PRACTICE

5

IN BRIEF

- The use of diagnostic tests for diagnosis and screening
- The philosophy of screening
- The statistical tools for assessing the effectiveness of a diagnostic test
- The choice of the optimal diagnostic test

Further statistics in dentistry Part 5: Diagnostic tests for oral conditions

A. Petrie[1] J. S. Bulman[2] and J. F. Osborn[3]

A diagnostic test is a simple test, sometimes based on a clinical measurement, which is used when the *gold-standard* test providing a definitive diagnosis of a given condition is too expensive, invasive or time-consuming to perform. The diagnostic test can be used to diagnose a dental condition in an individual patient or as a screening device in a population of apparently healthy individuals.

FURTHER STATISTICS IN DENTISTRY:

1. Research designs 1
2. Research designs 2
3. Clinical trials 1
4. Clinical trials 2
5. Diagnostic tests for oral conditions
6. Multiple linear regression
7. Repeated measures
8. Systematic reviews and meta-analyses
9. Bayesian statistics
10. Sherlock Holmes, evidence and evidence-based dentistry

[1]Senior Lecturer in Statistics, Eastman Dental Institute for Oral Health Care Sciences, University College London; [2]Honorary Reader in Dental Public Health, Eastman Dental Institute for Oral Health Care Sciences, University College London; [3]Professor of Epidemiological Methods, University of Rome, La Sapienza
Correspondence to: Aviva Petrie, Senior Lecturer in Statistics, Biostatistics Unit, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London WC1X 8LD
E-mail: a.petrie@eastman.ucl.ac.uk

Refereed Paper
© **British Dental Journal 2002; 193: 621–625**

When clinicians want to make a **diagnosis** for a particular patient, they are often faced with a number of alternative possibilities. A diagnostic test, usually performed in conjunction with a clinical examination, may be used to exclude some diagnoses or categorise the patient as either having or not having a specific disease. Such a diagnosis is rarely definitive but it is possible for the clinician to use the test result to decide whether a disease is unlikely or probable in a particular patient. Diagnostic tests can also be used for **screening,** the objective of which is to determine whether members of an apparently healthy target population are likely to have the disease or condition under investigation. Often the screening test will be a first step in selecting people likely to have the condition, and this may be confirmed later using more refined procedures. The reasoning is that if such conditions can be easily detected in the early, pre-symptomatic stages, then subsequent treatment, cure, or prevention may be easier, less costly, and may have a better chance of succeeding.

A diagnostic test may suggest that a disease (eg oral cancer) is present on the basis of a *categorical outcome* (eg whether or not a red patch, white patch or ulcer of greater than 2 week's duration can be detected (Downer *et al.*, 1995)).[1] Sometimes the diagnostic test is based on a *continuous* measurement and the patient classified as having some disease if the level of the measurement exceeds (or is less than) a particular value, the **cut-off** value. For example, there is a suggestion (Streckfus *et al.*, 2001) that it might be possible to use the concentrations of the salivary protein c-*erb*B-2 as a marker for the initial detection and follow-up screening for the recurrence of breast cancer in men and women.[2] The cut-off is usually the upper (or the lower) limit of the reference range for that measurement. The reference range is the range of values which includes a large proportion (usually 95%) of the healthy (disease-free) individuals in the population. If the cut-off is set too low (in the instance in which high levels of the measurement indicate disease) then some people will be classified as having the disease when, in reality, they are disease-free. This can be costly in terms of money, time and the unnecessary psychological stress induced. If, however, the cut-off is set too high then patients with the disease will be missed, and this may have dire consequences for the individual, or in the case of an infectious disease, others in the population. An approach to choosing the optimal cut-off is explained in greater depth later in this paper.

