Validation of measurement procedures

R. Haeckel and I.Püntmann
Zentralkrankenhaus
Bremen

The new ISO standard 15189 which has already been accepted by most nations will soon become the basis for accreditation in many European countries. Subclause 5.5.2 of this standard claims that the medical laboratory should only use validated procedures: "The laboratory shall use only validated procedures to confirm that the examination procedures are suitable for the intended use. Validations shall be as extensive as are necessary to meet the needs in the given application or field of application." The term validation has already been defined in ISO 9000:2000 and has been differentiated from verification:

- verification: confirmation, through the provision of objective evidence, that *specified requirements* have been fulfilled

- validation: confirmation, through the provision of objective evidence, that the *requirement for a specific intended use or application* have been fulfilled

The definitions of both terms sound very similar. The only words which differ are set in italics. In other words: verification means testing characteristics which are typical for a instrument independent of a particular procedure, such as e.g. the quality of spectrometer filters; validation is testing the capability of an instrument for a particular procedure, e.g. for the measurement of the blood glucose concentration.

The term validation is not familar to clinical chemists. In former times we always used the term evaluation. Validation only mean confirmation of claims, whereas evaluation may also include setting claims by experimental work. However in practice, both terms are used synonymously.

In the new concept of the Directive 98/79 of the European Commission on *in vitro* diagnostic medical devices, industry will be responsible for the validation of commercially available test procedures. Validation must be performed by the laboratory only for in-house, or non-standardized, measurement procedures.

A decade ago, the European Committe for Clinical Laboratory Standards (ECCLS) has developed a multicentre evaluation concept which is still valid. It consists of three steps:

1) *Evaluation* of prototypes by the manufacturer, perhaps in cooperation with user(s).

2) *Multicentre evaluation*: confirmation of the manufacturer´s claims (product ready for introduction on the market).

3) *Validation*: short evaluation by each customer after purchase of the product.

In step 3, some national accreditation bodies claim end-user protocols. The ECCLS document for procedure comparison evaluation is complemented by a new standard from the European Committee of Standardization (*Performance evaluation of* in vitro *diagnostic medical devices*) which is currently prepared.

In the a note in the subclause 5.4.5.3 of the new standard ISO 17025, which is a sister document to ISO 15189, designed for all laboratories, performance criteria for validation are listed: "The techniques used for the determination of the characteristics of a method should be one of, or a combination of, the following: calibration using reference standards or reference materials, *comparison of results with other methods*, interlaboratory comparisons, systematic assessment of the factors influencing the result, *assessment of the uncertainty* of the results based on scientific understanding of the theoretical principle of the method and practical experience."

All characteristics and the tools how to study them, are well known. However, two characteristics which have changed somehow, will be discussed in more detail: uncertainty and method comparison.

Three types of uncertainty have been defined:

- standard uncertainty (*u*): imprecision (standard deviation)
- combined uncertainty (*uc*): $(u_1{}^2 + u_2^2 + u_3^2)^{0,5}$
- expanded uncertainty (*U*): ku_c (if coverage factor *k*=2, level of confidence ≈ 95%)

The standard uncertainty, that means the precision determined in the laboratory, is on one side a very useful operational quantity, but on the other side an artificial quantity which does not satisfy the clinician´s need. The clinician requires an estimation of the total variability called uncertainty budget which also includes the preanalytical phase and other components.

Because of the complexity of modelling the measurement procedure, influences of various input quantities have to be considered as possible sources of uncertainty:

- incomplete definition of measurand,

- sample heterogeneity

- inexact value of calibrators (insufficient traceability)

- matrix differences between calibrators and samples

- stability of the sample, the analyte or reagents used

- presence of interfering compounds in the sample (lack of specificity)

- imprecision of statistical algorithms used on results on calibrators

- random variability inherent in the measurement process.

The information of a large expanded uncertainty may not be very valuable in the treatment of an individual patient. Therefore, it will be necessary to rely on standard uncertainty (imprecision) and in-house reference ranges until the uncertainty is reduced to a reasonable level (Kristensen and Christensen, 1998). The major disadvantage of the new concept is that it is based on many (may be too many) assumptions. On the other side, it opens a new way of thinking.

Another essential part of any measurement procedure validation is the comparison with another measurement procedure which is usually applied by the laboratory (see also note 2 of subclause 5.4.5.3 in ISO 15189).

