Int. Arch. Arbeitsmed. 32, 167—190 (1974)

Validity of Biological Tests in Epidemiological Toxicology

R. L. Zielhuis* and M. M. Verberk**
Coronel Laboratory, Faculty of Medicine, University of Amsterdam

Received July 13, 1973 / Accepted September 26, 1973

Summary. The authors emphasize the need to introduce the concept of validity (sensitivity and specificity) of biological test methods in epidemiological toxicology (occupational and public health). Up till now too often relevant information is lost, because the frequency distribution of individual data is not taken into account. The method of calculating parameters of validity is demonstrated. These parameters add relevant information for determining the feasibility of test methods; they provide valuable information not presented by classical statistical treatment of data. Several examples have been worked out to elucidate the approach.

Key words: Epidemiological toxicology — Biological monitoring — Validity — Sensitivity — Specificity.

1. Epidemiological Toxicology

Epidemiological toxicology studies the relationship between parameters of exposure and parameters of response in groups of human subjects, taking into account endogeneous and exogeneous co-determinant factors. Exposure may be described in terms of external load (EL: amount of agent, offered to the body per unit of time) and parameters of internal load (IL: levels of noxe itself or of its metabolites in biological specimen, e.g. blood, urine, hair, saliva). Parameters of response (R) indicate the effect of IL on biological systems; such parameters are to be found e.g. in blood cells, biochemical levels, functional performance, physical signs, subjective symptoms. Epidemiological toxicology studies R as function of EL or IL, in workers or in the general population.

Biological tests are used to measure qualitatively and quantitatively exposure or response. In recent years application of biological tests in monitoring groups of exposed subjects receives increasingly more atten-

* Professor in public health, particularly in regard to occupational medicine and environmental health, Faculty of Medicine, University of Amsterdam.

** Research worker Coronel Laboratory.

tion. Biological threshold limit values are being developed and — at least for some types of exposure — they may take a prominent place in preventive health care alongside TLV's, MAC's (workroom air), Air quality standards (ambient air), ADI's (food stuff), etc. In this paper the authors discuss the validity of these tests. Emphasis will be put upon the concepts of sensitivity and specificity (as an indication of validity).

2. Validity, Sensitivity, Specificity

Up to now the concept of validity has been used in epidemiology mainly to evaluate biological tests as predictors of disease, e.g. electrocardiogram as predictor of (probable development of) coronary infarction. According to Mac Mahon and Pugh (1970) *validity* refers to "the extent to which subjects in a case control study are correctly classified as to the extent to which a situation as observed reflects the true situation". *Sensitivity* is "the extent to which patients who truly manifest a characteristic are so classified". *Specificity* is "the extent to which patients who do not manifest such a characteristic are correctly classified". A high sensitivity corresponds to a low probability of false negative data, a high specificity implies a low probability of false positive data. The "extent to which" and the probability can be expressed in *quantitative terms*.

The following table presents the general frame of approach [biological data referred to as "indicator (I)", exposure as "true situation (T)"]

True situation (T)	Indicator (I)		
	present $I+$	not present $I-$	
present $T+$	$i+t+$	$i-t+$	$T+$
not present $T-$	$i+t-$	$i-t-$	$T-$
	$I+$	$I-$	

$$\text{sensitivity} = \frac{i+t+}{T+} \quad (= \text{percentage of true positive data})$$

$$\text{percentage of false negative data} = \frac{i-t+}{T+}$$

$$\text{specificity} = \frac{i-t-}{T-} \quad (= \text{percentage of true negative data})$$

$$\text{percentage of false positive data} = \frac{i+t-}{T-}$$

Now we define validity as the sum of *sensitivity and specificity*.

For the purpose of simplicity indicators and true situation are supposed to be — qualitatively — *present* or *not*; it means that in most cases one has to define on a quantitative scale a "point of discrimination" between presence or absence of indicator or true situation.

Epstein (1967) added other indices of validity:

Predictive value+ $= \frac{i+t+}{I+}$, i.e. probability for subjects with indicator present ($I+$) to belong to the specific situation $T+$, in other words percentage true positives within total number of subjects with positive indicators.

Predictive value— $= \frac{i-t+}{I-}$, i.e. probability for subjects with indicator absent ($I-$) to belong to the specific situation $T+$, in other words percentage false negatives within total number of subjects with negative indicators.

Risk ratio $= \frac{\text{predictive value}+}{\text{predictive value}-}$, i.e. relative probability for subjects with indicator present ($I+$) compared with subjects with indicator absent ($I-$) to belong to the specific situation ($T+$).

For calculation of predictive values one should have an equal number of subjects in $T+$ and $T-$; otherwise the predictive values also become dependent on the number of $T+$ and $T-$.

3. Mathematical Approach

To get an idea of normal and extreme values of sensitivity (*se*), specificity (*sp*), positive predictive value ($p+$) and negative predictive value ($p-$), Fig. 1 may be useful. Every possible 2×2 diagram in which $T+ = T- = a$ can be represented by one special point in this figure, depending upon the value of *se* and *sp*; $p+$ and $p-$ are directly derived from the value of *se* and *sp*.

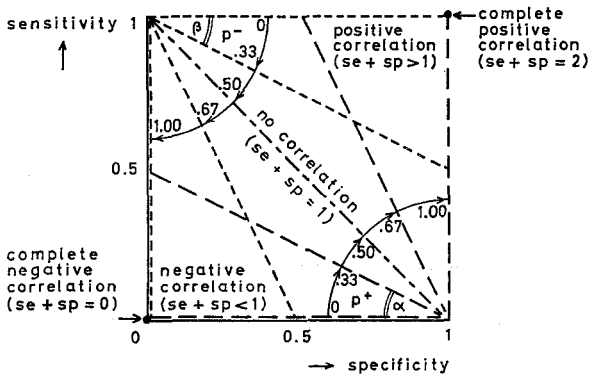

Fig. 1. Validity (V) as function of sensitivity (*se*) and specificity (*sp*); $V = se + sp$

We observe different situations[1]:

a) Complete Positive Correlation. All positive indicators are inside the true situation and all negative indicators are outside the true situation. There are no false negative or false positive test data.