THE PHILOSOPHY OF SCREENING

On the face of it, the concept of screening seems entirely laudable providing, of course, that suitable diagnostic tests exist. However, there are several points which have to be carefully considered. There may be serious ethical considerations. When an individual patient seeks help from a doctor or dentist, the implication is that any advice or treatment is being provided at the

request of the patient. In a screening exercise, however, the initiative may not come from the patient, but from the doctor or dentist proposing the screening. In other words, it is not the patient making the first move by saying to the dentist *'I think I have something wrong with me – please do what you can to help me'* but rather the dentist making the first move by saying to the patient *'Although you have no symptoms of the disease now, the result of this test could indicate that you may have a problem which could be treated'.*

Sensitivity and specificity

- Sensitivity is the proportion of individuals with the disease who are correctly diagnosed by the test
- Specificity is the proportion of individuals without the disease who are correctly identified by the test

It follows that if further investigation and treatment are being offered to patients following the positive result of a screening test, then the practitioner must be sure that:

- Adequate facilities for such investigation or treatment are available.
- There is an agreed policy on the stage in the disease process at which active treatment is needed.
- The proposed treatment will actually benefit the patient.
- Adequate funds are available to cover the costs involved.
- The subjects tested fully understand the implications of the result of a positive test. So, for example, if a child scored positive in a school screening for dental caries it would be wrong to say *'This examination has shown that you need dental treatment'*, but more correct to say *'This examination has indicated that you could probably benefit from a more detailed examination by your own, or the school, dentist, if you have not had such an examination recently.'*

Furthermore, before a mass screening procedure is implemented, the organisers need to satisfy themselves that:

- The disease or condition constitutes an important public health problem. This means either a condition with a high prevalence in the population, such as dental caries, or a condition which although not commanding a high prevalence is so serious as to be life-threatening or disabling, such as oral cancer.
- The cost of the screening programme for a relatively minor condition is not going to divert resources needed for the routine treatment of more serious conditions.
- The condition under investigation lends itself to screening, in that there is a recognisable latent or early symptomatic stage in the disease process. Dental caries, periodontal disease and oral cancer all clearly fulfil this criterion.
- A satisfactory and viable screening test exists. This should be:
  i) *Cheap:* that is, with a low per-capita cost, for obvious reasons
  ii) *Fast:* since lengthy tests mean not only that subjects will be reluctant to participate, but also that operative costs per capita will be greatly increased and the disease may progress.
  iii) *Acceptable:* in other words non-invasive, painless and not subjecting the subjects in any way to embarrassment or humiliation.
  iv) *Reliable:* in that different operators will always obtain similar results on the same subjects.

SENSITIVITY AND SPECIFICITY OF A TEST

Diagnostic tests used for both diagnosis and screening need to have a high degree of **validity,** and it is here that statistical analysis comes in. Validity in this context means the ability of the test to fulfil the required objectives; to indicate the presence of the disease or condition (i.e. to give a **positive test** result) for those with it, and to give a **negative test** result for those who are free of it.

An ideal test for a given condition would, of course, be positive for every person in whom the condition was present, and negative for those in whom it was absent. Sadly, this ideal is rarely, if ever, achieved. If it were perfect, the test would not be a test in this sense, but the gold standard diagnosis. Some tests will accurately identify all the positive cases, but only at the expense of returning a **false positive** result on some individuals who are free of the condition. Other tests may successfully identify all those who are free of the condition, but may also miss some of those who actually have it, returning a **false negative** result on some individuals with the disease. And others may be less than 100% successful in both directions.

In order to provide a measure of the relative validity of diagnostic and screening tests, the terms sensitivity and specificity have come into use.

Sensitivity is the probability (usually expressed as a percentage) that a subject with the disease will have a positive test result. With a perfect test, all those with the disease will have a positive test result and the sensitivity will then be 100%. A test with low sensitivity will fail to indicate disease in many of those that have it. The rate at which this occurs is called the *false negative rate*; sensitivity is equal to one hundred minus the false negative rate.

Specificity is the probability (usually expressed as a percentage) that a subject who is free of the disease will have a negative test result. Once again, with a perfect test all those free of the disease will have a negative test result and the specificity of that test will be 100%. A test with low specificity will falsely indicate the presence of disease in many of those that are free of it. The rate at which this occurs is called the *false positive rate*; specificity is equal to one hundred minus the false positive rate.

In algebraic terms:

Sensitivity = $Pr(T+|D+)$
Specificity = $Pr(T-|D-)$

where $Pr(A|B)$ is the probability of A given that B is true and is called a conditional probability. $Pr(T+|D+)$ therefore indicates the proba-

bility that the test (T) is positive, given that the disease (D) is positive (ie present).