Measurement procedure comparison studies are widely used in laboratory medicine to assess agreement between two measurement procedures which measure the same analyte. One measurement procedure is usually considered as the reference, the other one as the test measurement procedure, both of them measure with a certain degree of uncertainty. Usually they disagree to some extent. The question is: can the disagreement be tolerated.

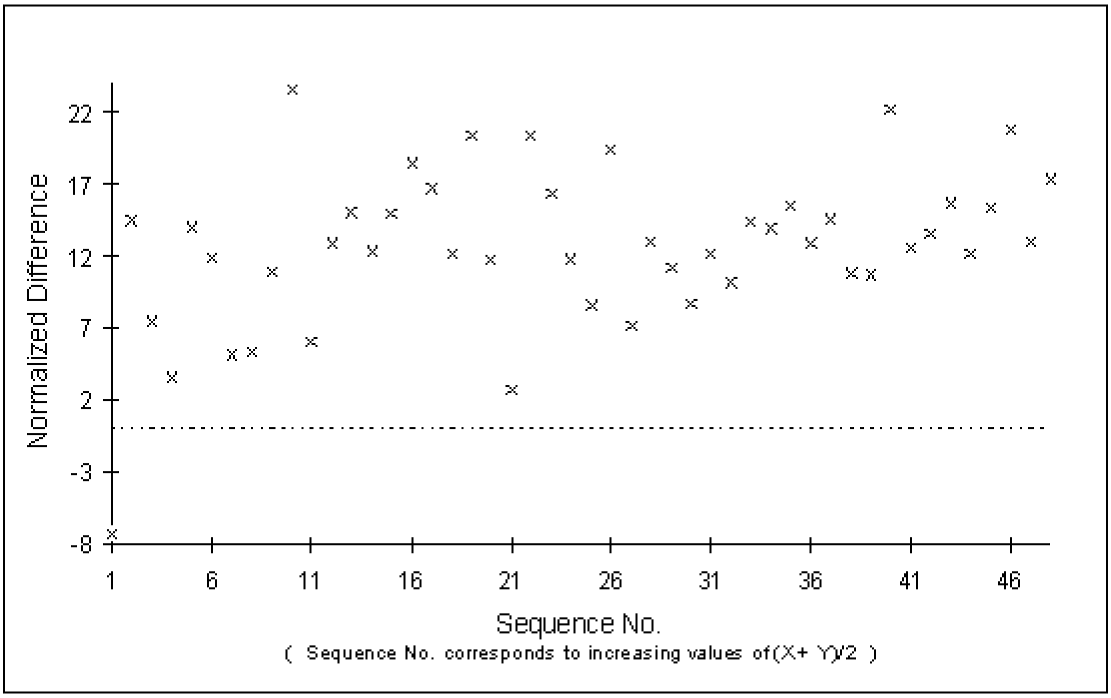
The results from comparison studies must be evaluated by two steps:

- Step 1: definition of the analytical (dis)agreement (conventional concept)
- Step 2: estimation of diagnostic relevance (future task)

