Diagnosis: Toxic! – Trying to Apply Approaches of Clinical Diagnostics and Prevalence in Toxicology Considerations

Sebastian Hoffmann, and Thomas Hartung

European Commission, JRC–Joint Research Centre, Institute for Health & Consumer Protection, ECVAM–European Centre for the Validation of Alternative Methods, 21020 Ispra (VA), Italy

Received November 30, 2004; accepted January 28, 2005

The assessment of relevance of toxicological testing was compared with approaches of diagnostic medicine, a discipline that faces a comparable situation. Considering the work of a toxicologist as setting a diagnosis for compounds, assessment tools for diagnostic tests were transferred to toxicological tests. In clinical diagnostics, test uncertainty is well accepted and incorporated in this assessment. Furthermore, prevalence information is considered to evaluate the gain in information resulting from the application of a test. Several common toxicological scenarios, in which test uncertainty and prevalence are combined, are discussed including the interdependence of test accuracy, prevalence and predictive values or the sequential application of a screening and a confirmatory test. In addition, real prevalences derived from prevalences determined by an imperfect test are presented. We conclude that information on prevalences of toxic health effects is required to allow a complete assessment of the relevance of toxicological test. In this process, lessons can be learned from evidence-based approaches in clinical diagnostics.

Key Words: prevalence; validation; evidence-based medicine; biometry; reference standard; risk assessment.

THOUGHT STARTER

Let’s assume you are a doctor asked to carry out an HIV test on a healthy European who has no specific risk factors. You choose the best test available, which is 99.9% accurate. Unfortunately, the result is positive. Bad news for your patient? Not yet: The prevalence of HIV infection—i.e., proportion of infected people in the general population—in Europe, is about 1:10,000 inhabitants. This means, if testing 10,000 people you will pick up one real-positive but your best available test will show 10 false-positives. A positive result will thus only be correct in one out of 11 cases; i.e., the probability that your patient is really HIV infected is about 9%. Similar reasoning can be found in Gigerenzer et al. (1998), who pointed out the need to communicate carefully any diagnosis to these patients.

What does this teach us as toxicologists? Our problem is, in most cases, that we do not even know how accurate our test methods are (certainly less then 99.9%), and we have no indication of the prevalence—i.e., the proportion of toxic chemicals for a given health effect—in specific populations of chemicals.

This thought prompted us to elaborate on what diagnostic medicine can teach toxicology in handling the uncertainty in our test methods of setting the diagnosis of a substance exerting a given toxic effect.

THE ACCURACY OF THE DIAGNOSIS—TRANSLATION TO TOXICOLOGY

Setting a diagnosis in a medical clinic is an art that involves three aspects: the patient, the physician, and the diagnostic measures (Haynes et al., 1996, 2002). Difficulties in setting a diagnosis arise from incompatibilities and limitations of each component. When identifying a toxic hazard of a chemical, similar to the role of the physician, the expert assessor has to overcome the many limitations of what is known of the nature of the phenomenon. In both cases a number of problems have to be considered (Table 1).

It is impossible to judge the relative contribution of the many variables. The conclusion is simple: Our tools to assign a toxic health effect are imperfect. This is well accepted in the field of clinical diagnostics (Boyko et al., 1988; Sackett et al., 1991; Hunink et al., 2001; Knottnerus et al., 2002). However, in toxicology, we are not used to estimating and incorporating uncertainty, but we base our conclusions (i.e., labeling/classification as well as use or non-use of the substance for certain purposes) on this imperfect assessment.

It must be pointed out that this review omits any discussion of the relationship of a toxicity test to any adverse human health effect. The analysis is based only on the interplay of test
quality and prevalence for a given test. This is principally applicable to any test, be it diagnostic in humans, in animals, or in vitro.

