Archives of Disease in Childhood 1990; 65: 792–795

CURRENT TOPIC

Evaluating screening tests and screening programmes

A E Ades

New laboratory and genetic technology is rapidly expanding the possible role of screening, and it has become increasingly important for paediatricians to evaluate critically all proposals for new screening programmes. There seems to be a wide belief that 'early diagnosis is beneficial, screening leads to early diagnosis, therefore screening is beneficial'. But in fact the benefits of screening are never self evident.[1] This article uses some recent examples from paediatrics to sketch out the key concepts needed to evaluate screening tests and programmes.

Sensitivity, specificity, and predictive value

There has been much interest in identifying infants at risk of sudden infant death syndrome (SIDS). Table 1(a) gives the results of a retrospective study of 145 children aged 8 days to 2 years who died unexpectedly at home and 154 controls.[2] After the death parents were questioned about symptoms observed in the previous week. A 'major symptom' was one that would require the attention of a doctor, though not necessarily hospitalisation or treatment. We can analyse this by considering the presence or absence of symptoms as a kind of prognostic test for unexpected death. Of the 145 deaths, 69 (48%) were preceded by major symptoms, while only 19 out of 154 (12%) control parents reported major symptoms. To some extent, then, the presence of symptoms can discriminate the cases from the controls.

The authors point out that only 12 of the 69 infants who died following symptoms had been seen by a doctor in the preceding 24 hours. They concluded 'that many deaths . . . might be prevented if doctors and parents were more aware of the importance of non-specific symptoms as markers of life-threatening illness'. But how would this work in practice?

Diagnostic tests are described in terms of their sensitivity, the proportion of true positives (cases) correctly identified by the test (0·48), and specificity, the proportion of true negatives (controls) correctly identified as negative (0·88). While sensitivity and specificity completely describe a test, its performance as part of a screening programme is also determined by the prevalence of the condition or event being screened for, in this case unexpected death.

Table 1(b) illustrates what would happen if parents were to apply the 'major symptoms' test in a typical population of 3 million infants, in which 1/600 children would be expected to die of SIDS each year,[3] or 1/30 000 each week. The probability of SIDS in children with major symptoms, known as the positive predictive value, would be 48/360 036 or 1·3/10 000. This can also be thought of as the proportion of occasions on which doctors' visits, if they were invariably effective, would be able to prevent a death. The remaining 99·987% of episodes of symptoms would be false positives: this clearly represents an unacceptable cost both in parental anxiety and in unnecessary medical call out. (Other issues that would need addressing are the effectiveness of a doctor's visit in preventing SIDS, and the possibility that in this study parents' recall of symptoms could be biased following the sudden death of their child.)

While subsequent papers pointed out the problems in the original concept,[4–6] this is not an isolated instance. All too often tests are developed in a high prevalence laboratory setting or case-control study and then proposed for use in a low prevalence population without first considering the implications, or even the numbers, of incorrect diagnoses.

Typically, the initial screen is an inexpensive but sensitive test intended to pick out those who are most likely to have the condition. The definitive diagnosis is then produced by one or more diagnostic tests, which may be costly or invasive. In the SIDS example the parent's detection of symptoms would have been the initial screen, triggering a doctor's visit, a clinical diagnosis, and hopefully the prevention of a death.

The same concepts of sensitivity, specificity, and positive predictive value apply equally to diagnostic tests that follow an initial screen. Reflex anal dilatation has been considered as a diagnostic test for child anal sexual abuse.[7] In this case the 'initial screen' is effectively performed by the police or social services who refer selected children already suspected of having been abused. The key issue here is the prevalence of child anal and sexual abuse in the groups referred for subsequent examination. In children referred by police the prevalence might be 50%. Assuming that the sensitivity of reflex anal dilatation is 0·60, a specificity of 0·99 would give a positive predictive value of 0·984. But if a wider population were to be examined, with a prevalence of say 5%, then the positive predictive value would be 0·76, and nearly a quarter of those tested positive would be false positives. Positive predictive value can be calculated from false positive and false negative rates, or equivalently from sensitivity and specificity, and a formula is given in table 2. Such calculations are a simple and powerful way of analysing

Institute of Child Health, London

Correspondence to:
Dr A E Ades,
Department of Paediatric Epidemiology,
Institute of Child Health,
30 Guilford Street,
London WC1N 1EH.