	$I+$	$I-$	
$T+$	a	0	a
$T-$	0	a	a
	a	a	

It is evident that

$$se = \frac{i+t+}{T+} = \frac{a}{a} = 1$$

$$sp = \frac{i-t-}{T-} = \frac{a}{a} = 1$$

$$p+ = \frac{i+t+}{I+} = \frac{a}{a} = 1$$

$$p- = \frac{i-t+}{I-} = \frac{0}{a} = 0$$

$$\boxed{se + sp = 2}$$

In Fig. 1 this situation is represented by the point of the upper right corner.

b) Complete Negative Correlation. All positive indicators are outside the true situation and all negative indicators are inside the true situation. There are only false negative and false positive test data.

	$I+$	$I-$	
$T+$	0	a	a
$T-$	a	0	a
	a	a	

It is evident that

$$se = \frac{i+t+}{T+} = \frac{0}{a} = 0$$

$$sp = \frac{i-t-}{T-} = \frac{0}{a} = 0$$

$$p+ = \frac{i+t+}{I+} = \frac{0}{a} = 0$$

$$p- = \frac{i-t+}{I-} = \frac{a}{a} = 1$$

$$\boxed{se + sp = 0}$$

In Fig. 1 this situation is represented by the point of the lower left corner.

1 a, b and c indicate number of subjects in each category; any number may be assumed.

c) No Correlation. As many (b) positive indicators are inside the true situation as there are outside (b). There are 50% false test results.

	$I+$	$I-$	
$T+$	b	$a-b$	a
$T-$	b	$a-b$	a
	$2b$	$2(a-b)$	

It is evident that $se + sp = \frac{i+t+}{T+} + \frac{i-t-}{T-} = \frac{b}{a} + \frac{a-b}{a} = \frac{a}{a}$

$$p+ = \frac{i+t+}{I+} = \frac{b}{2b} = \frac{1}{2} \qquad \boxed{se + sp = 1}$$

$$p- = \frac{i-t+}{I-} = \frac{a-b}{2(a-b)} = \frac{1}{2}$$

In Fig. 1 these situations are represented by the line from the upper left to the lower right corner.

d) Any Positive (or Negative) Correlation. More (or less) positive indicators are inside the true situation than there are outside. There are less (or more) than 50% false test results.

	$I+$	$I-$	
$T+$	b	$a-b$	a
$T-$	c	$a-c$	a
	$b+c$	$2a-b-c$	

$b > (\text{or } <) c$

$$se + sp = \frac{i+t+}{T+} + \frac{i-t-}{T-} = \frac{b}{a} + \frac{a-c}{a} = \frac{a+b-c}{a} = 1 + \frac{b-c}{a}$$

$b > (\text{or } <) c \rightarrow \frac{b-c}{a}$ is positive (or negative)

so $\boxed{se + sp > (\text{or } <) 1}$

$$p+ = \frac{i+t+}{I+} = \frac{b}{b+c} \rightarrow \frac{1}{p+} = \frac{b+c}{b} = 1 + \frac{c}{b}$$

$b > (\text{or } <) c \rightarrow \frac{c}{b} = < (\text{or } >) 1$

so $\frac{1}{p+} < (\text{or } >) 2 \rightarrow p+ > (\text{or } <) \frac{1}{2}$

$p- = \qquad \rightarrow p- < (\text{or } >) \frac{1}{2}$

In Fig. 1 this situation is represented by the upper right (or lower left) triangle.

Moreover it can be calculated that a certain value of $p+$ (or $p-$) is represented in Fig. 1 by a line through the lower right (or upper left) corner. This implies that a certain combination of *se* and *sp* automatically determines $p+$ and $p-$. To facilitate calculation a is defined as 1.

	$I+$	$I-$	
$T+$	b	$1-b$	1
$T-$	$1-c$	c	1
	$1+b+c$	$1-b+c$	

now $se = b, \quad sp = c$

$$p+ = \frac{b}{1+b-c} = \frac{se}{1+se-sp} \rightarrow se = p^+ + p^+se - p^+sp \rightarrow se(1-p^+)$$

$$= p^+ - p^+sp \rightarrow (se-0) = \frac{-p^+}{1-p^+}(sp-1)$$

This aequation represents a collection of lines which all go through the point ($se = 0$, $sp = 1$), and are determined by p^+ as follows:

$$tg\,\alpha = \frac{p^+}{1-p^+}.$$

For instance: $p^+ = \frac{1}{2} \rightarrow tg\,\alpha = \frac{0.5}{0.5} = 1 \rightarrow = 45°$ (line of no correlation); if the p^+ line goes through the point ($se = 0.5$, $sp = 0$) $\rightarrow p^+ = \frac{0.5}{1+0.5} = 0.33$.

In the same way it can be calculated that: $(se-1) = \frac{p^-}{p^- - 1}(sp-0)$ representing a collection of lines that all go through the point ($se = 1$, $sp = 0$) and are determined by p^- as follows: $tg\,\beta = \frac{p^-}{p^- - 1}$.

To summarize: to be valid, an indicator has to have the following characteristic: the sum $se + sp > 1$ and as much approaching 2 as possible. For instance when one allows $\leqq 10\%$ false test data, this sum should be $\geqq 1.80$. As for the composition of this sum it should be said that when it is important to detect (people exposed to) the true situation (being a threat to health) *se* should approach to 1 at the cost of *sp*. In this case *se* could be 0.98 and $sp = 0.82$; this means 2% false negatives (overlooked threats) and 18% false positives (no real threats).

4. Validity as a Selection Criterion

Epidemiological toxicology has to a great extent shifted from the study of disease distribution due to excessive exposure amongst workers to evaluation of subclinical effects due to long term low level exposure amongst workers and the general public. These effects nowadays are often sought at biochemical, microscopical and molecular level. Attention has also shifted to prevention: determination of early indicators of internal exposure levels and of early precursors of more serious effects.