THE QUALITY OF OUR “DIAGNOSTIC” TOOLS

In the field of carcinogenicity testing, the rodent bioassay’s 50% positive rate triggered a detailed discussion of the restrictions and limitations of its predictive capacity for humans (Ames and Gold, 1990; Gold et al., 1998). However, few toxicity tests have been studied with this scrutiny. The area of validation of alternative methods has pioneered the assessment of the quality of methods employed in toxicology (Balls et al., 1990, 1995). The crucial achievement here was the concept of relevance, i.e., assessing not only the reliability/reproducibility but also the predictive capacity of a method. This implies, however, a point of reference (usually termed the “reference standard”). In clinical diagnostics this reference is often included in systematic studies (Walter et al., 1999; Knottnerus and Muris, 2003) and new assessment tools have even been developed for study evaluation (Whiting et al., 2003). In toxicology this optimal way of direct comparison most often is not applied for reasons of cost or animal welfare. If the reference is evaluated at all, a retrospective comparison (e.g., based on information from databases) is carried out (Fentem et al., 1998). In both disciplines, this reference standard is usually but not necessarily another test. In toxicology a consensus of experts on the toxicological properties or classification of a given substance could substitute for the standard. Likewise in clinical diagnostics: sometimes the reference standard for assessing the performance of a diagnostic measure is established by consensus, when an independent expert panel establishes a patient’s diagnosis from established clinical criteria (Weller and Mann, 1997; Knottnerus and Muris, 2003).

It is of utmost importance to understand that validation assesses the reliability and relevance of methods. Figure 1 illustrates the validation process and which type of information constitutes the validity of a test method. Because the primary test result is often not expressed in the same way as the reference test, a prediction model is required to translate one into the other (Worth and Balls, 2001). For example, results of the test method might be continuous but must be classified into positive/negative or negative/mild/moderate/severe employing the thresholds. In case of alternatives to animal experiments, the prediction model in toxicology is established by consensus, when an independent expert panel establishes a patient’s diagnosis from established clinical criteria (Weller and Mann, 1997; Knottnerus and Muris, 2003).

### TABLE 1

**Problems of Setting a Diagnosis in Clinical Medicine and Toxicology**

<table>
<thead>
<tr>
<th>Clinical medicine</th>
<th>Toxicology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient-related problems</td>
<td>Substance-related problems</td>
</tr>
<tr>
<td>Mix-up of patients</td>
<td>Mix-up of substance identities</td>
</tr>
<tr>
<td>General health status</td>
<td>Purity of the substance including synergistic and antagonistic effects of bystanders/impurities</td>
</tr>
<tr>
<td>Possibility of carrying out invasive diagnostics</td>
<td>Stability of the substance</td>
</tr>
<tr>
<td>Individual reasons not permitting a diagnostic test (no consent of the patient, phobias, allergies)</td>
<td>Chemicophysical properties of the substance interfering with the test; size of crystals and solubility in carrier as well as biological fluids</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>Further toxic effects of the substance beside the tested one</td>
</tr>
<tr>
<td>Temporal changes: compliance; stage of disease; fluctuation of symptoms</td>
<td>Bioavailability, i.e., in vivo toxicokinetics or in vitro biokinetics</td>
</tr>
<tr>
<td>Test-related problems</td>
<td>(describing the fate of a substance in cell culture including nonspecific binding)</td>
</tr>
<tr>
<td>The variability and reproducibility of the test</td>
<td>Test-related problems</td>
</tr>
<tr>
<td>The reliability of reference standards employed</td>
<td>The variability and reproducibility of the test</td>
</tr>
<tr>
<td>The data interpretation procedure, e.g., the thresholds for classifications</td>
<td>The reliability of reference standards and materials employed</td>
</tr>
<tr>
<td>The pathophysiological relevance of the test used</td>
<td>The data interpretation procedure, e.g., the thresholds for classifications</td>
</tr>
<tr>
<td>The limit of detection of the test</td>
<td>The mechanistic relevance of the test used</td>
</tr>
<tr>
<td>The predictive capacity of the test for a disease (accuracy, sensitivity, specificity) for the human health effect</td>
<td>The limit of detection of the test</td>
</tr>
<tr>
<td>Personnel-related problems</td>
<td>The predictive capacity of the test accuracy, sensitivity, specificity)</td>
</tr>
<tr>
<td>Conducting the test correctly: linked to the quality and training of personnel</td>
<td>Personnel-related problems</td>
</tr>
<tr>
<td>The bias of the technical personnel and the physician when interpreting the test</td>
<td>Conducting the test correctly: linked to the quality and training of personnel</td>
</tr>
<tr>
<td></td>
<td>The bias of the experimenter/assessor with regard to the outcome of the test for a given substance</td>
</tr>
</tbody>
</table>
normally developed before the validation study, using a training set of substances (Bruner, 1996). The quality and properties of this set pre-determines the quality of results and applicability of the test method. Similarly, the selection of patients to establish a diagnostic method determines its quality for its intended use, the so-called patient spectrum (Irwig et al., 2002). If the selection is not representative or is somehow flawed—e.g., if it includes only severe cases—the relevance of the test will be impaired or restricted. Consequently, it is instrumental that validation studies in toxicology include a sufficient number of weak toxicants. Optimally, although often a dichotomous (i.e., positive vs. negative), test outcome is chosen, the selection should cover the whole range of toxic potency. This allows a better test assessment by expressing probabilities (e.g., of being positive or negative), for each chemical: A highly toxic compound will more likely be classified as such than a moderately toxic compound. Another difference in prediction model development/threshold definition in clinical diagnostics is that the sample sizes are usually substantially larger. This eases biometrical assessment, but several advantages of setting a diagnosis of toxicity can compensate here:

Testing of substances can be synchronized.
Testing can be repeated.
Positive and negative controls are readily available.
Replication and related substance testing are feasible.
The number of toxic health effects is limited.

THE IMPACT OF PREVALENCE

As demonstrated in our thought starter, the prevalence of a disease is a key determinant of the practical value of a diagnostic measure (Buck and Gart, 1966; Linnet, 1988; Grimes and Schulz, 2002). If you are looking for something rare, even the best test will produce too many false-positives to provide a reliable result. It is therefore crucial to use descriptors of test relevance, which take the prevalence into account. In most simple cases (two outcomes of the test as well as of your reference standard, which will be mainly considered here for reasons of simplicity), this means describing the relevance by the PPV (positive predictive value) and NPV (negative predictive value) instead of the sensitivity (i.e., the probability of a correct negative result) and the specificity (i.e., the probability of a correct positive result). It would be challenging but also demanding to expand this concept of including prevalence information to multiple-class outcomes, as we have recently demonstrated for the case of skin irritation (Hoffmann et al., 2005). The predictive values estimate the proportion of correct positive/negative test outcomes in all positives/negatives and are thus an indication for the reliability of a positive/negative test result. In toxicology, however, we often forget that our panel of test compounds is not reflecting the real world but was designed to produce efficiently reliable estimates for sensitivity and specificity. For example, we choose 20 negatives and 20 positives, not caring what the toxic effect is. Thus, the predictive values based on the artificial study prevalence are only telling us the predictive capacity of the test if the same distribution of positives and negatives is found in the real world. This is usually not the case. Unfortunately, for most toxic health effect, we have no idea about their actual prevalence, for example in chemicals of general use. Not only are we lacking complete information on basic toxic properties for a large number of high production volume chemicals on the market (EPA, 1998; Allanou et al., 1999), but even for the existing data sets no such analysis is available. Therefore, efforts should be made to retrieve reliable estimates of those prevalences for the most relevant areas of toxicology.
Taking the example of skin irritation, this problem of estimating prevalences was approached (Hoffmann et al., 2005). Although in that report a detailed distribution of skin irritating potential was presented and analyzed, here we restrict ourselves to the prevalence analysis of the dichotomous outcome, i.e., irritant vs. non-irritant. In the New Chemicals Database of the European Chemical Bureau, which includes 3121 chemicals mainly registered in the last 15 years, the prevalence of skin irritating substances (according to EU-regulation) assessed by an animal experiment was 7.9%. The applicability domain with this prevalence would be the population of newly developed chemicals, whereas its use for other domains would have to be discussed. Because the database contains only results from one test in one laboratory for each chemical, the predictive capacity of the in vivo experiment could only be modeled for the outcome of a repetition of the same experiment. This resulted in a specificity of 99.7%, i.e., three out of 1000 nonirritating chemicals would be classified false-positive, and a sensitivity of 94.1%, i.e., 59 out of 1000 irritating chemicals would be classified false-negative. Thus a NPV of 99.5%—i.e., only one of 200 chemicals classified negative would in fact be an irritant—and a PPV of 96.8%—i.e., out of 1000 chemicals classified as irritating 32 would be not irritating—were calculated. It is evident, that modeling further aspects of variability—e.g., the within- and between-laboratory reproducibility—would decrease the predictive capacity estimates, resulting in the respective decrease of the predictive values. For some combinations, the effect of prevalences and test accuracy—i.e., assuming that sensitivity equals specificity—on the predictive values is illustrated (Table 2). The most important consequence of these considerations is that, for rare toxic events, we can rely on the negative test results but not on the positive ones.