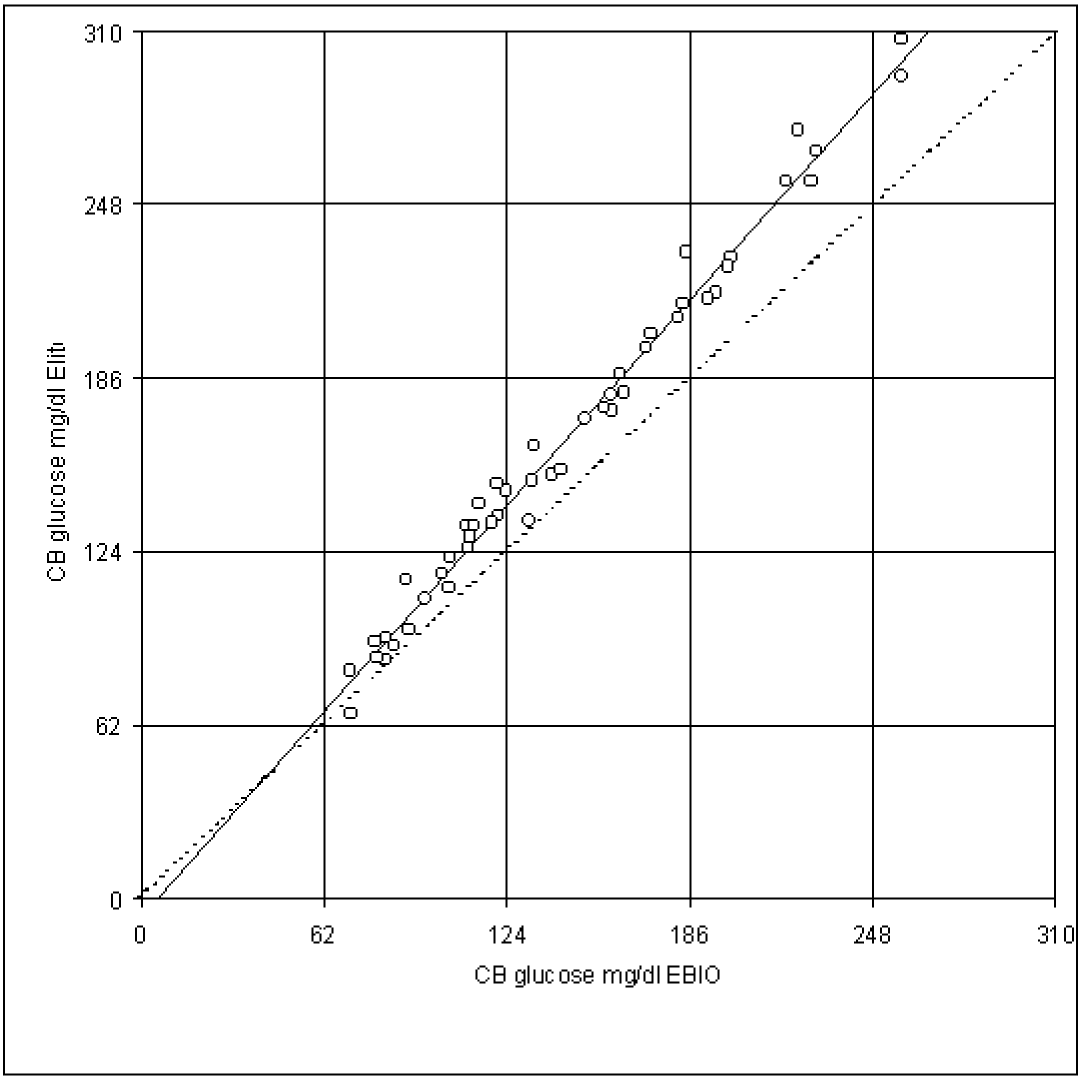
The first step is the definition of disagreement, the second step is to consider the acceptability of the detected disagreement concerning the intended purpose.

The step 1 is probably familiar to most clinical chemists. It usually starts with the graphical presentation of the paired data obtained from several independent subjects. Two types are still en vogue:

- the difference plot or one of its modifications such as the normalized differences (Fig.1):



- the x/y plot (Fig.2):



Both figures represent the same data set (glucose concentration in capillary blood dertermined with an Elite and an Ebio analyzer). Both plots have their advantages. However, they are both more or less equally useful. Several characteristics are usually examined from these plots that can be judged by individual statistical tests:

- linearity over the entire range: visually or by Cusum test;
- slope;
- intercept calculation of the fitting line (regression analysis): e.g. by (classical linear regression with or without of weighing, standardized principle component analysis, Deming, Passing-Bablok;
- spread around the fitting line: e.g. md(95) or $s_{y.x}$;
- outliers;
- maximum/minimum interval;

- (correlation coefficient).

It is still a debate which mathematical procedure should be applied for calculating the fitting line. The Deming and the Passing-Bablok procedures have the advantage that they do not require a normal distribution on both axes. Whereas Americans still prefer the classical regression with some kind of weighing the data, the majority of Europeans appears to use the Passing-Bablok method. The correlation coefficient does not provide any further information and can even be misleading. However, clinicians still like this characteristic because it is a single value. They do not accept if we provide them with a set of several analytical characteristics. The spread around the fitting line can be characterised by several statistical parameters:

Example of precision indices in measurement procedure comparisons (Hallé et al.1986): Elite versus EBIO in capillary blood (mg/dl)

- $y = 1{,}167x – 8{,}167$; $r = 0{,}981$; $n = 143$
- Standard deviation of residuals: $s_{y.x} = 12{,}0$
- Dispersion of residuals:

  - Passing-Bablok regression: *md(95)* = 15,4

  - Standardized principle component: *SE* = 8.00
- Mean percent differences:*(y-x)% of x* = 11,5
- Precision index: *1- 1/100(mean % diff.)* = 0,89

Despite the different statistical tests available, the fastest and perhaps most reliable testing tool is the eye of an experienced clinical chemist.

