Current trends and problems in health screening

J. M. G. WILSON

From the Department of Health and Social Security, London

SYNOPSIS Automation in clinical chemistry is a powerful tool. Like most innovations it is posing many questions as well as resolving some. Its use in population screening has led to questioning the validity of the old system of ‘normal values’ and to the search for better methods of indicating the significance of laboratory tests. This in itself is likely to have a good effect in that it is leading to a rigorous examination of the clinical value in diagnosis of the ever increasing number of laboratory tests.

The ease and cheapness with which laboratory screening can be carried out is in itself a danger; it can readily lead to an increasingly large diagnostic effort but diminishing returns, including some inconvenience or even harm to a fair proportion of the people screened. It is, therefore, important to examine thoroughly, through population surveys and intervention studies, the validity of all novel forms of prescriptive screening, including those of clinical chemistry. Cheapness of testing needs to be examined in the wider context of the overall cost-effectiveness of screening.

One great benefit of laboratory automation has been that it has resulted in new knowledge. For example, people have taken a new look at what are still usually called ‘normal values’ but which we might do better to call ‘values of interest in diseases’ (Murphy and Abbey, 1967) or ‘reference values’ (Whitehead, 1972).

Elveback, Guillier, and Keating (1970) have pointed out with clarity that many, if not most, clinical chemical variables do not, as has often been assumed, follow a Gaussian (or normal) frequency distribution, and also that values which are actually indicative of disease (for example, the serum calcium level in parathyroid disease) may not attract the attention of the clinical chemist or clinician because the population denominator on which the so-called ‘normal values have been based has been drawn from a biased sample, such as hospital patients. Moreover there is increasing recognition of the importance of such demographic factors as age and sex which are associated with bigger variations from population means than had been previously realized.

However, there are disadvantages too. The ability to offer multiple tests as a form of screening, whether as ‘admission profiles’ or in other contexts, has brought us sharply up against the problem of the large number of analyses which are far enough from the population mean to present the probability of some departure from health but of which the clinical significance is unclear. Often extensive, costly, and, for the patient, disturbing and time consuming, further investigations fail to reveal a reason for the supposed abnormal level. This problem is surely inescapable when screening is carried out as opposed to tests requested arising from the clinician’s suspicions of disease. Wise clinicians have long known, perhaps intuitively, when to ignore test values of doubtful significance but in the screening situation there is no similar ‘fall-back’ position. Murphy has pointed out that when the common practice is used of defining the limits of normal as plus or minus 2 standard deviations from the mean value, and a number of independent tests are carried out together, the chances of one test falling outside this range, and thus being classified as ‘abnormal’, is high. Thus, if 10 tests are performed there is a 40% chance of one test proving ‘abnormal’. If 20 tests are carried out this probability rises to 64% of tests that would be classified as ‘abnormal’ on this basis. This probability surely presents an appalling load for patient and investigator to bear as the product of screening, and should provoke thought on how to avoid such a situation. Murphy and Abbey (1967) have already done some very helpful thinking on these lines.

Two examples of clinical chemical screening are illustrative. The first is the report of the Metropolitan Life Insurance Company of America (1971)
Table I  Biochemical profiles of 1000 employees over 35 years of the Metropolitan Life Insurance Company of America (Metropolitan Life, 1971)

Table II  Results of 31 439 admission profile tests on 2071 hospital inpatients at Queen Elizabeth Hospital, Birmingham (after Whitehead, T. P., 1971b)

Table III  Results of admission profile tests classified as expected, diagnostic, or unexplained, in 2071 hospital inpatients at Queen Elizabeth Hospital, Birmingham (after Whitehead, T. P., 1971b)

on the screening by clinical chemical tests of 1000 of its employees over the age of 35. Table I shows that 25% of those screened had 'positive' findings based on 'normal' values previously established. This 25% was made up as follows: 1% was discovered to have some previously unknown disease; a further 1% had what is termed a 'patho-physiological syndrome'—or positive risk factor; 4% had positive tests indicating the presence of already known disease; while a further 19% had so-called 'unclassified' positive tests when no disease was discovered corresponding to the positive test. The new positive cases—11 in all—included three diabetics, two cases of alcoholic cirrhosis, two patients with Paget's disease of bone, and two cases of hyperparathyroidism.