Table 1 Occurrence of major symptoms in a one week period as a predictor of unexpected death at home (a) in 145 cases and 154 controls[2] (b) the same sensitivity and specificity applied to a typical population with SIDS rate of 1/30 000 per week

(a) Case-control study	*SIDS cases*	*Controls*
Major symptoms:		
No (%) with	69 (48)	19 (12)
No (%) without	76 (52)	135 (88)
Total	145	154

Sensitivity= 69/145=0·48
Specificity=135/154=0·88

(b) Population, during one week	*SIDS*	*Not SIDS*	*Total*
Major symptoms:			
No (%) with	48 (48)	359 988 (12)	360 036
No (%) without	52 (52)	2 639 912 (88)	2 639 964
Total	100	2 999 900	3 000 000

Sensitivity=48/100=0·48
Specificity=2 639 912/2 999 900=0·88
Prevalence=100/3 000 000=1/30 000 per week
Positive predictive value=48/360 036=0·00013

Table 2 Parameters of screening tests and formulas to estimate them

		True diagnosis +	−
Test result	+	a	b
	−	c	d

N

Properties of the test independant of prevalence:

False negative rate	$\beta = c/(a+c)$
False positive rate	$\alpha = b/(b+d)$
Sensitivity	$1-\beta = a/(a+c)$
Specificity	$1-\alpha = d/(b+d)$

Properties of the test dependant on prevalence:

Positive predictive value (PPV)	a/(a+b)
Negative predictive value (NPV)	d/(c+d)
Proportion correct	(a+d)/N
True population prevalence, P	(a+c)/N
Observed prevalence	(a+b)/N
Bias	(a+b)/(a+c)

PPV and bias in terms of false positive and negative rates and prevalence:

$$PPV = \frac{1}{1+\frac{\alpha(1-P)}{P(1-\beta)}} \qquad Bias = \frac{1-\beta+\alpha(1-P)}{P}$$

To compare sensitivity, specificity in two or more groups, use standard methods for differences between, or ratios of, proportions.[15 22] Use paired methods to compare two tests applied to the same individuals.[22]

screening problems, and a recent paper applying them to reflex anal dilatation illustrates in a dramatic way how rapidly the positive predictive value deteriorates with low prevalence.[8]

Specificity or sensitivity

In many instances although the test produces a dichotomous positive/negative result, it is based implicitly or explicitly on an underlying continuum, sometimes reflecting the severity of the condition. Here it is important to distinguish between the inherent resolution of the test, which is its ability to discriminate between true positives and true negatives, and the rules used to interpret the results. In the hypothetical situation in the figure, we assume that there are two distributions of the underlying test measure X, in the unaffected population and in the diseased. There is also a cutoff level C. Responses to the right of C are declared positive, and those to the left negative; this is called the decision rule. By moving C to the left we can improve sensitivity as successively more of the weakly diseased are correctly declared positive. But at the same time more of the unaffected are declared positive, lowering specificity and positive predictive value.

The essential point is that sensitivity and specificity are in a trade off relation: either can be improved but only at the expense of the other. It is not possible to improve both simultaneously except by improving resolution. By contrast, the resolution does not depend on the cutoff at all, but remains a constant for any population of those with the condition (true positives) and normal controls (true negatives). Resolution can be estimated by taking the mean difference between controls and diseased, and dividing by the standard deviation of measurements within these two groups. The formula (figure) is similar to a *t* test of the difference between the normal and diseased group: higher resolution results from either a greater difference between diseased and normal groups, or from less individual variation within groups.

Choosing the appropriate decision rule

Ultimately it is the medical logic of the situation as well as the population prevalence of the condition that determines the choice of decision rule. A few examples illustrate that no set of principles is likely to replace common sense.

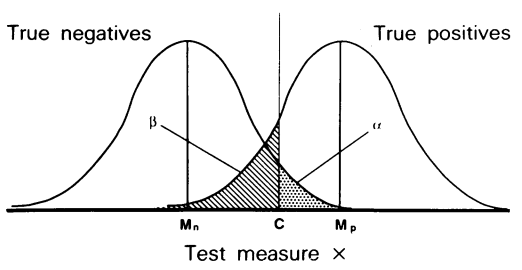
Antenatal sera are tested for rubella antibody, and women with titres under 15 IU are vaccinated postpartum. The cutoff separating 'posi-

Underlying distributions of a test measure X in true negatives (mean M_n), and true positives (mean M_p), with standard deviation (S).

Properties of the test and of the decision rule:

False positive rate	α=proportion of true negatives with X>C
False negative rate	β=proportion of true positives with X<C
Specificity	$1-\alpha$=proportion of true negatives with X<C
Sensitivity	$1-\beta$=proportion of true positives with X>C

Inherent properties of the test:
Test resolution $(M_p-M_n)/S$



tive' and 'negative' has effectively been shifted to higher levels, in a sense 'lowering sensitivity', so that women who may be weakly positive or borderline are vaccinated as well as those who are clearly negative. Because the procedure is both completely safe and an effective way of preventing congenital rubella syndrome, the inefficiency of vaccinating a number of women who are probably already immune is a small price to pay.

