FROM CLINICAL PERFORMANCE TO PROGRAM PERFORMANCE

Helmut WENZEL, M.A.S
Kurt NEESER, MD

1Health Economist, Konstanz, Germany
2SIMEDES - Swiss Institute for Medical Decision Support, Basel, Switzerland

Some words on test performance

Diagnostic technologies enable medical professionals to find the right diagnosis and finally allocate the right treatment to the right patient. Diagnostic examinations use tests to sort the population into two groups: healthy and diseased. This distinction is based on threshold values. Depending on this threshold value the test result classifies the population into sick and healthy.

In practice both populations, diseased and healthy are not totally separated. Figure 1 shows such an overlap. As a consequence of this overlap there are two more categories with patients that possibly will not be diagnosed correctly: Persons who actually are sick but who are classified as healthy, false negatives (FN), and those who are falsely tested as being sick, false positives (FP) (see figure 1 b). Changing the threshold values that separate diseased from non-diseased will change the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Depending on the goal of diagnostic testing and the expected consequences of falsely categorizing people the acceptance of the test performance may vary and the expectation on the performance of the test may be different. Performance is comprises two important aspects. Firstly, the clinical performance, described by the diagnostic accuracy, which enables the user to classify patients into relevant health states and, secondly, the usefulness of the diagnostics test results (program performance).

The performance in terms of test accuracy usually is shown by its sensitivity and specificity. Sensitivity or true positive rate is the proportion of diseased people with a positive test and specificity is the proportion of non-diseased with a negative test. This performance information is easily accessible and it is derived from clinical evaluations with a fixed proportion of diseased persons and sometimes is also named diagnostic performance, analytic performance or testing performance (2). A very illustrative plot that shows the accuracy of a test (pairs of TPF and FPF), depending on different threshold values is the Receiver Operating Curve (see ROC in figure 2).

Usefulness on the other hand aims at documenting the value that testing might bring to decision-making.

Figure 1. Distributions of test results for diseased and non-diseased populations with categories defined by testing threshold (DT), adapted from A.J., COOPER, N.J., GOODACR, S., and STEVENSON, M [1]
In contrast to accuracy measures it is a conditional measure, incorporating the peculiarities of an application, too, especially the influence of various prevalence rates at various locations of the disease under survey. Additionally, the economics of testing are getting more and more in the focus and now are perceived to be an important part of the usefulness. In this context not only the cost of testing are examined carefully, moreover the cost consequences, i.e. additional diagnosing procedures, therapeutic measures, due to the classification (and possible errors) are included in calculations.

The value of a test, and consequently the usefulness of test results are measured by the predictive values. The predictive values, i.e. the probability that a given test result correctly classifies into diseased and non-diseased populations depends on the prior probability (prevalence of the disease in a population under study) as well. A test which performs well in a laboratory environment might be less good in a situation with a different prevalence. A test can be used to confirm a tentative diagnosis (rule-in) or to reject the prior diagnosis (rule-out). Figure 3 shows the relationship between prevalence (p), the positive predictive value (ppv), and the negative predictive value (npv).

It is quite obvious that with changing likelihood that a certain disease might be present in a specific location and population - depicted by the prevalence - the correctness and usefulness of a diagnosis will vary. Therefore technical testing performance should be distinguished from the potential performance in a health program (e.g. screening), or in a doctor's office with a different proportion of people with the target disease. Furthermore the suitability of test for diagnostic purposes, i.e. ruling in or ruling out a disease, is determined by the magnitude of its sensitivity or specificity. When a diagnostic test has a high sensitivity, a negative result rules out the diagnosis, in case that a diagnostic test has a high specificity, a positive result rules in the diagnosis. Taking this into account, choosing and applying a diagnostic test requires some planning and makes it necessary to ask a few questions about the potential outcomes (see figure 4).

How to evaluate and to anticipate a program performance?
Diagnostic test results mark the beginning of a „production process of health“. In figure 4 this process is shown using the example of a population screening. Having in mind figure 1, questions arise on the correctness of a classification, and the subsequent consequences within this „production process of health“ in a specific setting. Diagnostic procedures use health care resources, may have adverse effects, and - as mentioned above - can provide wrong results.Erroneous classifications may lead to unnecessary and possibly harmful treatment or even prevent or delay access to the right treatment. Diagnostic are suitable subjects for economic evaluations (3).

Principally there are two basic approaches in the evaluation of diagnostic tests proposing procedures to analyze the value (comprehensive usefulness including economics) of this kind of processes. The first is in the tradition of classical discriminant analysis which uses the Receiver Operating Curve (ROC) as a standard tool to examine the test performance, as outlined above. The second category could be more precisely characterized as a child of "rational clinical theory" or „clinical decision analysis“(6). The focus here lies on the quantification of expected benefit from testing. Based on pre-posterior analysis (or Bayesian Revision) e.g. predictive values are calculated. Phelps et al. (7) developed an approach integrating ROC analysis with clinical decision analysis and provided a surface of expected value of clinical information (EVCI) with cut-off points establishing the region where the use of the diagnostic test is desirable. Laking et al. (3) discussed an approach that explicitly aims

![Figure 2. Receiver Operating Curve (ROC) as a result of changes in the test threshold](image)

![Figure 3. Relationship between prevalence (p), positive predictive value (ppv), and negative predictive value (npv). Own calculations based on a sensitivity of 90% and a specificity of 99%](image)
at strengthening the economic side, especially by using economic evaluation techniques and by combining both the predictive information with a decision analytic-based approach.