Toxicology does not necessarily deal with patients, but mostly with subjects who do not feel ill in the classical sense. However, it is a well known fact that levels of biological parameters (e.g. enzyme activity, subjective symptoms, number of blood cells) are affected by a multitude of external and internal factors; low-level exposure may be only one of the intervening factors; the probability of false positive test data increases: not exposure, but maybe season, physical activity, sex, age determine occurrence of positive test data.

In addition, indicator levels may point to changes in several biological systems, e.g. increase in reticulocytes due to blood loss ánd to disturbed porphyrin synthesis, LDH due to heart ánd liver disease, decreased psychomotor performance due to toxic effects on nervous system ánd to cerebral atherosclerosis. In epidemiological toxicology one therefore needs tests which are sensitive and specific both to exposure and to effects on biological systems.

The predictive validity of tests add an extra element in the choice of methods in modern epidemiological toxicology. It may enlarge the probability of receiving adequate (i.e. true) answers to following questions:

1. Given a specified parameter, does there exist exposure to specified agents, and if so, to what extent?

2. Given a specified exposure, does there exist any specified response, and to what extent?

If a biologic indicator "truely" indicates exposure to a specific agent, then it is also "true" that subjects exposed to this particular agent (to the degree as relevant to the valid test method) probably have as a characteristic the presence of this parameter, and that non-exposed subjects probably do not have this characteristic. If a parameter of response has a high validity in predicting exposure, one may conclude that this exposure causes this response (qualitative and quantitative). One can use the data of a particular study to calculate the validity of $I+$ as *indicator* of the true situation $T+$. If this validity also appears to exist in a number of other studies, one can attach a general significance to the validity of $I+$: it can be used as a general *predictor*. In statistics one also distinguishes σ and μ as parameters of one study, and s and m as generally valid parameters.

Validity of a test may also be a criterion in selecting the most appropriate test from a set of possible tests. Information to be gained should not only be valid, but should also be gathered with due regard to input of manpower, budget, expenditure of workers' time (efficiency). The validity of simple or cheap test methods should be determined in regard to specified "true situations". Williams *et al.* (1968a) introduced the concept of comparative merits of various tests indirectly measuring

Pb-exposure; they also calculated the comparative merits per unit cost. If validity is known, costs of measurement may add another element in the choice of methods.

5. Other Uses of Same Terms

The terms sensitivity and specificity are also applied in other concepts:

analytical sensitivity: detection level of a specified technique

analytical specificity: absence of effects of other compounds on levels determined

Improvement of detection level and of discriminative power promotes sensitivity and specificity as epidemiological concepts.

Sensitivity is also used in the sense of hypersusceptibility, i.e. decreased biological threshold level; this indicates a property of exposed subjects, and not of test methods.

In this paper we only discuss *epidemiological* sensitivity and specificity.

6. Requirements in Application of Tests and Presentation of Data

If one wants to make use of biological tests, one has to observe some strict conditions:

1. One cannot determine validity of tests if predictive indicators (I) and true situation (T) are not exactly specified in qualitative and if possible in quantitative terms.

2. Data should not only be given as average (+ standard deviation), but as individual data; at least stratified frequency distributions should be presented. Many data do not follow a normal (i.e. Gaussian) distribution. Biological levels of xenobiotic metals usually do not follow a normal distribution in contrast to levels of e.g. essential metals (Liebscher *et al.*, 1968; Kubota *et al.*, 1968). Tails of distributions may give more relevant data than mean or median values. Cut off scores, e.g. percentage Hb < 13.0 g%, leucocytes $< 4000/\text{mm}^3$, SGPT levels exceeding 95% confidence range (in non exposed controls) may provide a sensitive indicator. Average levels do not specify the number of workers exceeding biological threshold limits; percentile scores (e.g. levels in $> 50\%$, $> 90\%$ of controls) may be of great value.

3. The methods for establishing indicator level (I) and true situation (T) should be clearly specified e.g. in regard to time of sampling in relation to exposure, method of sampling (personal versus static sampler); urinary elimination data based upon excretion pro time (μg/min, mg/24 hrs) or based upon creatinin excretion may provide a lower variation coefficient than levels based upon concentration as such, even if correction on specific gravity has been applied (van Rees *et al.*, 1968).

7. Exposure Tests

A multitude of tests is being applied for quantitative evaluation of the internal chemical load (*IL*): one determines either the toxic agent itself or its metabolites in blood, urine, saliva, hair, exhaled air. How valid are these tests (*I*) as predictors of the true external exposure (*T*)? In order to study validity one has to examine both *I* and *T* for the same group of subjects. One example will be discussed.

Williams *et al.* (1969) examined the relationship between exposure to lead (PbA = Pb in air, mg/m³, personal sampling) and PbB (μg Pb/100 ml) and PbU (μg Pb/l) in a group of workers from a lead accumulator factory (individual data). For sake of reasoning we assume PbA = 0.12 mg/m³ to be the acceptable level. We determined the validity of various PbB and PbU levels as indicators of unacceptable exposure (> 0.12 mg/m³). The data of this study can be regrouped as follows:

PbA	PbB				*n*
	0—40	> 40—60	> 60—80	> 80	
> 0.12	—	4	8	2	14
< 0.12	10	2	3	—	15

PbA	PbU				*n*
	0—60	> 60—120	> 120—160	> 160	
> 0.12	2	5	7	2	16
< 0.12	10	8	—	1	19

From these data we calculated the validity:

		Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
PbB	> 40	1.00	0.66	1.66
	> 60	0.72	0.80	1.52
	> 80	0.56	1.00	1.56
PbU	> 60	0.88	0.53	1.41
	> 120	0.56	0.95	1.51
	> 160	0.12	0.95	1.07

PbB levels have a higher validity than PbU levels as indicator of unacceptable exposure. PbB > 40 has a maximum sensitivity (all individuals with PbA > 0.12 have PbB > 40); however, specificity is moderate

(also PbB > 0.40 in subjects with PbA < 0.12); PbB > 80 is highly specific, i.e. no individuals with PbA < 0.12 have PbB > 80, however, sensitivity is moderate (many false negatives). If one wants to be certain that all subjects with PbA > 0.12 are selected out of a universe of exposed workers PbB > 40 will serve this objective, however, at the cost of a number of false positives. If on the other hand one wants to select only individuals with PbA > 0.12, then PbB > 80 will serve this objective, however, with many false negatives. It depends on the objective of the investigation, whether one puts emphasis upon sensitivity or upon specificity.