Example

Even though the preventative programmes against dental disease are extremely effective, there are still individuals who develop large numbers of caries lesions. A screening programme for the early detection of these potentially high-risk individuals is useful in that special preventative programmes can be instituted for them which will result in a low cost-effectiveness ratio. Several micro-organisms have high cariogenic potential and have served as a basis for identifying individuals susceptible to caries. In particular, the level of lactobacilli has been shown to have a positive association with the incidence of dental caries. A 17-month long longitudinal study (Kingman *et al.*, 1988)[3] of 541 US adolescents initially aged 10–15 years was conducted with a view to establishing a screening test for high risk individuals, taking a bacterial level of lactobacilli $>10^5$ as a positive test result. A saliva sample was taken from every subject at baseline and the number of lactobacilli recorded. These measurements were related to the caries increment after 17 months, where at least three new lesions in the period was recorded as a positive disease result. Of the 541 children screened, 116 actually were disease positive after 17 months and 425 were disease negative. It is possible to display these results in a 2x2 contingency table of frequencies (Table 1):

Table 1 Table of frequencies showing the results of a screening test

	D+	D–
T+	17	29
T-	99	396
Total	116	425

This table indicates that the test successfully identifies 17 of the 116 subjects with the disease as positive, but records the remaining 99 falsely as negative. Similarly, it successfully identifies 396 of the 425 disease-free children, but gives a false positive result for the remaining 29.

From this table it is possible to calculate:

Sensitivity = 17/116 = 0.147 (ie 14.7% or approximately 15%)

Specificity = 396/425 = 0.932 (ie 93.2% or approximately 93%)

Thus there is a 15% chance that, using this test, a child with disease will screen positive, and a 93% chance that a child who is disease-free will screen negative.

POSITIVE AND NEGATIVE PREDICTIVE VALUES

At this stage it is necessary to refer to a popular misconception, which is to assume that just because a test has a sensitivity of 15%, it follows that a child who screens positive has a 15% chance of having the disease. To show that this is not the case, consider a tree diagram (Fig. 1) which

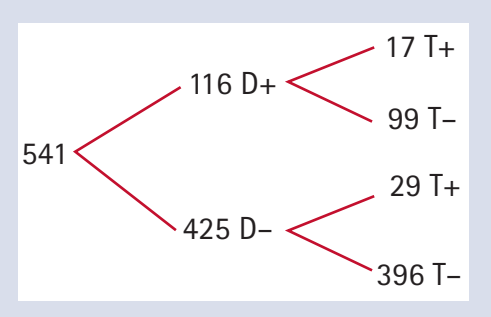

Fig. 1 Tree diagram for a test with sensitivity = 14.7%, specificity = 93.2% and estimated prevalence of disease = 21.4%

displays the same results as those in Table 1. This diagram clearly shows that although 46 children test positive, only 17 of them have the disease. Thus if a child screens positive, the probability that he/she will have the disease is 17/(17 + 29) = 0.370 or 37%, which is substantially greater than the sensitivity of 15%.

It is now possible to introduce two new terms, both of which assess the usefulness of the test in practice:

The **positive predictive value (PPV)** which is the probability (usually expressed as a percentage) that an individual who has a positive test result actually has the disease. In algebraic terms, PPV = *Pr*(D+|T+) which is 37.0% in the caries example

The **negative predictive value (NPV)** which is the probability (usually expressed as a percentage) that someone who has a negative test result does not have the disease. In algebraic terms, NPV = *Pr*(D–|T–) which is 80.0% in the caries example

Further investigation along these lines shows that the positive predictive value and the negative predictive value of a test depend both on the sensitivity and specificity of the test and also on the **prevalence** of the disease in the population. The prevalence of a disease is the proportion of individuals in the population who have the disease, which, in the context of a screening test, is taken as the *pre-test* probability or *a priori* probability that an individual has the disease. For a given test, the positive predictive value will be greater when the prevalence is high than when the disease prevalence is low. The reverse is true for the negative predictive value. Note that the sensitivity and specificity of a test are not affected by the prevalence of the condition. The prevalence is estimated in the sample by p which is D+ divided by the total number of individuals in the sample; so p = 116/541 = 0.214 (ie approximately 21%).

