Serodiagnostic tests for tuberculosis: a need for assessment of their operational predictive accuracy and acceptability

J.M. Grange¹ & A. Laszlo²

There have been numerous unsuccessful attempts to develop clinically useful serodiagnostic tests for tuberculosis. Although the large number of published reports clearly show that antibody levels are significantly higher in patients, as a group, than in a control population, little consideration is given to the value of the tests in various operational situations. In this paper we review the criteria generally used to assess the usefulness of a diagnostic test and introduce two new concepts—namely, operational predictive accuracy and operational acceptability.

Background

The accurate diagnosis of active tuberculosis is essential for the control of this disease in both high-incidence and low-incidence regions and for the welfare of individual patients. Despite an enormous amount of research since the time of Robert Koch we still have no simple, sensitive and specific test that will distinguish most or all patients with active tuberculosis from those with quiescent lesions, previous BCG vaccination, or other diseases or even from those who are completely healthy.

Being an obligate pathogen, the demonstration of the tubercle bacillus in clinical material by microscopy or cultural techniques is a clear indication of active tuberculosis. Unfortunately, microscopy is relatively insensitive and culture techniques are time-consuming and expensive. Attempts have been made to improve the sensitivity and speed of the detection of tubercle bacilli, or components thereof, by such techniques as radiometric determination of bacterial growth (1), gas chromatography/mass spectroscopy (2) and DNA hybridization (3). All of these have met with problems of either sensitivity or cost. Accordingly, there is still considerable interest in the development of tests based on cell-mediated or humoral immune responses to (supposedly) specific antigens extracted from the tubercle bacillus. Although cell-mediated immunity may be assessed in vitro by several techniques, these usually require sophisticated laboratory facilities: the only practical way to investigate cell-mediated immunity to mycobacteria in most clinical situations is to perform skin tests with appropriate antigens. The disadvantages of tuberculin testing as a diagnostic tool are well known (4); in addition to various problems of specificity, the patient needs to be seen on two occasions, 48 or 72 hours apart. For these reasons, numerous (hundreds, or possibly thousands) of attempts have been made to develop a serological test for tuberculosis, yet none has proved suitable for routine clinical use.

Development of serological tests for tuberculosis

The history of serodiagnosis of tuberculosis commenced with Robert Koch's unsuccessful attempts to diagnose this disease by direct agglutination of tubercle bacilli. Later workers applied virtually every available immunoassay to the problem: agglutination of sensitized erythrocytes and other carrier particles, complement fixation, precipitation in fluid and gels,
and radioimmunoassay (5). Since the pioneering work of Nassau in 1976 (6), virtually all studies have been based on the enzyme-linked immunosorbent assay (ELISA) (7).

The early attempts to diagnose tuberculosis serologically were based on the use of crude or partially fractionated antigens, and invariably there was an unacceptably high overlap between the results of tests on proven cases of tuberculosis (usually smear positive) and healthy control subjects. This overlap was even more evident in comparisons of antibody levels between patients with smear-negative tuberculosis and patients suffering from other diseases likely to be confused clinically with tuberculosis. Accordingly, these tests were particularly unhelpful in those very cases for which a good diagnostic test is most needed. As the tubercle bacillus contains both shared and species-specific antigens (8), the logical step forward was to extract and purify the latter by various physico-chemical separation techniques. In practice, such purification proved notoriously difficult although a few apparently pure species-specific antigens have been isolated and evaluated (7, 9). Although highly purified antigens are more specific than crude bacillary antigens such as PPD, culture filtrates, or sonicates, they have nevertheless proved disappointing in clinical use. Other workers have attempted to refine diagnostic tests using unpurified antigens by assaying the antibody response in the various immunoglobulin classes (10) and in the subclasses of IgG (11). Although these procedures increased the discrimination of the assay, the introduction of such complex, time-consuming and expensive techniques into routine clinical use would be unjustifiable.