It should be pointed out that the data of Williams *et al.* have only been used as an example; the conclusions are only valid for this study, and for the levels chosen. For general study of predictability more data should be scrutinized.

Most publications on biological monitoring for exposure to various agents do not give individual data or stratified distributions; very few studies have been set up with the objective to study validity. If one wants to establish Biological Threshold Limit Values (BTLV) based upon estimation of body burden, and so of exposure, one should know the validity of such limits, in other words: one should know quantitatively the probability that subjects with unacceptable exposure exceed BTLV, and subjects with acceptable exposure do not exceed BTLV.

For some agents it is also possible to use parameters of response as indicator of internal and external exposure; this will be dealt with further on.

8. Haematological Responses

Haematology offers many parameters to be used in evaluation of response to toxic agents: haemoglobin, erythrocyte-, leucocyte-, thrombocyte-, reticulocyte-count, Heinz bodies, basophilic punctation of erythrocytes, lymphocytosis, and so on. Some parameters apparently are affected by many external factors, i.e. low specificity (many false positives).

One of the authors (Zielhuis, 1959) examined reticulocytosis and basophilia in lead exposed workers ($n = 513$) and in non-exposed controls ($n = 117$) *(example 2)*. If one uses as cut off scores the level not exceeded by 80% or 99% of the controls validity can be calculated:

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
retic-80	0.72	0.80	1.52
retic-99	0.42	0.99	1.41
bas-80	0.73	0.80	1.53
bas-99	0.31	1.00	1.31

Validity of reticulocytosis-80 and basophilia-80 as indicator of (response to) lead exposure as such is moderate, and even worse if high (99%) cut off scores are used. However, other data from the same investigation show that *within* the exposed group basophilia is a better indicator of *high* exposure than reticulocytosis. Evaluation of validity requires a rigid definition of the true situation to be predicted.

An other example comes from a field adjacent to toxicology: screening of workers exposed to ionizing radiation. Carpay (1970, 1972) examined exposed workers (group R) and non-exposed workers (nR) *(example 3)*; R did not differ from nR in leucocytes- and differential count, but there were significant differences in reticulocytes and Heinz bodies count (>4 bodies/eryth.):

	Reticulocytosis (‰)			Heinz bodies (%)		
	n			*n*		
R	392	5.03		101	4.8	
nR	597	4.55	$P < 0.05$	57	2.3	$P < 0.05$

Within R both levels increased with doses received. The data given already suggest a higher validity for Heinz bodies-count.

If we establish the cut off scores at levels not exceeded by about 50 and 90% of control subjects, validity is as follows:

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
reticuloc.-90 ≧ 9.0‰	0.10	0.91	1.01
-50 ≧ 5.0‰	0.49	0.55	1.04
Heinz b.-90 ≧ 4.0%	0.63	0.90	1.53
-50 ≧ 2.0%	0.99	0.56	1.55

These data show much more clearly (and quantitatively) the higher validity of Heinz bodies count in comparison to reticulocytes count in indicating (response to) exposure versus non-exposure, although P-values were similar. Leucocytes and differential count did not differ; these parameters apparently have no validity in the exposure range studied. If one wants to select exposed (affected) subjects from a universe of non-exposed + exposed workers, Heinz bodies count >2.0% yields few false negatives, i.e. exposure (response) is indicated, but one makes many mistakes (specificity per definition about 0.50). It depends on the risk of missing exposed (affected) subjects, whether one pays more attention to sensitivity or to specificity. One should try to develop tests which

have both a high sensitivity and a high specificity. It might be added that *within* exposed group R both reticulocytosis and Heinz bodies were dose-dependent.

9. Subjective Symptoms

Questionnaires are used to evaluate differences in subjective health; questions and test procedure require a rigid standardisation. Despite the pitfalls questionnaires are important tools to measure subjective health, particularly so because subjective experience of decreased wellbeing may give an early indication of response, before objective signs can be elucidated.

Ensberg (1973) measured subjective health in groups of workers exposed to a cocktail of pesticides (mainly in agriculture and horticulture) by means of a questionnaire *(example 4)*; the score (24 items) had the following distribution:

	Exposed $n = 85$	Matched controls $n = 86$	
Score 0—1	12	28	
2—3	24	30	
4—5	25	15	
6—7	10	5	
8—9	7	4	
>10	7	4	
Total number of + items	393	261	$P < 0.05$

Validity of various cut off scores could be calculated:

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
Cut off score >1	0.85	0.33	1.18
>3	0.58	0.67	1.25
>5	0.28	0.85	1.13
>7	0.17	0.91	1.08

With higher cut off score sensitivity decreases very fast, i.e. many false negatives (low score, exposed), and specificity increases, i.e. fewer false positives (high score, no exposure). A cut off score >3, i.e. dividing subjects into classes with 0—3 and >3 symptoms yields a sensitivity and specificity both about 0.60; this cut off score may serve as a first approach of evaluating subjective response due to exposure.

10. Physical Signs

Physical examination generally does not provide much relevant information in epidemiological toxicology. However, examination of the nervous system (hyporeflexia, areflexia, incoordination, tremor, muscular jerking, etc.) may offer early indicators of response in e.g. exposure to pesticides. Sophisticated instrumental methods (EEG, EMG) may even provide more pertinent information.

