should make their judgements explicit. If this is not done, decisions will be made on the basis of implicit judgements which are less likely to be questioned and may be based on incorrect criteria, ignoring some of the consequences of the screening programme.

One of these criteria is the yield of the test, that is, the number of true positive cases detected. It is plausible to suggest that since the purpose of screening is to detect disease, the test which does this best (or has the highest sensitivity) is the one that should be used. This criterion ignores the cost of the test; hence it would lead to a preference for test $X$ rather than test $Y$ when $X$ has, say, a 0.1% higher sensitivity, but costs twice as much as $Y$. This criterion also ignores the specificity of the test. If, for example, the majority of screened individuals are incorrectly referred for further investigation, this large number of false positives becomes important.

A somewhat more sophisticated criterion, which remedies one of the defects of the yield criterion, is to examine the cost per case detected. This has the merit that it does not ignore the cost of the test, but it is again based on the assumption that only the true positive outcomes influence the decision of the test to be used. It is a cost-effectiveness criterion in which the only effect or output that is assumed to be relevant is the number of true positives.

Youden (1950), by contrast, proposed an index which depends on both the sensitivity and the specificity of the test. The index is $I$ minus the sum of the false negative and false positive proportions. The index takes account of the correct diagnosis outcomes (true positives and negatives) but it ignores the cost of the test and assumes that true positive and true negative outcomes are equally valuable.

These three criteria and others like them are imperfect because they ignore some of the outcomes, value them arbitrarily, or ignore the cost of the test. In order to choose the best test, all the outcomes must be valued so that the total value of, and costs of, the different tests can be compared. However, even without specific information on the values of the outcomes, it is possible to choose between some tests on the grounds of efficiency. A test is inefficient if another test exists which either costs less and has at least as large a sensitivity and specificity, or costs no more and has greater sensitivity (specificity) and at least as large a specificity (sensitivity). Once inefficient tests have been eliminated, the choice among the efficient ones requires valuation of the four test outcomes.

**ALTERNATIVE SCREENING TESTS FOR BREAST CANCER**

To illustrate the issues involved in this kind of decision, evidence on sensitivities, specificities, and costs of alternative combinations of tests has been derived from preliminary results of a study of screening tests for breast cancer (Chamberlain et al., 1975). In this programme, each woman receives two independent clinical examinations, one by a doctor, and one by a nurse. The woman undergoes x-ray mammography and the films are subsequently read, independently and without knowledge of the clinical findings, by two radiologists. Thus, each of
four different observers makes an independent decision on whether or not the woman should be referred for a surgical opinion. If one of the four recommends it, the woman is referred through her general practitioner to a surgical outpatient’s department for further investigation and treatment if necessary. The screening test is classified as positive if referral is recommended, negative if it is not. All the women are followed up six months later, apart from those in whom cancer is diagnosed by screening (the true positives); and it is assumed that any cancers diagnosed during the six-month interval were, in fact, present at the time of the first screening but were missed by it (false negatives).

By comparing the verdict at the first screen with the diagnosis six months later, the women can be allocated to the four test outcomes—true positives, false positives, true negatives, and false negatives.

Because each observer records his or her verdict independently of the others, the sensitivity and specificity of each observer’s verdict can be calculated separately, and so can the eleven possible combinations of two or more observers, thus giving a total of 15 different combinations. As an example, Table 2 shows the verdicts for 1500 unselected women attending the clinic if screening had been limited to clinical examination by a nurse. Only eight of the 16 cancers would have been detected (sensitivity 50%), and 100 women without cancer would have been referred, giving a specificity of 93·3%.

The sensitivities and specificities of each of the 15 different combinations of observers are shown in Table 3. The negative relationship between sensitivity and specificity, whereby increases in sensitivity are usually gained at the expense of decreases in specificity, is also seen in Table 3. This is to be expected, because each observer will refer some women not referred by the others, so that the numbers referred will increase with increasing numbers of observers. Some of the additional women referred will have cancer (increased sensitivity) and some will be false positives (decreased specificity).

It must be emphasised that the figures quoted here should be looked at with caution. They are preliminary, and unreliable because the numbers are small. This applies particularly to those sensitivities based on only 16 cancers; for example, the difference in sensitivity between Tests 5 and 14 (each with the same specificity) is not statistically significant (McNemar, 1955). However, the purpose of quoting these figures here is not to report reliable estimates of sensitivity and specificity, but to illustrate the issues raised by decisions about screening strategy.

Estimates were made of the costs of the 15 different tests in 1974–75, the first financial year in which the clinic was operating. The price for all items of capital expenditure, such as x-ray equipment, on 1 April, 1975, was divided by the estimated life in years, and the resulting cost was assigned to one year of clinic operation. Running costs (including salaries, stationery and postage, x-rays and chemicals, heating and lighting, etc., but excluding specific research costs) were estimated on the assumption that prices on 1 April, 1975, remained constant for a whole year.