The availability of monoclonal antibodies to mycobacterial antigens (12), antigenic products of cloned mycobacterial genes (13), and synthetic carbohydrate antigens (14) again raised hopes for near-perfect serodiagnostic tests for both tuberculosis and leprosy but, although some progress has clearly been made, these hopes have not so far been realized.

Serological tests

Problem of specificity

The sensitivity of the assays for anti-mycobacterial antibodies is not a problem: modern immunoassay techniques such as ELISA are able to detect such antibodies in virtually every individual. Indeed there is no longer any doubt that ELISA is a simple, viable and reproducible technique for use in studies on the serodiagnosis of tuberculosis. The problems lie with the specificity and are the result of two paradoxes (15). First, despite the fact that mycobacteria are very good adjuvants, the humoral immune response in disease caused by such bacteria is usually a very weak one. Secondly, although many of the more recent studies have been based on purified antigens that appear to be specific for the tubercle bacillus, there is still an unacceptable overlap between antibody levels in diseased and healthy individuals and an even greater overlap between levels in the former and in individuals with lesions and symptoms that could be confused with those of tuberculosis.

In general, serodiagnostic tests are useful in acute diseases in which a rising titre of antibody may be demonstrated in sequential serum samples (e.g., many viral infections). This approach is unsuitable for chronic diseases such as tuberculosis, although in a patient with HIV infection and suspected tuberculosis, examination of a number of sequential serum samples (which, fortunately, had been saved) showed a distinct rise in anti-mycobacterial antibody even though the highest level was within the defined "normal" range for that assay (16).

Serodiagnostic tests are also useful in those infections caused by pathogens that contain dominant species-specific antigens that regularly elicit antibody responses in the diseased host (as, for example, in diphtheria). In the case of tuberculosis, there is apparently no such dominant specific antigen—indeed most of the antibody response is directed towards shared mycobacterial antigens (17). Furthermore there is, probably as a result of genetic factors or previous sensitization by environmental mycobacteria, a considerable patient-to-patient variation in the antigens to which an individual responds. Thus, very specific tests based on binding competition with monoclonal antibodies to species-specific antigens of Mycobacterium tuberculosis are not positive in all cases of multibacillary (smear-positive) pulmonary tuberculosis (12). This problem of response variation can be partially overcome by the use of a number of different monoclonal antibodies (18) or purified, but not necessarily species-specific, antigens (19), but even such relatively complex and expensive multiple assays do not show a complete correlation with clinical, radiological and bacteriological diagnoses.

Design and evaluation

In addition to difficulties caused by the underlying biological nature of specific antibody production in tuberculosis, many published studies show gross inadequacies in their design and, more especially, in their evaluation. In many published reports antibody levels in sera from defined, usually bacteriologically confirmed, patients with tuberculosis are compared statistically with those of healthy controls, usually by Student's t-test when the data are parametric (or even
Serodiagnostic tests for tuberculosis in operational situations

when they are not, as is often the case). In fact, an assessment of the statistical significance of the difference is virtually valueless as it depends largely on the number of patients and controls included. The crucial factor from the diagnostic point of view is the degree of overlap between those with and without active disease. In most studies this overlap is considerable and serious problems as to the definition of the upper limit of “normal” are raised. This problem is accentuated when two or more separate antigens are used and patterns of “positive” results need to be determined. Although very high specificity has been claimed for a dual-antigen tuberculosis ELISA (20), the doubtful results (i.e., those positive with one antigen and negative with the other) were interpreted in the light of clinical and radiological data. An unbiased evaluation of that test shows it to be less specific.

**Criteria for assessment**

In a meaningful mathematical evaluation of a serodiagnostic test, four factors require consideration: sensitivity, specificity, efficiency (or accuracy), and predictive value.