As the negative predictive value is always close to 100%, this shows that the idea of controlling negative test results by means of a second test—e.g., confirm negative in vitro results in vivo, as suggested in the field of skin corrosion and irritation (OECD, 2002)—makes no sense at all for rare toxicities. Because the exposure events are rare, in most cases both tests will be done anyway, although the negative predictive value is high. In contrast, the positive test results have to be challenged; i.e., in regulatory toxicology it is necessary to avoid over-classification and unnecessary restrictions of substances. In this context, we are well aware of the most crucial safety aspect of false-negative classifications, but the calculation shows that a second test can hardly improve the NPV, which is impaired by the false-negative findings of the first test. For example, for skin irritation, even a test with only 70% accuracy will identify negatives correctly in more than 96% of the cases.

**CONSEQUENCES OF INACCURATE TESTS FOR PREVALENCE DETERMINATIONS**

An important question often overlooked is this: How reliable are prevalences of rare diseases/toxicities if they are being assessed with our imperfect tools? If we agree that an in vivo experiment is not 100% accurate, many false-positives will populate our databases in case of rare toxicities. This means that rare toxicities are rarer than we believe. For illustration, in Table 3 we present some combinations of prevalence determined by a test with a given accuracy. For example, applying a test with 90% accuracy and finding a prevalence of 20% means that the true prevalence is only 12.5%; i.e., only 5 of 8 positive test substances are truly positive. Similarly, if we assume that the rabbit skin irritation test is 95% accurate, more than half of the selected skin irritants (prevalence 7.9%) would be false-positives.

An obvious consequence is that the usefulness of databases for selecting the proper reference standard data is limited. As long as confirmatory testing in vivo is not carried out, it might be favorable to rely on the fewer, but more extensively studied substances from the scientific literature.

**THE USE OF CONFIRMATORY TESTS**

It is common practice to apply a second test to confirm or challenge the results of a first test. When retesting positives, the specificity of the test procedure can be improved; when retesting negatives, the sensitivity of the test procedure can be improved. As we have seen above, this makes sense for relatively low prevalences only for positives, the NPV being almost optimal anyway. However, sensitivity and specificity of

### TABLE 2

<table>
<thead>
<tr>
<th>Positive predictive value (PPV) of a test [%]</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevalence of toxicity</td>
<td>99.9%</td>
</tr>
<tr>
<td>0.01%</td>
<td>9.1</td>
</tr>
<tr>
<td>0.1%</td>
<td>50.0</td>
</tr>
<tr>
<td>1%</td>
<td>91.0</td>
</tr>
<tr>
<td>10%</td>
<td>99.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative predictive (NPV) value of a test [%]</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevalence of toxicity</td>
<td>99.9%</td>
</tr>
<tr>
<td>0.01%</td>
<td>100</td>
</tr>
<tr>
<td>0.1%</td>
<td>100</td>
</tr>
<tr>
<td>1%</td>
<td>100</td>
</tr>
<tr>
<td>10%</td>
<td>100</td>
</tr>
</tbody>
</table>

**NOTE:** Assuming a given test accuracy resulting from equal sensitivity and specificity, the consequences of different prevalences for the predictive values are calculated.
a test are interdependent: By defining, for example, the threshold value for a classification as positive or negative, one can be increased at the expense of the other, usually demonstrated with receiver operation curves (ROC) as illustrated in Figure 2 (McNeil et al., 1975; van der Schouw et al., 1995). This offers the opportunity to render tests extremely sensitive, accepting impaired specificity, a situation typical for screening tests.