The second example is the screening of just over 2000 hospital inpatients at the Queen Elizabeth Hospital in Birmingham by Professor Whitehead and his colleagues (personal communication, 1971a), using 16 clinical chemical tests. Table II shows that, of over 31000 tests, 24000 would not normally have been requested. Of these 24000 tests, 8% were abnormal tests which would not normally have been requested but for the profiling study. This 8% comprised 1% of tests which proved diagnostic, 2% which were abnormal but expected (equivalent to the Metropolitan Life study's 'confirmatory'); and 5% which were 'abnormal'. Table III deals with patients rather than with tests. Eighty-seven per cent of the patients had abnormal results but it must be remembered they were hospital inpatients. Of this 87%, 8% (i.e., about 170 consisted of new diagnostic findings whilst 43% had abnormal but expected test results. In the remaining 36% of patients the abnormal test finding was unexplained but in a random sample of 84 patients a good proportion of these had declared themselves on review at the end of one year, leaving 16% unexplained. Nearly half the new diagnoses were patients with...
either diabetes or iron-deficiency anaemia. Other diagnoses included hyperparathyroidism and thyroid disease but the number of these more serious and treatable conditions was small.

As in most useful studies, these examples of screening throw up more questions than answers; and it is questionable whether any health authority would be justified in introducing a continuing screening service of this kind on the strength of these results. Doubts can be summarized as follows.

The output, in terms of results achieved, is small compared with the input; in other words the proportion of life-or health-threatening conditions which are known to respond well to treatment is low, while the sum of effort of treatment (including that of the people screened) is high.

This unfavourable balance is compounded when we consider that we must add to the input side of the equation most of the effort spent either in arriving at conclusions, the importance of which for the individual is not known, for instance, moderately raised blood sugar and uric acid, or moderately low serum iron levels, and for which there is no corresponding output in the form of known prevention, control, or cure.

The unfavourable balance is further weighted by the need to add to the input side most of the people with 'unclassified' or 'abnormal but unexplained results'. These people will probably have needed diagnostic follow up to balance the trouble and cost of which no satisfactory output can be set—the best that can be done is either to ignore the results (after inevitably repeating them) or to offer the person a long-term follow up, using the findings as base-line information about that individual.

Implicit in these apparently ungrateful criticisms is the specific grumble that no specific measures of outcome have been built into these studies; that is, we do not know the effect, amongst the patients discovered by screening, on factors like death rate, survival, and quality of life in such terms as mortality and morbidity experience, use of medical and social services, and working life, compared with a similar group of people not offered the screening examination.

In seeking to avoid these pitfalls let us consider briefly six interlinked headings which one might use as tests of the value of screening for particular conditions. Figure 1 shows these headings in the form of a benzene ring in order to emphasize their interdependence.

**Importance**

Obviously, the condition should present an important problem, and its importance could be considered as a function of severity and prevalence. For example, the consequences of undetected and hence untreated phenylketonuria, though its incidence in a population is only about 1/10000-1/15000 births, are grave for the patient, the family, and society generally; while equally, although the consequences of, say, mild iron-deficiency anaemia in a population may be slight, the prevalence of low haemoglobin values is large, particularly in women of childbearing age, and it is therefore often considered worth screening for anaemia.

**Natural History**

We also need to understand the natural history of the condition well enough to know that we can effectively intervene at an early clinical or pre-clinical stage. To do this clearly involves the development of a valid test and the demonstration that early treatment is both effective and also influences the course of the disease more favourably than treatment started later after clinical diagnosis. These points can be illustrated in the diagnosis, screening, and treatment of phenylketonuria. It is important to remember that when we move from the usual clinical situation to the screening situation we are breaking new ground. In screening for phenylketonuria do we discover a population of infants with disease characteristics differing from those diagnosed by clinical means? Are there, for example, many patients with biochemical phenylketonuria, who do not become mentally handicapped? If so, the supposed results of treatment could be due, at least in part, to the discovery by screening of a new, previously unsuspected phenylketonuria population with a biochemical abnormality alone. If we forget this we may fall into an important error, as Birch and Tizard (1967) have shown by re-analysing the data of Berman, Waisman, and Graham (1966). Table IV shows the data as published by Berman *et al* (1966), from which it appears
Table IV  Phenylketonuria. Relationship between IQ and time treatment started (Berman et al, 1966)

<table>
<thead>
<tr>
<th>Group of Children</th>
<th>Mean IQ</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>First Test</td>
<td>Last Test</td>
<td></td>
</tr>
<tr>
<td>Unaffected siblings (n = 44)</td>
<td>111</td>
<td>113</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated before 4 months (n = 8)</td>
<td>84</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated 6-11 months (n = 5)</td>
<td>64</td>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated 12-24 months (n = 4)</td>
<td>57</td>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated 27-43 months (n = 5)</td>
<td>39</td>
<td>48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

That the earlier the diagnosis is made and dietary treatment started, the higher the intelligence quotient (IQ). However, when the data are looked at in the context of the nature of the population (Table V), whether diagnosed by screening, because of an affected relative, or because of mental subnormality, in relation to age we see a different picture. It is apparent that those diagnosed early were picked up by screening, while those diagnosed late were discovered because of mental subnormality. When the data are rearranged into the respective population categories, the apparent relationship between age at diagnosis and IQ disappears.

