ORIGINAL ARTICLE

A Measure Evaluating Relevance of a Validation Study of Alternatives to Animal Testing

Takashi Omori

Department of Biostatistics, Kyoto University, Japan

Abstract
Sensitivity, specificity and accuracy are well known measures for evaluating the relevance of an inter-laboratory validation study for alternative tests. It is not generally discussed that the measures are dependent on two determining factors: a set of chemicals and the number of laboratories. Furthermore, some alternative tests such as these for the phototoxicity test have an “Equivocal” category for judging the toxicity of chemicals. These facts have made it difficult to interpret the value of the measures.

Therefore, in this paper we propose new measures to evaluate the alternatives, which depend on a set of chemicals rather than on both factors, and can treat data which have “Equivocal” category. We also propose their confidence intervals, which are measures of their precision.

Key words: relevance, inter-laboratory validation study, sensitivity, specificity, accuracy, confidence interval

Introduction
Recently, due to an increasing social concern for animal welfare, a lot of alternative animal tests have been proposed, and in order to examine their feasibility and practicality various inter-laboratory validation studies have been conducted (e.g. Ray et al., 1994; Spielmann et al., 1998). Generally, the primary purpose of the validation study is to evaluate both the relevance and reliability of a proposed alternative test from the results of experiments using the alternative test (Balls et al., 1999). Sensitivity, specificity and accuracy are measures to determine the effectiveness of the alternative test when both the alternative and the animal tests have a binary classification for judging toxicity of chemicals, as “Positive” and “Negative”. These are well known measures which have been widely used to evaluate the relevance of the alternative test in many validation studies (e.g. Balls et al., 1990; Roy et al., 1994; Spielmann et al., 1998).

However, two points should be taken into consideration concerning the interpretation of the summarized data from validation studies. The first point is that the values in the 2 by 2 table, which summarizes data, depend not only on a selected set of chemicals in the study but also on the number of participant laboratories. The other point is that a category for “Equivocal” produced from some alternative tests such as these for the phototoxicity test, which is neither a “Positive” nor “Negative” category, is often provided. For instance, the test guideline of the in vitro 3T3 NRU phototoxicity test states that ‘a test substance with a PIF < 2 or an MPE < 0.1 predicts: ”no phototoxicity”. A PIF >2 and < 5 or an MPE > 0.1 and < 0.15 predicts: “probable phototoxicity” and a PIF > 5 or an MPE > 0.15 predicts: “phototoxicity”.’ where the PIF and the MPE are measurements of phototoxicity for the test (OECD, 2004). In this case, since there was a range suggesting similar performance when several cut-off points were examined, the
category “probable phototoxicity” as “Equivocal” was set (Peters and Holzhütter, 2002). Sugiyama, et al (1994) proposed a red blood cell hemolysis assay to predict phototoxicity of chemicals, and they classified photohemolysis into three categories, +, ± and -.

In this paper, we discuss the above two points for the measures, sensitivity, specificity and accuracy, and propose new measures for evaluating the relevance of an inter-laboratory validation study. We also construct an equation for their confidence intervals, which measure their precision of them (Altman, 2000a).

**Methods**

**Definition for sensitivity, specificity and concordance**

Table 1 shows a 2 by 2 table. Sensitivity is defined as the proportion of chemicals judged as positive by an alternative test in which the chemicals are identified as positive by an animal test. When data is summarized as in table 1, sensitivity is calculated by $a / (a + b)$. Specificity is defined as the proportion of chemicals judged as negative by the alternative test in which the chemicals are identified as negative by the animal test. The measure is $d / (c + d)$. Accuracy is defined as the proportion of a corresponding number of chemicals by the judgment of the alternative test in which all the chemicals are identified by the animal test. The measure is obtained as $(a + d) / (a + b + c + d)$.

It is rarely noted that the values of these measures depend on the selected set of chemicals. If the toxicity of the selected chemicals in a validation study has only the strongest classes and the weakest classes, the values of these measures would be expected to be higher when the assessed alternative test has a good correlation to the targeted animal test. If the researchers conducting the validation study can select test chemicals before the experiments on the alternative test, they can control the measures. On the other hand, if they choose many middle class chemicals in the study, the measures may show an inferior result compared to our expectation. Even if the chemicals are selected by an external person not directly involved in the study, the values of these are dependent on the selected chemicals. Thus, we should interpret the values of these as conditional proportions dependent on the set of selected chemicals in the study.

**Motivated data**

Table 2 shows a typical form of data from a validation study. The symbols "P", "E" and "N" in the Table mean “Positive”, “Equivocal” and “Negative” to be judged by *In vivo* test or the alternative test.