A commonly applied and simple strategy in clinics as well as in toxicology is the combination of a screening test followed by a confirmatory test. With an oversensitive test a population is screened in order to detect as many positives as possible. Inevitably, this approach produces a lot of false-positive results in the first step. In the second step, all positively screened patients/substances are subjected to a confirmatory test, which should be able to discriminate positives from negatives. The advantage of this strategy is often a reduction of costs, as screening tests with their lower overall predictive capacity are often substantially cheaper than their associated confirmatory tests. Nevertheless, the usefulness of this approach again strongly depends on the prevalence of the health effect (Buck and Gart, 1966) and the tests’ dependence (Marshall, 1989). Although, the overall testing costs are significantly decreased, the positive predictive value does not change substantially, when compared to the PPV of the confirmatory test. For example, let us assume a prevalence of 1%, an extremely sensitive screening assay with a sensitivity of 100%, but a specificity of only 50%, and a good confirmatory assay with an accuracy of 95%. Testing 10,000 substances, of which according to the assumed prevalence 100 are positive, the initial screen reduces the number of substances subjected to the confirmation test by 4950, i.e., all negatives. Applying subsequently the confirmatory test reduces the overall NPV from 100% to 99.95%; i.e., 9652 of 9657 negatives are true negative. But this results in a PPV of only 27.7% (Table 4); i.e., 95 of the 343 positives are true positives, compared to a PPV of the confirmatory test alone of 16.1% (Table 2). This means that, for rare health effects, a sufficient specificity of the screening has to be maintained; even close-to-perfect confirmatory tests cannot compensate. In the given example, an improved screening test specificity of 80% would result in a PPV of 49.0%, and a value of 90% would result in a PPV of 65.7% (Table 4).

A solution to further increase PPV is the application of a series of complementary tests in a sequence. This solution carries with it the problems of loss in cost savings and of the evaluation of the dependencies between tests, as complementary screens might be difficult to find. Furthermore the efficacy of combining tests for rare health effects is limited even under optimal conditions—i.e., assuming test independency (Table 5).

When combining a screening test and a confirmatory test, the screening test balancing sensitivity and specificity has to be carefully designed and adjusted because false-negative results in this step will have a negative effect on patient health or consumer safety. A more detailed insight into the prevalence is needed here to estimate consequences. One also has to consider the strength of a response and not only the dichotomized classification: It makes an enormous difference whether a large proportion of actual results are borderline to a given threshold or whether the negatives and positives are clearly distinct (Brenner and Gefeller, 1997; Bruner et al., 2002). Especially in

<table>
<thead>
<tr>
<th>Determined prevalence</th>
<th>99.9%</th>
<th>99%</th>
<th>95%</th>
<th>90%</th>
<th>80%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>0</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>1%</td>
<td>0.9</td>
<td>0</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>5%</td>
<td>4.9</td>
<td>4.1</td>
<td>0</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>10%</td>
<td>9.9</td>
<td>9.2</td>
<td>5.6</td>
<td>0</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>20%</td>
<td>19.9</td>
<td>19.4</td>
<td>16.7</td>
<td>12.5</td>
<td>0</td>
<td>n.d.</td>
</tr>
<tr>
<td>30%</td>
<td>30.0</td>
<td>29.6</td>
<td>27.8</td>
<td>25</td>
<td>16.7</td>
<td>0</td>
</tr>
</tbody>
</table>

n.d. = not defined, because for a given test accuracy, e.g. 90%, the determined prevalence cannot be smaller than 100% – accuracy, e.g. 100% – 90% = 10%.