Can this situation be resolved? There is circumstantial evidence that it can. First, if treatment only seemed to be effective because new patients, not destined to develop mental subnormality, were now being found by screening, there ought to be many adults in the general population with biochemical phenylketonuria who have never been diagnosed and whose IQ is normal or nearly so. Levy, Karolkewicz, Houghton, and MacCready (1970) found only three cases of phenylketonuria in their study of 250,000 adults routinely examined by serology for syphilis and at the same time screened for phenylketonuria, and all three proved to be clearly mentally handicapped. Thus we have reasonably good evidence that most untreated patients with phenylketonuria are mentally subnormal. If treatment is effective most of the patients treated efficiently by diet should develop normally. There is fairly good evidence that this happens (Clayton, Moncrieff, and Roberts, 1967; Koch, Acosta, Fishler, Schaeffler, and Wohlers, 1967). However, further confirmation is desirable; in Great Britain where screening for phenylketonuria by the Guthrie test is carried out on most newborn children, a phenylketonuria register has been instituted, partly with the aim of ascertaining the development of all cases of phenylketonuria in the country in relation to the efficiency and duration of treatment. Providing dietary treatment is effective this register should in due course demonstrate the virtual disappearance of mental handicap in phenylketonuria in those patients in whom efficient therapy has been maintained.

Looking back over the development of the phenylketonuria screening story we must surely agree that had a randomized controlled trial been carried out when dietary treatment was first advocated, the story could have been much simpler and have taken up much less effort and resources in spelling it out. It also illustrates the need for a higher degree of certainty about the value of screening than is necessary in clinical diagnosis and treatment. Had it turned out in the end that screening for phenylketonuria was not achieving effective results, one need only consider the world-wide effort, including that of all the afflicted children and their families, that has been made to realize the need for the firmest of evidence before a form of population screening is offered to the public. In the clinical situation, on the other hand, there is the much more straightforward traditional obligation of the doctor in relation to the patient who seeks his help simply to do the best he can in the light of current knowledge.

Validity

A test to be worth while must be reproducible and be a good index of the condition screened for, missing relatively few cases and not creating an
unacceptably large burden of false positive cases needing further investigation. As fig 2 shows, we may be dealing with a bimodal distribution of our index variable, where the abnormal state is determined by a single gene, as in phenylketonuria, or, more frequently, with a single frequency distribution such as we see with index variables like blood pressure, intraocular pressure, haemoglobin, urea, or glucose. The distribution appears to be continuous, with no visible division between a ‘diseased’ and a ‘normal’ population.

Table VI shows formally the data required to evaluate the discriminatory power of a screening test, which falls into the four categories of true and false positives and true and false negatives. To obtain accurate data as nearly as possible the whole of a defined population must be examined. The most difficult part of the table to obtain experimentally, and the part that is usually missing, is the determination of the number of cases of the disease screened for, which were missed by screening—the false negatives. The reason for this information being so frequently missing is, of course, that the whole population (or a representative sample) should be examined by some definitive method, or ‘gold standard’ in order to diagnose all the cases in the population.

Figure 3 illustrates table VI diagrammatically in the form of a frequency distribution. The ‘cut-off’ point between the ‘diseased’ and the ‘normal’

<table>
<thead>
<tr>
<th>Test</th>
<th>Diseased (D)</th>
<th>Not Diseased (ND)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (+)</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Negative (-)</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Totals</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

Table VI Validity of a biological test for disease in a population of healthy and diseased people giving positive and negative results to a test

\[
P(+/D) \text{ (sensitivity)} = \frac{a}{a+c}
\]

\[
P(-/ND) \text{ (specificity)} = \frac{d}{b+d}
\]

\[
P(D+/) \text{ (predictive value of positive test)} = \frac{a}{a+b}
\]

\[
P(ND-/) \text{ (predictive value of negative test)} = \frac{d}{c+d}
\]
the other indices. However, the wide degree of separation given by the plasma phenylalanine level between diseased and normal child populations is not given by screening, which illustrates well the importance of basing conclusions about the value of a screening test on a defined screened population. In Penrose’s paper the comparison is between known patients with phenylketonuria and normal children. Table VII, compiled from data kindly provided by Clayton (1972), shows the discriminating power of the Guthrie test in screening infants born in the two northern metropolitan hospital regions of London,