**Resolving the economic usefulness**

Researchers of both “schools” see the importance of economic appraisal in determining the value or usefulness of a specific test. Nevertheless the way how cost is considered varies as well as the effort needed. A comprehensive Health Technology Assessment probably will prefer the more complex way of the rational clinical theory approaches.

Cost can be seen from different view-points and have to be defined accordingly: patient, society, health care provider or payer. Metz (8) outlined in his cost-benefit approach the total cost appraisal as follows:

\[
C_{\text{avg}} = C_0 + C_{\text{TP}}*P(\text{TP}) + C_{\text{TN}}*P(\text{TN}) + C_{\text{FP}}*P(\text{FP}) + C_{\text{FN}}*P(\text{FN}),
\]

where \( p \) denotes the probability of occurrence of the cost element and \( C_{\text{TP}}, C_{\text{TN}}, C_{\text{FP}}, C_{\text{FN}} \) mark the cost of TP, TN, FP and FN and \( C_0 \) is the cost of the test. It turns out that this means to optimize the first derivative of the ROC (3;8;9). The optimal ROC slope occurs at the point where the savings of a marginal increase in TP balance the costs of the resulting increase in FP. With this information the optimal combination of TPF and FPF can be calculated, taking into account the cost of wrong classifications. Technically the slope of the ROC - the first derivative - equals the slope of the tangent. The slope of the tangent is defined as:

\[
m = \frac{(C_{\text{FP}} - C_{\text{TN}})}{(C_{\text{FN}} - C_{\text{TP}})} \cdot \frac{1-p}{p}
\]

where the relative cost are defined as:

\[
\frac{C_{\text{FP}} - C_{\text{TN}}}{C_{\text{FN}} - C_{\text{TP}}}
\]

Zweig et al (9) even use a simplified cost minimizing approach where they only use the ratio of FP and FN. For a quick, but nevertheless valid and informative procedure, we would like to present an approach where relatively little information is needed. The calculation is based on data of the Receiver Operating Curve, the supposed or known prevalence rate in the population and the relative cost caused by FP, TN, FN and TP. As only the ratio of cost is needed for such a rough calculation, this method is helpful for an ex ante evaluation of screening alternatives during a planning phase.

### A practical example

Let’s assume that we have to decide on a screening program for early detection of a disease. Within the framework of the program two tests with different performances are available. Furthermore we know that the prevalence of the disease is about 3% and the cost consequences of FN are more severe. Using the process of figure 4 as a guidance we would examine the appropriateness in terms of performance and adverse effects (questions 1 and 2). Figure 5 shows the Receiver Operating Curves of the two test options. The accuracy of test 1 (measured by the AUC) is 91.4%; the corresponding AUC of test 2 is 83.2%. Based on a prevalence of 3% and the relative cost of about 0.2, the optimal operating point (OOP) of test 1 is marked by a sensitivity of 61.4% and a specificity of 98.6%. This is where tangent 1 touches the curve of test 1. The comparable performance of test 2 is marked by a sensitivity of 25.7% and a specificity of 97.4%. With respect to the financial constraints of falsely classifying, test 1 classifies a test population of 1,000 persons correctly in 97.4%, the test 2 in 95.3% accordingly. In the case of test 1, 18 cases are true positive, 956 are true negative, 12 are false negative and 14 are false positive. In the case of test 2, 8 cases are true positive, 945 cases are true negative, 25 are false positive and 22 are false negative. Changing the prevalence to 5% and the relative cost to 0.6 will move the OOP of test 1 to a sensitivity of 56.2% and a specificity of 99.2%. The OOP of test 2 is at a sensitivity of 3.9% and a specificity of 99.9%. Changing the prevalence to 50% and the relative cost to 2 - which means that the cost consequences of false-positives are more relevant - the OOP of test 1 is at a sensitivity of 72.3% and a specificity of 95.4%. The OOP of test 2 would be at a sensitivity of 59.6% and a specificity of 87.8%.

If the cost of missing a diagnosis (FN) is high compared to the cost of FP, and a treatment would be safe, it makes

---

**Figure 4. Sample Clinical Analytic Framework: From Diagnosis to Patient Outcomes (4;5)**
sense to move the OOP to the right with a high TPF and the risk of having many FP. If one has to be more conservative because the possible outcomes of a treatment are at high risk or wouldn’t help very much, the OOP should be moved to the left. This would reduce the number of FP but also miss a higher number of positives. Moving OOP to the right means also a lower slope, conversely moving to the left means increasing the slope. Figure 6 shows the impact of relative cost and prevalence on the steepness of the slope (m of tangent and first derivative of ROC).

Conclusion
Diagnostic test results mark the beginning of a „production process of health“. They are important for the allocation of adequate therapeutic and/or preventive actions. However, in times of cost containment in the health care sector, the economic evaluation of diagnostic tests is of increasing importance to safeguard a fair allocation of the available resources. This paper has outlined different methods for economic assessments with different levels of complexity. High complexity and high level evaluations require comparatively considerable investments in time and budgets. The authors are convinced that an evaluation based on five parameters could be sufficient for a first but valid assessment, especially in the planning process of programs, where ex ante information on the performance will be needed most.

Reference List