Czegledi-Jankō *et al.* (1970) examined 37 workers, exposed to lindane *(example 5)*. Blood lindane levels ranged from 0.002—0.340 ppm; in non exposed controls blood levels ranged from 0.003—0.017 ppm ($\bar{x} =$ 0.008 ppm). The exposed workers can be divided into two groups: $n = 17$ with levels $\geqq 0.020$ ppm, and $n = 20$ with levels < 0.020 ppm. The authors examined the nervous system (minor symptoms and signs or more serious symptoms such as muscular jerking and myoclonia with emotional changes). Non specific EEG changes (increased variation in frequency and amplitude of wave pattern) also occur in 15% (10—20) of the general population; the number of EEG changes in the second group has to be corrected for this. The data of this study can be summarized as follows:

	n	Neurol. signs		EEG changes		EEG changes corrected		Neurol. or EEG or both not corrected	
		+(+)	—	+(+)	—	+	—	+	—
Exposure $\geqq 0.020$ ppm	17	12	5	15	2	15	2	17	0
< 0.020 ppm	20	3	17	1	19	3	17	4	16

The parameters of validity are:

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
Neurol. signs	0.71	0.85	1.56
EEG changes uncorrected	0.88	0.95	1.83
EEG changes corrected	0.88	0.85	1.73
Neurol. or EEG or both, uncorrected	1.00	0.80	1.80

In lindane exposed workers with blood lindane levels $\geqq 0.020$ ppm examination of the nervous system, particularly when combined with

EEG, yields a highly valid test; moreover, these data also demonstrate the effect of lindane on the nervous system, if blood lindane levels are > 0.020 ppm.

Ensberg (1973) examined a group of workers ($n = 85$) with long term exposure to a cocktail of pesticides, mainly in agri- and horticulture; they were compared with a matched (age, social economic class) control group ($n = 86$) *(example 6)*. A standardised neurological examination (no EEG) resulted in following data:

	n	Subj. with neurol. signs	Only hypo- or areflexia
Exposed	85	37	21
Non exposed	86	34	14

Differences not significant at $P < 0.05$.

Validity is as follows:

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
Neurological signs (hypo-areflexia, areflexia, tremor, incoordination)	0.44	0.60	1.04
Only hypo-areflexia	0.25	0.85	1.10

In this investigation the standard examination procedure apparently did not yield a valid indicator of neurological response in this group of workers with this type and this degree of exposure; disturbance of neurological health occurred only to a limited extent (many false negatives), and similar signs also were present in many non-exposed workers. Either neurological response was hardly present or the method used was not valid enough.

11. Enzyme Activity

Measurement of enzyme activity takes an increasingly prominent place in epidemiological toxicology. Some of the changes in activity occur as a consequence of many causitive factors, others may predominantly be due to uptake of one specific agent. The same may be said in regard to biological response: change in activity either due to altered function of various systems, e.g. LDH as indicator of disturbed heart- or liverfunction, or predominantly due to effect on one biological system, e.g. decreased δ-aminolaevulinic acid dehydratase in erythrocytes (ALAD) as indicator of disturbed porphyrine synthesis.

A change in activity may therefore indicate:

response of a specific system

1. due to a specific external factor,
2. due to a variety of external factors;

response of various systems

3. due to a specific external factor,
4. due to a variety of external factors.

1. and 3. will provide the most valid tests for indirect monitoring of internal and external chemical load.

Measurement of δ-aminolaevulinic acid dehydratase (ALAD) in erythrocytes — a parameter of response — is used as indirect parameter of internal exposure to lead (PbB). Hernberg *et al.* (1972) suggested to use it as "a poor man's method" for measuring PbB indirectly; the correlation coefficient ALAD-PbB = 0.9. Another parameter of response, δ-aminolaevulinic acid in urine (ALAU) also increases due to lead exposure. Both are indicators of disturbed porphyrinogenesis.

ALAD and ALAU may serve as indicators of excessive internal exposure, i.e. for general public PbB > 40 µg/100 ml, for workers PbB > 70 µg/100 ml. From data presented by Haeger-Aronsen (1971a, b) *(example 7)* we calculated the validity of decreased ALAD (50 or 25% of "normal" average, i.e. 60 or 30×10^{-3} µmol PBG/mill. ery/hr, $n = 135$) and of increased ALAU (2 or 4 times normal average, i.e. 6 and 12 mg ALA/l, $n = 110$):

Prediction of	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
PbB > 40 µg Pb/100 ml			
ALAD-50	0.94	0.93	1.87
ALAD-25	0.75	0.97	1.72
ALAU-2	0.77	0.58	1.35
ALAU-4	0.34	1.00	1.34
PbB > 70 µg Pb/100 ml			
ALAD-50	1.00	0.73	1.73
ALAD-25	0.82	0.82	1.64
ALAU-2	0.94	0.51	1.45
ALAU-4	0.65	0.84	1.49

For use in public health, i.e. prediction of excessive internal load (PbB > 40 µg Pb/100 ml), a cut off score of 50% of average ALAD-level has a high validity; as indicator of excessive occupational exposure (>70 µg Pb/100 ml), sensitivity is very high, but specificity is moderate,

i.e. also decreased activity (ALAD-50) below PbB = 70 μg Pb/100 ml; ALAU > 6 and > 12 mg/l have a lower predictive validity, higher in occupational than in public exposure.

Agents present in air (inside and outside industry), food and water may induce enzyme activity as a means of selfdefence (adaptation); the same phenomenon may be elicited by drug medication. This may result in increased breakdown of not only the causative agent, but also of other exogeneous (e.g. drugs) or endogeneous compounds. Some examples: DDT may increase breakdown of phenobarbital and decrease sleeping time; it may also decrease body burden of dieldrin. The biological half life ($t\,½$) of drugs may provide an indicator of (response to) exposure to toxic agents.