<table>
<thead>
<tr>
<th>Chemical</th>
<th>In vivo</th>
<th>Laboratory</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>B</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>C</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td>D</td>
<td>P</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>G</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>H</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>I</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

Table 1. The 2 by 2 table.

<table>
<thead>
<tr>
<th>Alternative test</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td>Negative</td>
<td>b</td>
<td>d</td>
</tr>
<tr>
<td>a+b</td>
<td>c+d</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. A motivated example of a inter-laboratory validation study.

Symbols: P, positive; E, equivocal; N, negative
ternative test. This data is from an actual validation study conducted in Japan which has not been published yet. In the study, nine chemicals were tested by six laboratories. In order to meet an increasing demand for assessing test chemicals, the laboratories used the alternative test for as many chemicals as possible. However, due to time and financial constraints, all the laboratories did not experiment applying the alternative test for all the chemicals. In view of animal welfare, data from animal tests is usually obtained from some published articles and/or databases including data from past experiments; animal tests are rarely conducted in validation studies. Therefore there is usually only one result for each chemical. On the other hand, some results for each chemical in an alternative test are obtained from the inter-laboratory study.

When the measures, sensitivity, specificity and accuracy, are calculated, data, as in Table 2, is summarized by a 2 by 2 table, in which a result from a chemical in a laboratory for an alternative test corresponds to a result from using the same chemical in an animal test; total for four cells in the 2 by 2 table is 36 as is the case in Table.

Consideration of two points
Furthermore, in addition to the fact that the measures are a conditional proportion of a set of chemicals, we also have to consider that these depend on the number of laboratories conducting inter-laboratory validation studies. However, when data is summarized by a 2 by 2 table, as in Table 2, distinguishing between the two factors, the set of selected chemicals and the number of participant laboratories is overlooked. Then the interpretation of the value is difficult. For instance, the sensitivity from a laboratory which has examined ten positive chemicals is 100% when all the chemicals are judged positive. The sensitivity from the ten laboratories which examined a positive chemical is also 100% when all laboratories judge positive for the chemical. Should we regard both sensitivities as the same?

Some people often use only the values of these measures from different validation studies without taking into consideration these factors, when they compare the alternatives.

The presence of an “Equivocal” category is another difficulty involved in interpreting the measures. Since these measures are based on the assumption that the results of both tests are expressed as binary categories, often data for “Equivocal” is artificially changed: these are eliminated from the numerator; data for “Equivocal” is relabeled as “Positive” (e.g. Sugiyama et al., 1994). The value of the measures depends on which treatment is used.

Proposed methods
We propose similar measures to sensitivity, specificity and accuracy, which take into consideration and deal with the previous two points.

Firstly, we consider the relationship between two factors; chemical and laboratory. Since several laboratories experiment using the alternative test for a same chemical in the inter-laboratory validation study, data from the validation study has a hierarchical structure between two factors. In the proposed methods, the factor of chemical becomes a basic unit.

Suppose \( y_{ij} \) is a variable to explain the result from an alternative test, and \( x_i \) is a variable to explain the result from an animal test, where subscript i and j mean the i th chemical (i = 1, 2, ..., n) and the j th laboratory (j = 1, 2, ..., m_i) respectively. The variable \( y_{ij} \) take 1 for the “Positive” result, 0 for the “Negative” and 0.5 for the “Equivocal”, when the alternative test is experimented for the i th chemical in the j th laboratory. The variable \( x_i \) is 1 for the “Positive” result of the targeted animal test, and 0 for the “Negative” result. We initially define \( p_i \) as a proportion for the number of positive results in the i th chemical for the alternative test, that is

\[
P_i = \frac{\sum_j y_{ij}}{m_i}.
\]

As shown the appendix A, we can calculate the variance, \( V(p_i) \), based on the assumption of trinomial distribution.

Using \( p_i \), we also define \( q_i \) as

\[
q_i = x_i p_i + (1 - x_i)(1 - p_i).
\]

Note that \( q_i \) is a measure for the reliability of the i th chemical. The alternative test shows good reliability when the value of \( q_i \) is close to 1.

Finally, we define three measures which correspond to sensitivity, specificity and accuracy, using \( p_i \), and call these measures \( Psn \), \( Psp \) and \( Pac \), respectively;

\[
Psn = \frac{\sum_i x_i p_i}{\sum_i x_i}
\]
\[ P_{sp} = \sum_i \left(1 - x_i \right) \left(1 - p_i \right) / \sum_i \left(1 - x_i \right). \tag{4} \]

and

\[ P_{ac} = \sum_i q_i / n = \left( \sum_i x_i p_i + \sum_i \left(1 - x_i \right) \left(1 - p_i \right) \right) / n. \tag{5} \]

These measures are also consistent with sensitivity, specificity and accuracy which are based upon the number of tested chemicals, when each \( p_i \) becomes 1 or 0.