The table presents the real underlying prevalences, when rare prevalences are determined with imperfect tests assuming accuracy = sensitivity = specificity.
the case of low prevalences, an extremely skewed distribution
toward the negative end of the scale can be expected.

PREVALENCE IN DISTINCT CHEMICAL CLASSES

So far we have handled the chemicals from the chemical
universe only as single entities. But each one is related to
others, e.g., by chemical structure, physiochemical properties,
or mechanism of action. There is an increased probability that
related chemicals will exhibit similar toxicological effects; thus
if information about a related chemical is available, it should be
considered. This can be compared to the clinical situation
where the integration of family anamnesis might change the
probability of a diagnosis dramatically. Consider, for example,
mutagenicity: in contrast to being a relatively rare toxic effect
among all chemicals, it is a common effect of the chemical
group of nitrosamines. Making use of this kind of a priori
information, structural alerts or read-across approaches should
help to assign chemicals to families with high or low
prevalence of health effects. This is by no means new but
reflects practices of priority setting by (Q)SAR and similar
computational approaches, as well as the experience of the risk
assessor. What we desperately need, however, are measures
of how closely two substances are related. As long as such
measures are lacking, groups of chemicals should be considered
mainly as classes with different prevalences and therefore
different certainty of test results. This approach allows us to
apply tests with different sensitivities and specificities, and it
calls for test strategies that take into account general prev-
allence, chemical classes with their individual prevalences, and
the use of proper tests with suitable predictive capacities.

CONCLUSIONS

Toxicological tests can be considered as diagnostic tools to
assess the toxicological properties of substances. Taking this
point of view, it is possible to draw parallels between diag-
nostic medicine and toxicology and to explore the possibility of
adopting evidence-based medicine methodology in toxicolog-
ical evaluations. To properly assess toxicological tests, their
reliability and relevance need to be explored, a process, in
which the reference standard is of crucial importance. If no
appropriate reference standard exists, expert consensus can be
a valid alternative. Nevertheless, reference standards will
always be imperfect. Accounting for this imperfection is
crucial for a complete test evaluation. In the present study,
the extent to which imperfect reference data might populate
databases is demonstrated, especially with false-positives for
low prevalence cases.

Furthermore, a systematic assessment of prevalences of toxic
health effects in the chemical universe as well as in defined
classes of chemicals is required. Only when taking into account
prevalence considerations, can “diagnostic” value of a test can
be estimated. For example, in low prevalence situations,
negative predictive values are almost optimal, which challenges
the approach of confirming negative results. In addition, we
emphasized that the use of confirmatory tests strongly depends
on prevalence and test accuracy. Additional information on
substances, including chemico-physical properties, chemical

<table>
<thead>
<tr>
<th>Prevalence</th>
<th>n = 1</th>
<th>n = 2</th>
<th>n = 3</th>
<th>n = 4</th>
<th>n = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>0.9</td>
<td>7.5</td>
<td>42.2</td>
<td>86.8</td>
<td>98.3</td>
</tr>
<tr>
<td>1%</td>
<td>8.3</td>
<td>45.0</td>
<td>88.0</td>
<td>98.5</td>
<td>99.8</td>
</tr>
<tr>
<td>5%</td>
<td>32.1</td>
<td>81.0</td>
<td>97.5</td>
<td>99.7</td>
<td>100</td>
</tr>
<tr>
<td>10%</td>
<td>50.0</td>
<td>90.0</td>
<td>98.8</td>
<td>99.9</td>
<td>100</td>
</tr>
</tbody>
</table>

This table shows how the PPV can be increased by combining several tests with 90% accuracy (assuming sensitivity = specificity).
structure, or classes, might affect the prevalence. Therefore, for toxicological hazard identification and testing strategies, integration of prevalence information is crucial. In this process, lessons can be learned from the medical diagnosis setting, and especially from the evidence-based evaluation of diagnostic measures, as a step toward evidence-based toxicology.

REFERENCES