The calculations of the PPV and NPV shown in this paper require knowledge of the diagnostic test result as well as the true diagnosis in every member of a group of individuals. A later paper in this series, Part 9: Bayesian Statistics, describes an alternative method clinicians can use in order to determine the PPV of a test if they only have knowledge of the pre-test probability of the disease and the test result for a given patient. In these circumstances, the PPV is usually called the *post-test* or *posterior* probability of the disease.

Predictive values

- The positive predictive value is the proportion of individuals with a positive diagnostic test result who have the disease
- The negative predictive value is the proportion of individuals with a negative test result who do not have the disease

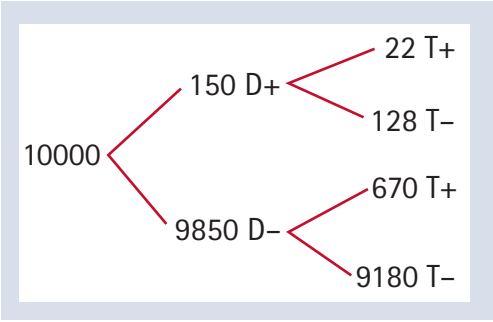

Fig 2(a) Tree diagram for a test with sensitivity = 14.7%, specificity = 93.2% and estimated prevalence = 1.5%

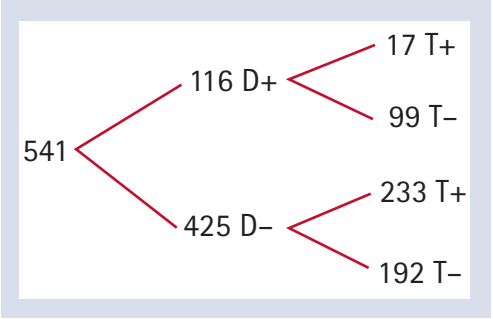

Fig 2(b) Tree diagram for a test with sensitivity = 14.7%, specificity = 45.2% and estimated prevalence = 21.4%

As an illustration of the relationship between the positive predictive value of a test and the prevalence of a disease, consider a sample of 10,000 children from a different population; the disease prevalence (ie a caries incidence of at least 3 new DMFS) in this population is only 1.5%. If the sensitivity and specificity of the test remain at 14.7% and 93.2% respectively, a quick calculation will show (Fig. 2a) that of the 150 children with the disease, 22 will screen positive and 128 will screen negative; of those 9,850 without the disease, 670 will screen positive and 9,180 will screen negative. Thus if a child screens positive, there would be a 22/692 = 0.032 or 3.2% chance that he/she has the disease, ie the PPV = (22)/(22 + 670) = 0.032. Hence, for this lactobacilli test, the PPV is reduced from 37% when the prevalence is 21% to approximately 3% when the prevalence is only 1.5%. The reverse is true for the NPV which increases from 80% to approximately 99% in these circumstances.

This result indicates the important difference between a diagnostic test when it is used for screening and diagnosis. Even if precisely the same test is used, when it is applied to screen a population which is apparently healthy (with respect to the disease or condition being studied), the *a priori* or pre-test probability of having the disease (ie the prevalence) is low. In this case, the positive predictive value or post-test probability is also likely to be low. On the other hand, if a patient complains to his doctor or dentist about a series of symptoms, and a case history leads the doctor or dentist to formulate a hypothesis that the patient has a certain disease, the probability that the hypothesis is correct should be relatively high before the test is performed, and a positive test result acts as a confirmation of the diagnosis. If, for example, with all the information given by the patient, the doctor or dentist thinks (with probability, say 0.75) that the patient has disease X, the predictive value of a positive test may be close to 100%. A good clinician formulates a hypothesis (and maybe an alternative) and requests one or two diagnostic tests. Note that a less skilled clinician might have no clear idea of the cause of the patient's problem, and may request a long battery of tests, not realising that, if a large number of tests are performed, there is a high probability that something will turn up positive purely by chance.