Using \( V(p_i) \), we can also calculate the variance for the proposed measures, \( V(P_{sn}) \), \( V(P_{sp}) \) and \( V(P_{ac}) \) (Appendix B). Furthermore we can construct the 95% confidence interval as follows;

\[ P_{sn} \pm 1.96 \times \sqrt{V(P_{sn})}, \quad P_{sp} \pm 1.96 \times \sqrt{V(P_{sp})}, \quad \text{and} \quad P_{ac} \pm 1.96 \times \sqrt{V(P_{ac})}. \]

Note that the proposed measures only depend on the set of chemicals in the validation study but not on the factor of laboratory.

**Results**

Table 3 displays a 2 by 2 table of the summarized data of Table 2, in which “Equivocal” for the alternative test is assigned to “Positive”. Based on Table 3, sensitivity, specificity and accuracy become 93.7% (15/16), 40.0% (8/20) and 63.9% (23/36) respectively. When “Equivocal” is eliminated from the numerator, sensitivity and accuracy are 62.5%(10/16) and 50%(18/36) respectively.

Table 4 shows \( y_{ij}, x_i, m_i, p_i, q_i \) and \( V(p_i) \) corresponding to Table 2. According to the table, the proposed measure, \( P_{sn}, P_{sp} \) and \( P_{ac} \) are obtained as 78.1% (3.38/4), 47.5% (2.37/5) and 61.1%(5.75/9) respectively. The 95% confidence intervals of the proposed measures become

\[
\begin{array}{|c|c|c|}
\hline
\text{Animal test} & \text{Positive} & \text{Negative} \\
\hline
\text{Alternative Test} & \text{Positive} & \text{Negative} \\
\hline
\text{Positive} & 15 & 12 \\
\text{(P-16, F-5)} & (P-9, F-3) \\
\hline
\text{Negative} & 1 & 8 \\
\hline
16 & 20 \\
\hline
\end{array}
\]

Table 4. Scores for data in Table 2, and \( p_i, q_i \) and \( V(p_i) \).
24.7 to 100%, 9.5 to 85.5% and 31.2 to 91.1%, respectively.

When we set 0.5 for “Equivocal” and calculate the values of sensitivity, specificity and accuracy, these are consistent with $P_{sn}$, $P_{sp}$ and $P_{ac}$, respectively. Though both measures are identical when the number of laboratories doing experiments for all the chemicals is the same, generally these return different values.

**Discussion**

There are several recommended statistical methods for data analysis when alternative tests are assessed (Festing, 2001). However, it is seldom that statistical methods for the inter-laboratory validation study have been developed. The proposed measures in this paper are for the inter-laboratory validation study.

Sensitivity, specificity and accuracy are commonly used in studies on diagnostic test studies in medicine (Altman, 1994). In these, the results of diagnostic tests between separate groups of patients with and without a target disease are summarized by a 2 by 2 table. One of the biggest differences compared to the situation for evaluating an alternative test is that the researcher is able to control the values of these measures by selecting the set of chemicals for the validation study. Our proposal measures also have the same feature.

This feature also affects the construction of the confidence intervals. Though the way to constructing the confidence intervals for these measures is also well known (Altman, 2000b), it is assumed that each observation is followed by an independent and identical distribution. It would be difficult to make the assumption that experimental results from different chemicals have the same distribution because toxicity for chemicals selected in a study is usually widespread. However, the proposed measures solved the problem and we can construct their confidence intervals.

The confidence interval is an index to show the precision of measures, and usually it is affected by the number of data. Thus, $V(p_i)$ depends on the number of laboratories on the $i$th chemical, $m_i$; $V(P_{sn})$, $V(P_{sp})$ and $V(P_{ac})$ also depend on the number of chemicals, $n$ beside $m_i$. Therefore, if a validation study with a few laboratories is conducted, the confidence intervals of the proposed measures, $P_{sn}$, $P_{sp}$ and $P_{ac}$, may become wider. Though this feature is a limitation of the proposed measure, we could obtain more reliable estimates if there are replications for a chemical in a laboratory.

**Acknowledgment**

This study was supported by a grant from the Ministry of Health, Labour and Welfare.