The capital expenditure assigned to one year (£2282) was added to the running costs (£17,712) to give a total annual expenditure of £19,994 for 1974–75. During this year, 2271 screening examinations were carried out, giving an estimated cost per screen of £8·80. This cost has been adjusted to the level of prices on 1 July, 1976, taking account of the rise of 28·3% in the general index of retail prices. The result is an estimated average cost of £11·29 in mid-1976. This is the cost of a test when all four observers are involved. In the same way, estimates have been made for each observer separately, and for the various combinations. These estimates have been adjusted to mid-1976 prices and are shown in Table 3.
Choice of screening tests

EFFICIENCY OF TESTS
The data in Table 3 show that some tests can be discarded as inefficient. Test 2 is of equal sensitivity to Test 6 but it is more specific and costs less. Test 9 is more efficient than Tests 11 and 13, and Test 7 is more efficient than Test 8.

THE EFFECT OF ASSUMPTIONS ABOUT VALUE
The choice among the remaining 11 efficient tests requires valuation of the true positive, false positive, false negative and true negative outcomes. In the absence of estimates of the benefits of outcomes, Table 4 gives some alternative, arbitrary, but perhaps not entirely implausible, assumptions about the values of $B_1$, $B_2$, $B_3$ and $B_4$. On assumption (a), the value of a true positive is £2000, of a false positive —£100, of a false negative —£1000, and of a true negative £10. The remaining assumptions alter these values in turn so that they then represent 50%, 150%, and 200% of those on assumption (a).

The expected net benefit was calculated for each test on the different assumptions. For example, with Test 1 on assumption (a):

\[
P = \frac{16}{1500} \pi_1 - 0.5 \pi_2 = 0.933 \ C = 63.09
\]

\[
B_1 = £2000 \quad B_2 = -£100 \quad B_3 = -£1000 \quad B_4 = £10
\]

Substituting these values in the formula derived earlier (see p. 167) gives an expected net benefit per screen of:

\[
\frac{1}{1500} (16 \times 0.5 \times 2000 + (1484 \times 0.067 \times -100) + (16 \times 0.5 \times -1000) + (1484 \times 0.933 \times 10) - 3.09 = £4.85
\]

Table 5 shows which of the tests has the highest expected net benefit on each of the different assumptions. It is clear that the choice of test is crucially dependent on judgements about the value of the screen outcomes. The higher the assumed value of a true positive outcome, or the greater (negatively) the value of the false negative outcome, the more likely it is that a test with a higher sensitivity will be chosen. For example, Test 2 is chosen on assumption (a), but a doubling of $B_1$ (assumption (d) ) or $B_3$ (assumption (j)) leads to the choice of Test 14, which has a larger sensitivity. Similarly, increases in $B_2$ or $B_4$ tend to favour tests with greater specificity. Compare, for example, the tests chosen on assumptions (a), (g), and (e) when the value of a false positive varies. The choice of test is also more responsive to changes in some outcome valuations than it is to changes in others. As Table 5 indicates, the choice of test is not affected by the range of variations in the value of a true negative.

By contrast, use of the three conventional criteria described earlier would lead to the choice of Test 15 (maximum yield), Test 12 (cost-effectiveness), or Test 15 (Youden index).

COMMUNITY PHYSICIANS’ CHOICE
It might be suggested that the points made above are valid in theory but unimportant in practice, because there would be little disagreement on the choice of test among experienced professional health policy-makers. Faced with the information on sensitivities, specificities, and costs of different tests, they would all make the same implicit valuations of test outcomes; so they would all choose the same test. In other words, there would be an ‘obviously’ best test, and therefore no need for attempts to estimate outcome valuations, or even to make judgements explicit.

To discover whether this is true of breast cancer screening, the data in Table 3 were sent to a small number of community physicians in academic and service posts, who were asked to select, in order of preference, the three tests they considered most appropriate for a screening service. They were not given details of the observers used in the various tests. The physicians were asked to make several assumptions: for example, that the figures shown in Table 3 would still hold if a larger number of women were screened, that the prognosis for cases detected
at screening would not vary between the alternative tests, and that all the tests were completely safe.

Of the 10 doctors approached, eight responded, and the answers they gave are shown in Table 6.

Table 6  Community physicians' choices of alternative tests

<table>
<thead>
<tr>
<th>Community physician</th>
<th>Test numbers (as in Tables 3 and 5)</th>
<th>1st choice</th>
<th>2nd choice</th>
<th>3rd choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>14</td>
<td>9</td>
<td>13*</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>9</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>10</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>9</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>9</td>
<td>8*</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>14</td>
<td>15</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>5</td>
<td>7</td>
<td>8*</td>
<td></td>
</tr>
</tbody>
</table>

*Indicates choice of inefficient test

Three of them put Test 5 first in order of preference (doctor and nurse, sensitivity 75%, specificity 89%, cost £5.49). The next most first popular preference was Test 14 (doctor and two radiologists, sensitivity 88%, specificity 89%, cost £10.10). There is clearly considerable disagreement among these professionals about which test would be most appropriate for a screening service: therefore the choice of test is not 'obvious'.

We thank Dr. J. L. Price, Dr. B. E. Nathan, Dr. Susan Ginks and other staff at Ravenor Park Clinic, Mrs. Rosemary Clifford, Mr. W. J. Corlett, and the community physicians. Thanks are due also to the Department of Health and Social Security for financial support of the study into the economic aspects of screening for breast cancer.

Reprints from Peter R. Simpson, Department of Community Medicine, University College Hospital Medical School, University Street, London WC1E 6JJ.

References