In order to investigate the relationship between the positive predictive value and specificity, consider a different test (relating to the level of *mutans streptococci,* say) in the first sample of children for whom the estimated disease prevalence is 21.4%; suppose that the sensitivity of this test is also 14.7% but its specificity is only 45.2% (Fig. 2b) rather than the 93.2% obtained for the lactobacilli test. Here (17 + 233) children test positive of whom only 17 have the disease. So if a child screens positive, the chance of him/her having the disease (ie his/her positive predictive value) is only 17/(17 + 233) = 0.068. Thus (rounding the percentages), if the prevalence (21%) and the sensitivity (15%) remain unaltered, a test with a specificity of 45% instead of 93% lowers the PPV from 37% to 7%.

CALCULATIONS OF THE MEASURES OF THE TEST EFFECTIVENESS

The calculations for these statistics, which are estimates of the true population values, may now be summarised using the notation of Table 2, a generalised 2 x 2 table of frequencies:

Sensitivity = $Pr(T+|D+) = a/(a + c)$

Specificity = $Pr(T-|D-) = d/(b + d)$

Positive Predictive Value = $Pr(D+|T+) = a/(a + b)$

Negative Predictive Value = $Pr(D-|T-) = d/(c + d)$

Prevalence, $p = (a + c)/n$ (This is the pre-test or *a priori* probability of having the condition)

Finally, it may be of interest to obtain an estimate of the prevalence of the condition if the sensitivity and specificity of the test are known (and expressed as probabilities), and $Pr(T+)$ is the proportion of individuals in the sample testing positive. Then,

$$p = \frac{\text{Specificity} + Pr(T+) - 1}{\text{Sensitivity} + \text{Specificity} - 1}$$

that is, if sensitivity = specificity = 0.9, and 12% of the sample test positive, then the estimated prevalence is:

Table 2 Table of frequencies with frequencies expressed in general terms

	D+	D-	Total
T+	*a* (true positive)	*b* (false positive)	*a + b*
T–	*c* (false negative)	*d* (true negative)	*c + d*
Total	*a + c*	*b + d*	*n*

$$p = \frac{0.9 + 0.12 - 1}{0.9 + 0.9 - 1} = 0.025 \text{ (ie 2.5\%)}$$

It is a naive mistake to confuse the proportion who have a positive test result with the prevalence of the disease. As shown above, if 12% of the tests are positive, this does not mean that 12% of the individuals have the disease!

Note that the sensitivity, specificity and positive and negative predictive values of a test are generally evaluated using sample data and are only estimates of their true values in the population. It is possible, using the statistical theory of the binomial distribution, to calculate the standard errors of the estimates, and use the latter to determine confidence intervals for the population values.

CHOOSING A TEST

Although the aim should be to devise a test which has both a high sensitivity and a high specificity, it must be recognised that often one has to be sacrificed in order to accommodate the other since they are inversely related, sensitivity increasing as the specificity decreases and *vice versa*. Hence it is necessary in any given situation to establish what is required from the test and the consequences of false positive and false negative results. For example, if the test is to be used to screen for a fatal non-infectious disease, then it is important to be able to reassure individuals that they do not have the disease and avoid the risk of false positives; here, specificity and the NPV are of prime importance. If, on the other hand, the disease is treatable but infectious, it will be important for the screening test to have high sensitivity, ensuring that the false negative rate is low and not many true cases of the disease are missed. A confirmatory test with high specificity, and therefore a low false positive rate, can then be used on those individuals who were positive on the initial screen. If the test is based on a continuous measurement so that a test result is positive if an individual's value exceeds (say) a particular cut-off value for the measurement, then it is possible to alter the sensitivity and specificity of the test by changing the cut-off. If this cut-off is raised, fewer individuals with the disease will be classified as positive, the sensitivity will decrease and there will be more false negatives. At the same time, more individuals will appear to be disease-free so that the specificity will increase and there will be less false positives. If the cut-off for this measurement is lowered, the reverse is true and the sensitivity will increase and the specificity will decrease.

As an illustration of these concepts, consider the lactobacilli example for the detection of potentially high risk caries children. When a bacterial level of $>10^5$ was chosen as the cut-off, the sensitivity and specificity of the test were approximately 15% and 93%, respectively. However, when a level $> 10^6$ was chosen as the cut-off, fewer high risk children were identified and the sensitivity and specificity of the test were 2% and nearly 100%, respectively. On the other hand, when a level $> 10^2$ was chosen as the cut-off, the sensitivity and specificity of the test were 79% and 50%, respectively.