![Schematic representation of discriminatory power of four indicators for phenylketonuria in a population of phenylketonuric and non-phenylketonuric children (from Penrose, L. S., 1951).](image)

**Table VII** Distribution of 216 116 Guthrie tests showing proportion of true and false positive tests and true and false negative tests.

<table>
<thead>
<tr>
<th>Guthrie Test</th>
<th>Phenylketonuria</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>Positive</td>
<td>16</td>
<td>527</td>
</tr>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>215 573</td>
</tr>
<tr>
<td></td>
<td>(c)</td>
<td>(d)</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>216 100</td>
</tr>
<tr>
<td></td>
<td>(a + c)</td>
<td>(b + d)</td>
</tr>
</tbody>
</table>

Screening level 4 mg phenylalanine/100 ml

Prevalence 1/13 500 births  Sensitivity 100% \( \frac{a}{a + c} \)

Specificity 99-8% \( \frac{d}{b + d} \)  Predictive value 3% \( \frac{a}{a + b} \)

Data (Clayton, 1972)
Current trends and problems in health screening

Table VIII  Effect on false positive rate given by Guthrie test of altering the screening level of serum phenylalanine

<table>
<thead>
<tr>
<th>Screening Level of Serum Phenylalanine</th>
<th>4 mg</th>
<th>6 mg</th>
<th>8 mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (%)</td>
<td>100-0</td>
<td>100-0</td>
<td>100-0</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>99-6</td>
<td>99-9</td>
<td>&lt;100-0</td>
</tr>
<tr>
<td>Predictive value (%)</td>
<td>2-0</td>
<td>24-0</td>
<td>36-0</td>
</tr>
<tr>
<td>False positives</td>
<td>419</td>
<td>28</td>
<td>16</td>
</tr>
</tbody>
</table>

Table VIII shows the advantage of raising the screening level to 6 and to 8 mg/100 ml. The number of false positives needing investigation is decreased, with the consequent benefits of relieving a number of families from anxiety and of cutting down the use of valuable laboratory and clinical resources. This benefit is indicated by an increase in the predictive value of a positive test, that is, the proportion of true positives to all positives, from only 2% at 4 mg to 24% at 6 mg, and 36% at 8 mg/100 ml; in actual cases reducing the false positives from 419 to 28 to 16. Figure 6 generalizes this situation, and shows the reciprocal

Fig 5    Distribution of serum phenylalanine levels in 117 444 tests on infants born in two hospital regional areas (Guthrie test) (from Clayton, B., unpublished).

Fig 6    Schematic representation of effect of altering the screening level of a variable, on true and false positive, and true and false negative tests.
relationship between false negatives and false positives—a case of swings and roundabouts.

**Effective Treatment**

It is axiomatic that early treatment should influence the course of the disease more favourably than treatment established at the time of clinical diagnosis. Unfortunately, however, the actual demonstration of the effectiveness of treatment is very difficult, as various studies of intervention, for instance in diabetes and coronary disease, have shown. As an example, fig 7 (from the data of Whitehead, 1971b) shows the distribution of nearly 4000 serum glucose levels in hospital inpatients. Those patients with levels between 120 and 199 g/100 ml may be regarded as 'borderline' cases in which the value of treatment needs to be investigated. As Keen has done (1971), this borderline group can be ethically randomized into treated and control groups and followed for effects on outcome. Keen's results, whilst suggestive of the effectiveness of treatment in the borderline patient, are so far not conclusive.

**Reaching the Population at Risk**

Indices of success or failure in population screening are the response rate of the population approached and, equally important, the yield of cases found, which is likely to be related to particular demographic risk factors. The use made of cervical cytology screening is a classic example of the difficulties of persuading those at most risk to make use of the preferred service. Inherent in a successful response is the careful preparation of the ground before starting screening by studies of attitudes and intentions. For instance, the Rotherham Multiple Health Screening Clinic study (Girt, Hooper, and Abel, 1969) found, in an interview sample of 200 people, that, although half said they intended to attend the screening clinic, in fact only 7% actually attended. Of these, three out of the 14 people attending had said at interview that they did not intend to be screened. Even attendance and receiving advice on the treatment of conditions found does not necessarily indicate success. Many people do not take the treatment provided; this has often been shown, for example, in treating symptomless hypertension, and in a randomized trial of treatment in raised intraocular pressure to establish its relationship with chronic glaucoma.