On the other hand, the same 15 IU cutoff would produce some depressing and misleading results if it were used in a survey of the prevalence of rubella immunity.[9] As it is believed that those with titres above 5 IU are effectively protected, this would be the most appropriate criterion.

A contrasting example is provided by screening blood donations for HIV antibody. Here, the cutoff C is moved over to the left, increasing sensitivity and decreasing specificity, so that borderline cases are treated as positive in the initial screen. A series of biologically distinct confirmatory tests can then be used to restore specificity to extremely high levels, weeding out false positives before disclosing a definitive positive diagnosis to a donor.

As a general principle, if the costs and benefits of true and false positives and negatives can be quantified in the same units, then a maximally cost effective cutoff can always be found, with its position again depending on the prevalence of true positives in the population. This merely emphasises that failure to detect and treat false negatives has to be weighed against the unnecessary anxiety and costs of confirmatory tests in false positives. Whether these factors can or should be quantified in practice in order to calculate a cutoff point, however, is open to question.

It is useful to distinguish properties of tests that depend on prevalence from those that do not (see table 2). It is not often appreciated that both the proportion of individuals who are correctly diagnosed and the bias depend on prevalence. Bias is the ratio of the observed prevalence to the true prevalence, and measures the extent to which the test overestimates or underestimates the true prevalence. The aim of a seroprevalence study must be to have the observed prevalence equal to the true prevalence (bias=1). This is achieved by having equal numbers, not equal rates, of false positives and false negatives. In a mass anonymous survey, with no possibility of repeat samples being taken, HIV seroprevalence in London neonates was found to be 0·0002.[10] Assuming 0·98 sensitivity, specificity would have to be better than 0·99995 if bias is to be effectively avoided. If specificity were to slip only to 0·999 (one false positive per 1000), the observed prevalence would be greater than the true prevalence by a factor of six. The implication is that prevalence must be taken into account even when choosing a test intrument for a prevalence survey. Table 2 gives an appropriate formula.

The individual patient

What can the individual patient be told about the probability of having the condition, given the results of the diagnostic test? The probability of being a true positive given a positive result is simply the positive predictive value, while the probability of being a true negative given a negative test result is the negative predictive value. If the test gives a continuous reading rather than a positive/negative result a more precise answer can be given so long as the distribution of the test measure is known for normal and for diseased populations, as illustrated in the figure. This calculation again takes into account the prevalence of the condition in whatever group or subgroup the patient belongs to. Dennis and Carter, who give a worked example for Duchenne muscular dystrophy, recommended that researchers prepare a graph from which they can read off the probability of being a true postive for any given test result.[11]

Evaluating a screening programme

The concepts developed so far provide a framework in which to evaluate a screening programme. However, sometimes a creative use of new measures might be important. In nearly every screening situation a proportion of the true positives are already known before, or without, the screen. For example, the prevalence of sensorineural hearing loss is approximately 1–2/1000. But those with congenital syndromes or infections may already have been detected before the 8 month screening, and a further proportion may have been brought to medical attention by parents. The remaining cases, who are not detected until screening, are likely to be those with milder forms of the condition and will be harder to detect. In terms of the figure, M_p is shifted to the left, lowering the resolution between the two groups, and if the cutoff remains fixed lowering sensitivity.

One approach has been to calculate sensitivity as the probability of picking up cases who have not already been detected. Effectively, all those already known to be positive are dropped from the set of true positives. In a similar vein, Rose has used the term yield to mean the proportion of those tested who actually benefit from having been screened.[12] A low yield could reflect not only a failure to pick up extra cases, but also ineffective treatment of those screened positive.

A careful decision analysis is invariably the best starting point for evaluating both existing and proposed programmes. Not only can it be performed before any screening programme is being operated, but carrying it out will immediately show what data is required to evaluate a programme, thereby pointing the way to the most useful research studies. The policy of screening women with a history of genital herpes and offering a caesarean delivery to those shedding virus in late pregnancy[13] was evaluated in a classic paper of this sort.[14] Drawing on published sources for data on the prevalence of a history, the proportion shedding virus, and the sensitivity of culture to detect virus, it could be shown that every case of neonatal herpes averted would cost $1·8 million, and that for every 11 neonatal deaths averted 3·3 women

would die from caesarean operations motivated by culture results.

Rather sophisticated study designs, perhaps including randomised trials of screening compared with not screening, are required to test programmes where there is doubt about the efficacy not only of screening but the treatment itself.[15 16] Once a programme is in place, however, it may be difficult to justify a formal evaluation of this sort. Nevertheless, in the traditional childhood screening programmes currently operating[17 18] evaluation is no less important—and in some instances, such as hearing and visual tests it may be long overdue. Fortunately, less exacting research designs are usually sufficient.