**References**


Suppose \( y_{ij} \) is a score of the \( i \)th chemical in the \( j \)th laboratory, which takes 1 for “Positive”, 0 for “Negative” and 0.5 for “Equivocal”; \( r_i^{(N)}, r_i^{(E)} \) and \( r_i^{(P)} \) are frequencies for “Negative”, “Equivocal” and “Positive” of the \( i \)th chemical; \( m_i \) is the number of laboratories which experimented the \( i \)th chemical. We assume that \( \left( r_i^{(N)}, r_i^{(E)}, r_i^{(P)} \right) \) follows a trinomial distribution with parameters, \( m_i, \pi_i^{(N)}, \pi_i^{(E)} \) and \( \pi_i^{(P)} \): 
\[
\left( r_i^{(N)}, r_i^{(E)}, r_i^{(P)} \right) \sim \text{tri}(m_i, \pi_i^{(N)}, \pi_i^{(E)}, \pi_i^{(P)}),
\]
where \( \pi_i^{(N)} + \pi_i^{(E)} + \pi_i^{(P)} = 1 \).

The sum of \( y_{ij} \) for the \( i \)th chemical can write use \( \left( r_i^{(N)}, r_i^{(E)}, r_i^{(P)} \right) \) is:
\[
\sum_j y_{ij} = 0 \times r_i^{(N)} + 0.5 \times r_i^{(E)} + 1 \times r_i^{(P)}.
\]
Therefore, the expected value of \( \sum_j y_{ij} / m_i \) is:
\[
E\left[ \sum_j y_{ij} / m_i \right] = 0 \times \pi_i^{(N)} + 0.5 \times \pi_i^{(E)} + 1 \times \pi_i^{(P)}
= 0.5 \times \pi_i^{(E)} + 1 \times \pi_i^{(P)}.
\]
The variance is:
\[
V\left[ \sum_j y_{ij} / m_i \right]
= 0 \times \frac{\pi_i^{(N)}(1-\pi_i^{(N)})}{m_i} + 0.5 \times \frac{\pi_i^{(E)}(1-\pi_i^{(E)})}{m_i} + 1 \times \frac{\pi_i^{(P)}(1-\pi_i^{(P)})}{m_i}
- 2 \times \left( 0 \times 0.5 \times \frac{\pi_i^{(N)}(1-\pi_i^{(N)})}{m_i} \right)
- 2 \times \left( 0.5 \times 1 \times \frac{\pi_i^{(E)}(1-\pi_i^{(P)})}{m_i} \right)
= \left( \frac{0.5}{m_i} + \frac{1}{m_i} \right) \pi_i^{(E)}(1-\pi_i^{(E)})
= \left( \frac{0.5}{m_i} + \frac{1}{m_i} \right) \pi_i^{(P)}(1-\pi_i^{(P)}).
\]

Since we can’t know the values of \( \pi_i^{(N)}, \pi_i^{(E)} \) and \( \pi_i^{(P)} \), we need to estimate these values as \( \hat{\pi}_i^{(N)}, \hat{\pi}_i^{(E)}, \hat{\pi}_i^{(P)} \). When the maximum likelihood method is used, the estimates are \( \hat{\pi}_i^{(N)} = r_i^{(N)} / m_i \), \( \hat{\pi}_i^{(E)} = r_i^{(E)} / m_i \) and \( \hat{\pi}_i^{(P)} = r_i^{(P)} / m_i \). Then, by applying the estimates to (A3) and (A4), the values of \( p_i \) and \( V(p_i) \) are:
\[
p_i = 0.5 \times \frac{r_i^{(E)}}{m_i} + 1 \times \frac{r_i^{(P)}}{m_i} = \sum_j y_{ij} / m_i, \quad (A5)
\]
\[
V(p_i) = \frac{1}{4} \times \frac{r_i^{(E)}(r_i^{(E)} - r_i^{(P)})}{m_i} + \frac{1}{4} \times \frac{r_i^{(P)}(r_i^{(P)} - r_i^{(E)})}{m_i}.
\]
Appendix B

Since $p_i$ are independent for each other, using $V(p_i)$ the variances of Psn, Psp and Pac, $V(Psn)$, $V(Psp)$ and $V(Pac)$, are obtained as:

$$V(Psn) = V\left(\sum x_i p_i / \sum x_i\right)$$
$$= \sum_i x_i^2 V(p_i) / (\sum x_i)^2$$
$$= \sum_i x_i V(p_i) / (\sum x_i)^2,$$  \hspace{1cm} (B1)

$$V(Psn) = V\left(\sum (1-x_i)(1-p_i) / \sum (1-x_i)\right)$$
$$= \sum_i (1-x_i)^2 V(1-p_i) / (\sum (1-x_i))^2$$
$$= \sum_i (1-x_i) V(p_i) / (\sum (1-x_i))^2,$$  \hspace{1cm} (B2)

$$V(Pac) = V\left(\sum x_i p_i + \sum (1-x_i)(1-p_i)/n\right)$$
$$= \sum_i x_i V(p_i) + \sum (1-x_i) V(p_i) / n^2$$
$$= \sum_i V(p_i) / n^2.$$  \hspace{1cm} (B3)