Kolmodin-Hedman (1969) examined $t\,½$ of antipyrine in workers exposed to a cocktail of pesticides (DDT, chlordane, lindane, etc.) *(example 8)*: control group, $n = 33$, $t\,½ = 13.1 \pm 7.5$ hrs, range 5.2 to 35.0 hrs; exposed group $t\,½ = 7.7 \pm 2.6$ hrs, range 2.7—11.7 hrs; decrease of $t\,½$ is significant ($P = 0.01$). The distribution of $t\,½$ in controls is skewed; therefore a percentile score may provide a more valid indicator. We calculated the validity of two scores: $t\,½$-90 and $t\,½$-50, i.e. the level exceeded by 90 and 50% of controls.

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
$t\,½$-90 (< 6.0 hrs)	0.31	0.90	1.21
$t\,½$-50 (< 10.7 hrs)	0.85	0.48	1.33

Although as a group exposed workers differ significantly from controls, use of $t\,½$-90 selects only one third of exposed workers (many false negatives); use of $t\,½$-50 indicates exposure in most exposed individuals, however, 50% of non-exposed subjects are also selected. The validity of both scores is moderate.

Friborska (1969) examined alkaline phosphatase (L.A.P.) in peripheral leucocytes in workers exposed to trichloroethylene (their groups A, B and D: $n = 25$) *(example 9)* (for more data see paragraph 14). From the data presented the validity of L.A.P. > 95 U can be calculated: sensitivity = 0.89, specificity = 1.00; $se + sp = 1.89$. There are many conditions which may induce L.A.P., e.g. defence mechanisms against bacterial infections. However, in a group of "healthy" workers L.A.P. level appears to have a high validity in indicating exposure to trichloroethylene, maybe even in indicating the intensity of exposure.

Ensberg (1973) examined a group A ($n = 20$) of workers intensively exposed to a cocktail of pesticides in flower culture for more than 4 years

(example 10); in addition he examined a matched non exposed control group D ($n = 20$). This investigation yielded the following data:

	A ($n = 20$)	D ($n = 20$)
SGOT in U/l	8.5 ± 3.7*	11.6 ± 4.4
SGPT in U/l	7.8 ± 4.3*	11.9 ± 7.2
Alk. phosphatase in U/l	29 ± 6**	42 ± 21

* A → D: $P < 0.05$; ** $P < 0.02$.

We calculated validity for 50 and 90% percentiles, i.e. levels exceeded by 50 and 90% of control group D.

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
SGOT-50	0.85	0.50	1.35
-90	0.25	0.90	1.15
SGPT-50	0.65	0.50	1.15
-90	0.45	0.90	1.35
Alk. phosph.-50	0.85	0.50	1.35
-90	0.15	0.90	1.05

Apparently 90 percentile scores have a much lower sensitivity than 50-percentile scores in indicating exposure; SGOT and AP (median levels of controls) may be more valid indicators of exposure (and response) than SGPT; this could be a hypothesis for further study. The median levels of enzyme activity may be used as a rather sensitive, however, not specific response in workers exposed to a cocktail of pesticides. Although decrease of alk. phosphatase activity in A differed more significantly from D than decrease of SGOT activity, sensitivity of AP-50 and SGOT-50 was similar.

If one makes a combination of SGOT-50 and AP-50, then sensitivity = 0.70, specificity = 0.85, $se + sp = 1.55$. Combination of these two parameters increases validity.

Sensitivity and specificity as calculated should not be taken as generally applicable: the numbers of subjects examined were rather small; moreover the groups examined (exposed and controls) were taken from "healthy" workers; false positives due to existing diseases were therefore excluded.

12. Other Biochemical Parameters

Aside from measurement of enzym activity a multitude of other methods is applied for evaluation of biochemical responses, e.g. excretion of corticosteroids as an index of stress on adrenals, electrophoretic protein pattern, cholesterol in blood, thymol turbidity tests, etc.

Ensberg (1973) examined a group of workers from a factory producing many pesticides for use in agriculture *(example 11)*; group A ($n = 17$) was intensively exposed for more than 4 years, group B ($n = 14$) either slightly to moderately exposed >4 years, or intensively exposed <4 years; a control group D ($n = 29$) was also examined. In the electrophoretogram of groups A and B the percentage α_2 globine differed significantly from that in group D:

% α_2 globuline	group A 9.2 ± 1.9	$P < 0.001$
	group B 8.8 ± 2.0	$P < 0.01$
	group D 6.9 ± 1.8	

We calculated the validity of the levels not exceeded by 50 or 90% of group D:

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
Group A			
α_2-50	0.94	0.50	1.44
α_2-90	0.24	0.90	1.14
Group B			
α_2-50	0.79	0.50	1.29
α_2-90	0.21	0.90	1.11

For workers intensively exposed to a cocktail of pesticides the median α_2 level of controls appears to yield a highly sensitive indicator of exposure and response; 90-percentile levels appear to be useless in this way. This conclusion as such is only valid for the investigation discussed; however, this conclusion is confirmed by investigation of other groups of workers in the Netherlands (Ensberg, 1973) and elsewhere (e.g. Warnick *et al.*, 1972).

13. Psychological Tests

In recent years a fruitful cooperation has developed between toxicologists and psychologists: tests for measuring intellectual capabilities, psychomotor performance and personality are introduced to explore central nervous system responses to toxic exposures.

Several investigators have studied psychophysiological responses in subjects exposed to carbon monoxide. However, only a few authors reported individual data. Post-Lingen (1962) exposed 27 subjects to CO *(example 12)*: COHb before exposure <4.0%, after exposure 6.7—27.4%. She explored an effect on the nervous system with two methods: Critical Fusion Frequency (CFF) immediately after CO-exposure, and CFF after an i.v. evipan dose the following morning (evipan tolerance test, ETT). CO induced decrease of CFF, and increase of "area difference" in ETT.

Validity of both methods in regard to CO exposure could be calculated, taking 50 and 100 percentile levels in non-exposed situations (subjects served as own control):

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
CFF-50	0.56	0.50	1.06
-100	0.04	1.00	1.04
ETT-50	0.78	0.50	1.28
-100	0.50	1.00	1.50

The original data already suggested higher validity of EET in indicating differences in level of COHb if compared with the CFF-test. The calculation as given quantifies this difference.