*Sensitivity* is a measure of the ability of the test to detect a characteristic of a disease when it is present, i.e., to avoid “false” negative results. It is usually defined as the proportion of patients with the disease who have a positive test and may be expressed as:

\[ \frac{TP}{TP + FN} \times 100 \]

where \( TP \) = true positive, i.e., diseased individual correctly classified by the test, and \( FN \) = false negative, i.e., diseased individual incorrectly classified by the test.

The *specificity* of the test is a measure of its ability to clearly discriminate healthy from diseased individuals, i.e., to avoid “false” positive results, and is expressed as:

\[ \frac{TN}{FP + TN} \times 100 \]

where \( TN \) = true negative, i.e., non-diseased individual correctly classified by the test, and \( FP \) = false positive, i.e., non-diseased individual incorrectly classified by the test.

It is generally agreed that in serological diagnostic procedures, the sensitivity is determined by the technique whereas the specificity is affected by the antigen.

Sensitivity and specificity are really only meaningful in those cases when the characteristic detected or assayed is universally present in disease and invariably absent in health or in unrelated disease processes. In such cases, it is theoretically possible to develop a test which is both 100% specific and sensitive. For example, tubercle bacilli are present in all cases of tuberculosis but are never seen in, or cultured from, the tissues or secretions of healthy individuals. In the case of the serodiagnosis of this disease, no antibody, or level of antibody appears to be disease-specific. Repeated experience has shown that an assay may be rendered sensitive enough to detect antibody to antigenic determinants of *M. tuberculosis* in all patients but, in so doing, it will also detect such antibody in a high proportion of those without the disease. In such circumstances, sensitivity and specificity must be defined by selecting a suitable upper limit of “normal” values. Clearly, a high cut-off point will yield a test with high specificity and low sensitivity and vice versa.

The *efficiency* (or accuracy) of the test indicates the percentage of individuals (diseased and non-diseased) correctly classified by the test and is expressed by the following formula.

\[ \text{Efficiency} = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \]

The *error* of the test is the converse, i.e., the percentage of all subjects who are misclassified. A related value, the *certainty of diagnosis*, is the sum of the sensitivity and specificity (20). These values are of use when comparing two or more tests but have limited application in the evaluation of the test for clinical or epidemiological use.

For any preselected sensitivity or specificity, the significance of a positive or negative value is critically dependent on the prevalence of the disease in the region under study. Thus the chance of a negative value being correct, i.e., the *predictive value of negativity* (PVN), is high in regions where the disease is uncommon, while that of a positive value being correct, i.e., the *predictive value of positivity* (PVP), is high where the disease is common. The predictive values are therefore expressions of the ratio of true negatives to total negatives, and true positives to total positives. If the prevalence of disease is known, these predictive values may be calculated by use of Bayes' theorem (14, 21):

\[ \text{PVN} = \frac{(1 - \text{prevalence}) \times (\text{specificity})}{(1 - \text{prevalence}) \times (\text{specificity}) + (\text{prevalence}) \times (1 - \text{specificity})} \]

\[ \text{PVP} = \frac{(\text{prevalence}) \times (\text{sensitivity})}{(\text{prevalence}) \times (\text{sensitivity}) + (1 - \text{prevalence}) \times (1 - \text{sensitivity})} \]

These predictive values are usually calculated for tests that are applied indiscriminately to the entire population and were used, for example, in the evaluation of mass miniature-radiography campaigns (21). In indiscriminate screening for disease, the predictive value of positivity (PVP) is of the utmost importance.
because a reported positive result usually leads to investigation or treatment of the individual which may be unnecessary. This circumstance dictates high specificity for any acceptable test.