A basic recipe for evaluation of ongoing screening programmes will include three ingredients. First, a reporting scheme to identify any cases who were missed on the screen (false negatives), followed by an inquiry into whether they were genuinely missed (and, if so, why), or whether the results were misread or misrecorded. Secondly, the rate of false positives must be monitored in terms of extra time and costs required to produce a definite true negative diagnosis. Counselling might also be required to help avoid long term psychological consequences of false positives.[19 20] Thirdly, while efficacy of treatment cannot be proved without a formal clinical trial, an ongoing follow up of the true positives can provide data about the relation between the long term outcome and factors such as disease severity and age at start of treatment. At the same time, disadvantages of detecting a true positive must be weighed against early treatment advantages.[20] It was adverse psychological effects of positive diagnoses on parents that led to the abandonment of neonatal α_1 antitrypsin deficiency screening in Sweden.[21]

The computerisation of child health data is now sufficiently complete in some districts to make these types of evaluations perfectly feasible.

Conclusion

In any screening or testing programme, actual or proposed, a thorough analysis of costs and benefits is essential. Careful consideration must be given not only to the sensitivity and specificity of the tests, but also to the prevalence of the condition, and to the implications of—and numbers of—false positives and false negatives. The benefits of early diagnosis must be weighed against the unnecessary tests and procedures that may be carried out on false positives.

1 Hennekens CH, Buring JE. *Epidemiology in medicine*. Boston: Little, Brown, 1989.
2 Stanton I, Downham MAPS, Oakley JR, Emery JL, Knowelden J. Terminal symptoms in children dying suddenly and unexpectedly at home. *Br Med J* 1978;**ii**: 1249–51.
3 Balarajan R, Soni Raleigh V, Botting B. Sudden infant death syndrome and postneonatal mortality in immigrants in England and Wales. *Br Med J* 1989;**298**:716–20.
4 Watkins CJ. Can general practitioners prevent the sudden infant death syndrome? *Br Med J* 1989;**298**:1333–4.
5 Wilson AD, Downham MAPS, Forster DP. Acute illness in infants; a general practice study. *J R Coll Gen Pract* 1984; **34**:155–9.
6 Wright A, Luffingham GH, North D. Prospective study of symptoms and signs in acutely ill infants in general practice. *Br Med J* 1989;**294**:1661–2.
7 Hobbs CJ, Wynne JM. Child sexual abuse – an increasing rate of diagnosis. *Lancet* 1987;**ii**:837–41.
8 Harvey IM, Nowlan WA. Reflex anal dilatation: a clinical epidemiological evaluation. *Paediatric and Perinatal Epidemiology* 1989;**3**:294–301.
9 Orenstein WA, Herrmann KL, Holmgreen P, *et al*. Prevalence of rubella antibodies in Massachusetts schoolchildren. *Am J Epidemiol* 1986;**124**:290–8.
10 Peckham CS, Tedder RS, Briggs M, *et al*. Prevalence of maternal HIV infection based on unlinked anonymous testing of newborn babies. *Lancet* 1990;**335**:516–9.
11 Dennis NR, Carter CO. Use of overlapping normal distributions in genetic counselling. *J Med Genet* 1978;**15**:106–8.
12 Rose G. Epidemiology for the uninitiated: screening. *Br Med J* 1978;**ii**:1417–8.
13 Committee on Fetus and Newborn. Perinatal herpes simplex virus infection. *Pediatrics* 1980;**66**:147–8.
14 Binkin NJ, Koplan JP, Cates W. Preventing neonatal herpes: the value of weekly viral cultures in pregnant women with recurrent genital herpes. *JAMA* 1984;**251**:2816–21.
15 Armitage P, Berry G. *Statistical methods in medical research* 2nd Ed. Oxford: Blackwell Scientific Publications, 1987.
16 Morrison AS. *Screening in chronic disease*. Oxford: Oxford University Press, 1985.
17 Hall DMB, ed. *Health for all children*. Oxford: Oxford University Press, 1989.
18 Butler J. *Child health surveillance in primary care*. London: HMSO, 1989.
19 Sorenson JR, Levy HL, Mangione TW, Sepe SJ. Parental response to repeat testing of infants with 'false-positive' results in a newborn screening program. *Pediatrics* 1984;**73**: 183–7.
20 Marteau TM. Psychological costs of screening. *Br Med J* 1989;**299**:527.
21 McNeil TF, Thelin T, Aspegren-Jansson E, Sveger T, Harty B. Psychological factors in cost-benefit analysis of somatic prevention. *Acta Paediatr Scand* 1985;**74**:427–32.
22 Gardner MJ, Altman DG. Confidence intervals rather than p values: estimation rather than hypothesis testing. *Br Med J* 1986;**292**:746–50.