Ettema *et al.* (1970) administered different amounts of alcohol to 40 subjects *(example 13);* blood alcohol levels ranged from 0.05 to 0.66 g/l; subjects performed several psychomotor tests, one of which was the Bourdon-Wiersma test: subject has to discriminate between several groups of 3—5 dots, 15 lines of 25 groups each; the observer records number of mistakes and performance time for each line; the average time per line and the variance of time per line is calculated. One expects a slowing down of speed (increased average time) and a larger instability of performance (increased variance) with increasing blood-alcohol levels. Both parameters were correlated with blood-alcohol level (tests on trend: $P = 0.03$). Predictability of blood-alcohol > 0.3 g/l was calculated for several cut off scores.

	Sensitivity (*se*)	Specificity (*sp*)	*se* + *sp*
Average time in seconds			
$\geqq 14$	0.81	0.39	1.20
$\geqq 18$	0.31	0.83	1.14
Variance in seconds			
$\geqq 1$	0.93	0.23	1.16
$\geqq 2$	0.67	0.41	1.08
$\geqq 4$	0.47	0.73	1.20

Tests on instability of performance appear to be more sensitive than tests on speed of performance; total validity, however, is still poor. Future investigations should try to develop more specific tests for measurement of stability of performance. This difference in sensitivity between both parameters was not revealed by the classical statistical treatment of data.

14. Method of Calculation

In this paragraph we give the methods of calculation of sensitivity, specificity, and other criteria of validity as mentioned in paragraph 2. We choose two of the examples discussed before, one with a high and one with a low validity.

Friborska (1969) *(example 9)* found activity of leucocyte alkaline phosphatase (L.A.P.) to be increased in workers continuously exposed to trichloroethylene. We derived the data from the figure as given by the author:

Exposed subjects	A: 4	L.A.P. > 95: 3	L.A.P. < 95: 1
	B: 18	16	2
	D: 3	3	—
	$n = 25$	22	3

Control subjects: $n = 20$, range L.A.P. 35—95, mean 58. $I+ = 22$ (L.A.P. > 95). $I- = 23$ (L.A.P. < 95). $T+ = 25$. $T- = 20$.

Original data

	$I+$	$I-$	
$T+$	$i+t+$ 22	$i-t+$ 3	25
$T-$	$i+t-$ 0	$i-t-$ 20	20
	22	23	45

Data converted into same number of subjects for $T+$ and $T-$

	$I+$	$I-$	
$T+$	88	12	100
$T-$	0	100	100
	88	112	200

$$\text{sensitivity} = \frac{i+t+}{T+} = \frac{22}{25} = 0.88$$

$$\text{specificity} = \frac{i-t-}{T-} = \frac{20}{20} = 1.00$$

$$\text{sensitivity} + \text{specificity} = 1.88$$

$$\text{predictive value} + = \frac{88}{88} = 1.00$$

$$\text{predictive value} - = \frac{12}{112} = 0.11$$

$$\text{risk ratio} = \frac{\text{predictive value} +}{\text{predictive value} -} = \frac{1.00}{0.11} = 9$$

The validity of the L.A.P.-test as indicator of exposure to trichloroethylene and of a typical response in a universe of "healthy" exposed + unexposed subjects is very high, because:

there are very few false negatives (12%);
there are no false positives;
all subjects with L.A.P. >95 belong to exposed group;
few subjects with L.A.P. <95 belong to exposed group;
subjects with L.A.P. >95 have a 9 times higher probability;
to belong to the exposed group than subjects with L.A.P. <95.

By means of a standard questionnaire Ensberg (1973) determined the number of subjective symptoms in workers exposed to a cocktail of pesticides *(example 4);* the frequency distribution of the score is given in paragraph 8. If we calculate the validity of score >3, we have the following data:

exposed subjects, $n = 85$, score >3: 49 score $\leqq 3$: 36
non exp. subjects, $n = 86$ 28 58

$I+ = 77$ (score >3)
$I- = 94$ (score <3)
$T+ = 85$
$T- = 86$; $T+$ is about the same as $T-$

	$I+$	$I-$	
$T+$	$i+t+$ 49	$i-t+$ 36	85
$T-$	$i+t-$ 28	$i-t-$ 58	86
	77	94	171

$$\text{sensitivity} = \frac{i+t+}{T+} = \frac{49}{85} = 0.58$$

$$\text{specificity} = \frac{i-t-}{T-} = \frac{58}{86} = 0.67$$

$$\text{sensitivity} + \text{specificity} = 1.25$$

$$\text{predictive value} + = \frac{49}{77} = 0.64$$

$$\text{predictive value} - = \frac{36}{94} = 0.38$$

$$\text{risk ratio} = \frac{\text{predictive value} +}{\text{predictive value} -} = \frac{0.64}{0.38} = 1.7$$

The validity of subjective score >3 as indicator of exposure and response in an universe of "healthy" exposed + unexposed subjects is rather poor because:

sensitivity + specificity is only 1.25;

there are 42% false negatives;

subjects with score >3 have only 1.7 times higher probability to belong to the exposed group than subjects with score $\leqq 3$.

The validity scores calculated are only pertinent to the method used, the exposure given, and the group studied.

15. Comparison with Statistical Approach

In many of the examples given test data were significantly correlated with exposure date. This already indicates that the test as such has a certain degree of validity in indicating the "true situation". What then is the advantage of calculating validity in addition to the classical statistical treatment of data?

The procedure appears to add important information relevant for evaluation of the "value" of the test method:

1. In toxicology nowadays we have to deal with usually low exposure intensities; only some individuals will respond, either due to a somewhat higher exposure than the group mean, or due to a decreased capability to cope with exposure. These individuals may hardly affect group averages and correlation coefficients; however, they receive attention when one uses percentile distribution scores, and calculates validity of extreme levels. This particularly becomes important if one wants to study the relevance of occupational data for evaluation of exposure of the general public to the same agent at low exposure levels: variability in biological capacity. Because in public health — even more so then in occupational health — environmental quality standards should take account of these *deviant groups* (aged subjects, pregnant women, children), much attention should be given to the deviant worker (Zielhuis, 1972).