Even in high endemic areas, the prevalence of tuberculosis in the general population is only in the order of 1%. If a good test with, for example, specificity and sensitivity both of 95% were to be used on such a population, it would only yield a PVP of 16.1% (22). To achieve acceptable PVP values in the 90–95% range, one would have to apply such a test to a population with a prevalence of disease close to 50%. In practice, higher PVP values may be achieved by increasing the specificity, i.e., by modifying the cut-off point separating positive from negative test results. However, as discussed above, increased specificity can only be gained at the expense of sensitivity. Thus, to achieve higher PVP values it is necessary to find ways of increasing the prevalence of disease in the screened population. The actual prevalence of disease cannot be altered but the relative prevalence can be increased by only testing individuals in preselected groups. These include those with clinical signs or symptoms compatible with tuberculosis (e.g., chronic cough, radiological abnormalities, or features suggestive of non-pulmonary tuberculosis) or those in high-risk groups, such as contacts of index cases or those with conditions that predispose to tuberculosis (e.g., HIV infection, other immunodeficiencies, and dust-associated pulmonary disease).

Another way to increase the PVP is to screen a population with a test of high sensitivity and then to apply an unrelated test of higher specificity to those who were positive in the primary screen. This method of pre-screening, also called combination testing, increases the specificity in series testing and the sensitivity in parallel testing. This approach could, if properly applied, make tuberculosis serodiagnosis a viable proposition.

**Operational situations**

**Value of serodiagnostic tests**

Useful though estimates of PVP may be, we are of the opinion that they do not truly evaluate the applicability of a test for routine clinical use. We therefore introduce two new terms: *operational predictive accuracy* and *operational acceptability*. These will vary according to the principal purpose of the test and the circumstances under which they are used.

In this respect, all new tests have two major areas of application. First, in those developing countries where tuberculosis is common but health care resources are limited and where the major aim of disease control is to detect the open or infectious cases and to render them rapidly and permanently noninfectious by modern short-course chemotherapy. Secondly, in the industrially developed nations where tuberculosis is relatively uncommon and therefore often overlooked.

In developing countries with a high prevalence of tuberculosis and limited resources for therapy, sputum microscopy is the mainstay of diagnosis of infectious cases since, for practical purposes, a patient with enough bacilli (at least 5000 per ml of sputum) to be detected by a competently performed microscopic examination may be regarded as infectious (23). The advantages of microscopy are the relatively simple technical procedures, the robustness and straightforward maintenance requirements of the equipment, the rapidity of the results and, as mentioned above, the close association of smear positivity and infectiousness. The major disadvantage is the human error due to fatigue and boredom induced by the lengthy and monotonous process of examining the smears, especially when ultraviolet microscopes are not available. Despite this, any serodiagnostic test will have to compete closely with microscopy for *operational acceptability* in respect of case, reliability, rapidity, cost, maintenance of equipment, safety (particularly in obtaining, handling and transporting blood that is possibly infected with HIV or hepatitis viruses), and acceptability to the patients. In addition, the test would have to compete closely with microscopy for its ability to detect open and infectious cases. For this purpose it would be necessary to determine the *operational predictive accuracy* by determining the fraction of smear-positive individuals who are seronegative and of those with minimal or inactive disease who are seropositive. Clearly, both the *operational predictive accuracy* and *operational acceptability* will differ when the tests are used for special purposes in the same region, e.g., for the diagnosis of tuberculous meningitis (24).

In the industrially developed nations, tests are principally required for the differential diagnosis of chronic cough and weight loss, pyrexia of unknown origin, radiological shadows, night sweats, lymphadenopathy, space-occupying lesions and a range of vague "orthopaedic" symptoms. In many cases, the cause will not be tuberculosis and therefore the occurrence of misleading "positive" results would have to be very low or clinicians would rapidly lose confidence in the test. Thus the estimation of the *operational predictive accuracy* would not be based on the discrimination between smear-positive and healthy individuals but between those with smear-negative or non-pulmonary disease and those with compatible signs and symptoms. With existing tests,
Serodiagnostic tests for tuberculosis in operational situations

the upper limit of "normal" values would have to be set so high that many individuals with active tuberculosis would have antibody levels within this defined range. Furthermore, as tuberculosis is often overlooked as a diagnostic possibility, the test would preferably be one that would be readily included in the routine investigations conducted locally rather than one requiring a special request or referral to a specialist laboratory. Thus, simplicity, reliability, robustness, and extreme specificity would be mandatory requirements for a test in this situation.