**Cost Effectiveness**

To obtain the evidence needed for choosing one form of medical care for a particular condition rather than another requires the successful measurement of output which, as we have seen, has almost never been achieved. So we are usually left with the problem of making important decisions on incomplete evidence. Factors of importance in addition to those already mentioned in making a rational choice between different methods of screening for the same condition, or between screening and normal clinical practice, are the respective costs of diagnosis;
the costs of treatment; the 'cost', in terms of disturbance as well as financial, to the individual, the family, and the community; and the relationship of the overall cost of screening to the total resources available. It was estimated, for example, in the Rotherham screening study that, if the 60,000 people eligible had all been screened within one year, the cost would have absorbed about two-thirds of the Rotherham Health Department’s annual budget for normal services.

One particular aspect of screening costs perhaps merits special mention. It is said at times that it is as cheap to perform a number of tests together in response to a request for perhaps a single test as to perform one or two discretionary tests, and that therefore the battery of tests should be provided to the clinician. This is surely poor logic, and there is a strong case for only carrying out tests of known worth which are relevant to the particular situation, whether clinical or screening. If this means fewer tests, the cost saved can be put in to some other aspect of clinical chemistry or even into other aspects of medical care. Perhaps the answer is to group tests very carefully together, as is now being done with increasing frequency so that a group of automated tests can be offered for any one of a number of different clinical or screening needs.

As a final illustration of the difficult problem of cost-effectiveness, Whitehead, White, and Smith (1972) have tested the validity of three different methods of detecting bacteriuria in school girls, comparing these screening with a standard bacterial culture technique. Table IX shows the validity, in terms of sensitivity, specificity, and predictive value, of the automated urinary glucose examination, the Uriglox dip-paper urinary glucose test, and the Uricult dip-slide urinary culture test. Whereas the dip-slide test has no missed cases, it has the lowest predictive value, though in no instance was the false positive rate high. Table IX shows also the cost per test and, more important, the cost per case found. The automated glucose test has the lowest cost per test as well as per case found. Given the greater sensitivity of the dip slide technique the slightly higher cost could probably be accepted. The costing does not, of course, take into account the cost of further diagnosis and treatment, and makes no pretense of providing an index to the true cost of finding and treating patients with bacteriuria. In order to obtain a meaningful cost-effectiveness equation, the real cost of diagnosis and the effectiveness of treatment would, of course, then need comparison with the results of normal clinical diagnosis.

In summary, the problems of successful screening are formidable, and few conditions really satisfy all the requirements, not least the form of screening by a group of clinical chemical tests. However, greater priority could be given to examining the undisputed advantages of automation in clinical chemistry, not only as to its efficiency in the performance and analysis of tests, but also in its effectiveness in relation to clinical medicine and surgery. We should seek answers to such questions as, ‘What are the most useful tests to perform?’ Which might be relegated to a non-priority category? ‘What values, in what conditions, should be regarded as of what degree of clinical interest?’; and ‘How reliable are our tests?’ To the spectator of the clinical chemical field there seems a real need for clinical chemists and clinicians to work out these and other questions in close and continuing discussion and trial. Perhaps it is here that the epidemiologist can make his most direct and most speedily effective contribution.

**Table IX. Cost per 100 tests and cost per case found in three screening methods for asymptomatic bacteriuria of differing validity (Whitehead, White, and Smith, 1972).**

<table>
<thead>
<tr>
<th>Test</th>
<th>Automated Glucose</th>
<th>Uriglox</th>
<th>Dip Slide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (%)</td>
<td>76-0</td>
<td>68-0</td>
<td>100-0</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>98-0</td>
<td>99-0</td>
<td>97-0</td>
</tr>
<tr>
<td>Predictive value (%)</td>
<td>62-0</td>
<td>76-0</td>
<td>58-0</td>
</tr>
<tr>
<td>Cost £/100 tests</td>
<td>12</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>Cost £/case found</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

References


Current trends and problems in health screening

J. M. G. Wilson

doi: 10.1136/jcp.26.8.555

Updated information and services can be found at:
http://jcp.bmj.com/content/26/8/555

These include:

**Email alerting service**
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/