2. If parametric (e.g. Student t-test) or non-parametric (e.g. rank sign test) statistical methods are used, the *information* given by *individual data* is lost to a great extent; these methods only regard the group as a whole. Calculation of sensitivity, etc. takes into account individual data, and particularly emphasizes those data exceeding given levels in controls. Parametric tests assume a normal distribution of data in exposed and non-exposed subjects; however, very often xenobiotic agents induce skewed distributions of blood and urine levels, of levels of enzyme activity, and so on. Non parametric tests often only regard the rank sign, disregarding the level as such. The procedure suggested can be used irrespective of normality of distribution.

3. Differences between groups may become highly significant, if the number of subjects increases; validity scores are to a great extent *independent of the number of subjects* examined. So, even tests showing a highly significant difference between groups of subjects, may have a poor validity, i.e. hardly predicting the true situation (exposure or response) in members of a universe of exposed + non-exposed subjects.

16. General Discussion

Validity scores of various biological tests have been presented. The advantages of introduction of this approach in epidemiological toxicology may be summarized as follows:

It takes into account *individual data*, paying due attention to deviant responses; loss of information is less than with many usual biostatistical methods.

It allows a better insight into the *predictive power* of tests than calculation of correlation coefficients and of statistical significance.

It allows insight into the *difference in discriminative power* of tests used for predicting the same true situation (degree of exposure or response).

It allows the *choice of adequate scores*, based upon percentile distribution rather than on average values.

It may give *indications for future research* directed at development of valid test methods.

It allows determination of predictive power of *combination of test* methods.

It allows determination of the *comparative validity* of *different indices* derived from *one test* and so ultimately may make a future investigation less time consuming.

It adds a necessary *criterium in the choice* of adequate tests for prediction of true situations, aside from other criteria (e.g. analytical sensitivity and specificity, input of manpower, nuisance for subjects examined).

The examples have been worked out in order to show the approach as such, and to elucidate the feasibility of it. The calculated validity scores should not be regarded as universally applicable; they are only valid for the investigations from which the original data were drawn, i.e. for a certain type of exposure (qualitative and quantitative), for a certain group of subjects (usually "normal" adults), for certain test methods used. When validity scores are going to be presented in literature, consensus may arise; the comparative validity of different methods can be judged more adequately. In order to achieve this, the investigators should either present individual data (or at least stratified distribution frequencies), or should calculate validity scores themselves.

References

Carpay, W. J. M.: Reticulocytes counts on radiation and non radiation workers. 2nd IRPA Congress, Brighton, U.K. (1970)

Carpay, W. J. M.: Biological indicators of radiation (in Dutch). J. belge Radiol. **55**, 467 (1972)

Czegledi-Janko, G., Avar, P.: Occupational exposure to lindane: clinical and laboratory findings. Brit. J. industr. Med. **27**, 283 (1970)

Ensberg, I. F. G.: Health and exposure to pesticides (in Dutch). Thesis, University of Amsterdam (1973)

Epstein, F. H.: Predicting coronary heart disease. J. Amer. med. Ass. **201**, 795 (1967)

Ettema, J. H., Burer, E.: Alcohol and mental capacity; investigation. Amsterdam, Coronel Laboratory (1970)

Friborska, A.: The phosphatases of peripheral white blood cells in workers exposed to trichloroethylene and perchloroethylene. Brit. J. industr. Med. **26**, 159 (1969)

Haeger-Aronsen, B.: An assessment of the laboratory tests used to monitor the exposure of lead workers. Brit. J. industr. Med. **28**, 52 (1971a)

Haeger-Aronsen, B., Abdulla, M., Fristedt, B. I.: Effect of lead on δ-aminolevulinic acid dehydratase activity in red blood cells. Arch. environm. Hlth **23**, 440 (1971b)

Kolmodin-Hedman, B., Azarnoff, D. L., Sjöqvist, F.: Effect of environmental factors on drug metabolism: decreased plasma half life of antipyrine in workers exposed to chlorinated hydrocarbon insecticides. Clin. Pharmacol. Ther. **10**, 638 (1969)

Kubota, J., Lazar, V. A., Losee, F.: Cu, Zn, Cd and Pb in human blood from 19 locations in the United States. Arch. environm. Hlth **16**, 788 (1968)

Liebscher, K., Smith, H.: Essential and non-essential trace elements. Arch. environm. Hlth **17**, 881 (1968)

Mac Mahon, B., Pugh, Th. E.: Epidemiology, principles and methods. Boston: Little Brown 1970

Post-Lingen, M. L. von: An experimental study of the effect of CO on CFF and ETT in healthy persons. Stockholm: Norstedt 1962

Rees, H. van, Wink, A.: Biological sampling on workers exposed to toxic agents (in Dutch). TNO-Nieuws **23**, 194 (1968)

Warnick, S. L., Carter, J. E.: Some findings in a study of workers occupationally exposed to pesticides. Arch. environm. Hlth **25**, 265 (1972)

Williams, M. K., King, E., Walford, J.: Method for estimating objectively the comparative merits of biological tests in lead exposure. Brit. med. J. **1968 I**, 618

Williams, M. K., King, E., Walford, J.: An investigation of lead absorption in an electric accumulator factory. Brit. J. industr. Med. **26**, 202 (1969)

Zielhuis, R. L.: Industrial lead intoxication (in Dutch). Thesis, University of Leiden (1959)

Zielhuis, R. L.: Industrial toxicology and public health. Pracov. Lék. **24**, 112 (1972)

Prof. Dr. R. L. Zielhuis
Coronel Laboratorium
voor Arbeidshygiene
Eerste Constantijn Huygensstraat 20
Amsterdam, The Netherlands