Conclusions

Although numerous published serological tests for tuberculosis show highly significant differences between antibody levels in the chosen patient and control groups, very few of them have been subjected to a careful analysis of their potential value in clinical practice and even fewer have actually been evaluated prospectively in a clinical setting. Much will be saved in time and effort, as well as in technical and financial resources, if attention is given to such practical considerations during the planning, conduct and evaluation of any future investigations.

Résumé

Nécessité d'évaluer la valeur prédictive et l'acceptabilité opérationnelles des épreuves de sérodiagnostic de la tuberculose

Pour lutter contre la tuberculose, il est essentiel de pouvoir diagnostiquer avec précision les formes évolutives de la maladie; or, malgré de nombreuses recherches, il n'existe encore aucune épreuve simple, sensible et spécifique répondant à ce besoin. De nombreuses tentatives ont été faites pour mettre au point des épreuves de sérodiagnostic, mais aucune n'est utilisable en routine. Parmi les méthodes publiées, très peu ont été soumises à une analyse de leurs avantages potentiels en clinique, et moins nombreuses encore sont celles qui ont fait l'objet d'une évaluation prospective en pratique médicale courante.

Les principales caractéristiques qui doivent être déterminées lors de l'évaluation d'une épreuve de diagnostic sont la sensibilité (aptitude à détecter une manifestation de la maladie lorsqu'elle est présente) et la spécificité (aptitude à distinguer entre les personnes bien portantes et les malades). Cependant, ces caractéristiques ne permettent pas de juger de l'intérêt réel d'une épreuve en pratique clinique courante. Pour cela, il faut prendre en compte la valeur prédictive opérationnelle et l'acceptabilité opérationnelle qui varient selon l'objectif de l'épreuve et les circonstances dans lesquelles elle est utilisée.

Toutes les épreuves de diagnostic de la tuberculose doivent pouvoir être utilisées dans deux principaux types de situation: premièrement, dans les pays en développement où la tuberculose est courante et où la lutte s'appuie sur la détection des cas d'infection, généralement par microscopie; deuxièmement, dans les pays industriellement développés, où il s'agit d'une maladie rare et où, par conséquent, elle passe souvent inaperçue. Dans le premier cas, l'évaluation de la valeur prédictive d'une épreuve sera fondée sur son aptitude à détecter les cas d'infection en déterminant la proportion des individus à frottis positif qui sont séronégatifs et des personnes faiblement infectées ou en phase latente de la maladie qui sont séropositives. D'autre part, l'épreuve devrait pouvoir soutenir la comparaison avec la microscopie tant en ce qui concerne l'acceptabilité opérationnelle (facilité de mise en œuvre, fiabilité, rapidité, coût, entretien du matériel, problèmes de transport) que l'acceptabilité pour les patients.

Dans le second cas, l'épreuve servira principalement au diagnostic différentiel d'un large éventail de symptômes non spécifiques dont la tuberculose n'est que rarement la cause. L'évaluation de la valeur prédictive opérationnelle sera alors fondée sur l'aptitude à distinguer clairement les patients à frottis négatif ou atteints d'une forme non pulmonaire de la maladie de ceux qui présentent des symptômes compatibles avec la tuberculose mais d'étiologie différente. Étant donné que la tuberculose est souvent oubliée lors de l'établissement du diagnostic, l'épreuve, pour être opérationnellement acceptable, devrait pouvoir s'inscrire facilement dans les examens de routine pratiqués localement plutôt que de nécessiter une demande spéciale ou le recours à un laboratoire spécialisé. Dans de telles conditions, la simplicité, la fiabilité, la robustesse et une extrême spécificité sont des caractéristiques essentielles.

References