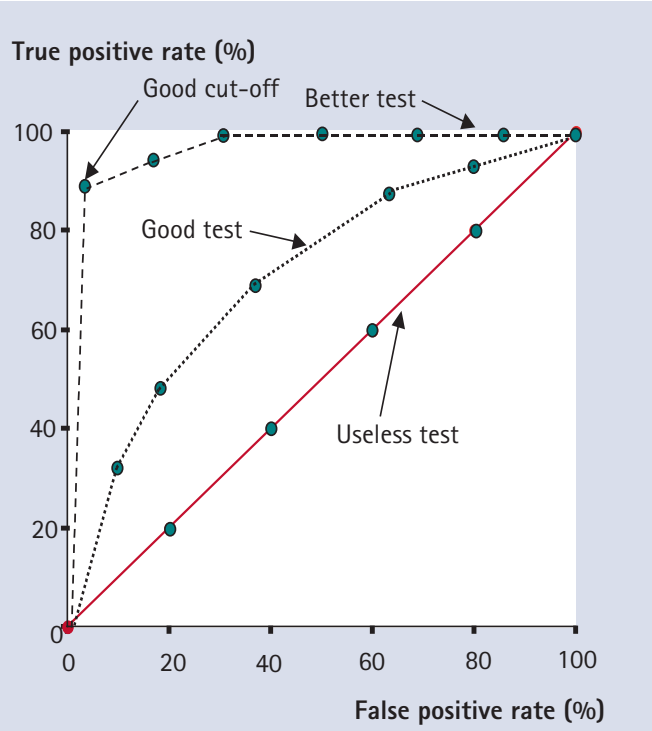

Fig. 3 ROC curves for 3 tests. The dots represent various cut-off values for each of the tests

In order to decide on an optimal cut-off for a given test, it is possible to draw its **receiver operating characteristic** (ROC) curve. If rates are expressed as percentages, this is a plot (Fig. 3) of the sensitivity (ie the true positive rate equal to 100 minus the false negative rate) on the vertical axis against 100 minus the specificity (ie the false positive rate) on the horizontal axis for various cut-off values for the measurement. If the true and false positive rates are equal for all cut-off values for a test, the resulting curve is a diagonal straight line from the bottom left-hand corner to the top-right hand corner. The ROC curve should never pass below the diagonal, as this would imply that the false positive rate is greater than the true positive rate. Choosing a cut-off from the ROC depends on the specific requirements of the test and the implications of false negative and false positive results The perfect cut-off for any test is one which produces no false positives and no false negatives and so is the point at the top left hand corner of the ROC diagram. If a very good test can be regarded as that with a high true positive rate and consequently very few false negatives, its curve would rise steeply from the bottom left-hand corner, almost reaching the top-left-hand corner, before flattening out. It is possible to use the ROC curve to choose the optimal cut-off value for a particular test by specifying the required sensitivity and specificity of the test, a non-statistical decision based on the clinical implications of a false negative and false positive result. In addition, two or more tests may be compared by evaluating the area under each of the ROC curves; generally, the test with the greater area is the 'better' test overall.

1. Downer M C, Evans A W, Hughes Hallet C M, Jullien J A, Speight P M, Zakrzewska J M. Evaluation of screening for oral cancer and precancer in a company headquarters. *Community Dent Oral Epidemiol* 1995; **23:** 84-88.
2. Streckfus C, Bigler L, Dellinger T, Dai X, Cox W J, McArthur A, Kingman A, Thigpen J T. Reliability assessment of soluble c-erbB-2 concentrations in the saliva of healthy women and men. *Oral Med* 2001; **91:** 174-179.
3. Kingman A, Little W, Gomez I, Heifetz S B, Driscoll W S, Sheats R, Supan P. Salivary levels of *Streptococcus mutans* and lactobacilli and dental caries experiences in a US adolescent population. *Community Dent Oral Epidemiol* 1988; **16:** 98-103.