The statistical characteristics mentioned describe the performance of a measurement procedure in the laboratory, it is an analytical validation, they are sufficient to judge the comparability for the laboratory. The customer of the laboratory requires one characteristic. He wants to know, how many errors does he has to expect when one measurement procedure is replaced by another one.

In the example just presented (Fig.2), a low variation occured around the regression line. The md(95) value was 15,4, and the correlation coefficient r = 0,981.However, a significant slope of about 1,167 was observed. Are the discrepancies to be expected due to the total error of clinical relevance?

In this example the comparison measurement procedure is a routine measurement procedure for the glucose concentration measurement of a central laboratory, and the test measurement procedure is a typical point of care testing device for blood glucose measurements near the patient.

Several consensus documents on the allowable error have been published; in the case of performance goals for glucose:

Error of measurement *CV* (%)

Consensus groups:

ADA (1994)[1)]	< 5 %	
NCCLS (1994)	< 20 % [>5,5 mmol/L]	
	< 0,8 mmol/L [≤5,5 mmol/L]	
CMT (1991)[2)]	< ±15 % [>6,7mmol/L]	< 5,0
	< ±18 % [<6,7mmol/L]	
ISO-TC 212(1998)	< +1,1 mmol/L [<5,5 mmol/L]	
	< ± 20 % [>5,5 mmol/L]	
RiliBÄK[3)](1993)	< ± 15 % (of assigned value)	≤ 5,0
CLIA (1992)	< ± 0,3 mmol/L or + 10 %	

Individual groups:

Fraser *et al*. (1993)	≤ 1,9 %	≤ 2,2 %
Cok et al. (1985), Koschinsky *et al*. (1988)[4)], Weiss *et al*. (1994)[5)]	≤ 20 %	
Price *et al*. (1988)	≤ 10 %	

1) American Diabetes Association
2) Center for Medical Technology (Leiden)
3) Guidelines of the German Association of Physicians
4) ³ 95 % of all test values, below regression line: £ 17 %
5) 3,3 mmol/L, 15 % ³ 4,7 mmol/L

The allowable error varies from 2 to 20 %. There is no consensus among the various consensus proposals. All proposals more or less are based on percentage values. They belong to class 2 in the hierarchy of objectivity suggested by Fraser and Petersen (Clin Chem 1999;45:321):

1. Assessment of the effect of analytical performance on clinical decision making.

1.1 Quality specifications in specific clinical situations.

1.2 General quality specifications based on medical needs: e.g. based on biological variation.

2. Professional recommendations from expert or institutional groups.

3. Quality specifications laid down by regulation or by external quality assessment schemes organizers.

4. Pubslished data on the state of the art from proficiency testing schemes or published methodology.

Class 1 has the highest hierarchical level, objectivity requires assessment of the effect of analytical performance on clinical decision making. Boyd and Bruns (*Clin Chem* 2001;47:209-14) very recently have related performance characteristics of glucose analyzers to error rates in insulin dosage. We have developed a new procedure relating performance characteristics to error rates in diagnostic decision making.

This new test answers the question: how many discordant classifications have to be expected at a particular decision limit which discriminates between non-diseased and diseased subjects if a laboratory switches from one measurement procedure to another one.

If two procedures of measuring the glucose concentration are compared, a clinically relevant decision is the diagnosis of diabetes mellitus in the fasting state (Fig.3). For capillary blood, WHO has recently recommended a decision limit of 6,1 mmol/L to discriminate between non-diabetic and diabetic glucose concentrations.

If a quantity is measured with two measurement procedures, both results may be below the decision limit c_d (x_1/y_1 in Fig.3) or above c_d (concordant classification). In a few cases, the results of one procedure (x_2) may be below c_d and of the other measurement procedure (y_2) above c_d (discordant classifications: x_2 = non-disease, y_2 = disease).

The number of discordant classifications depends on 2 probabilities (Fig.4):

(i) the probability that xp occurs in the population (population probability): $P_p(x_p)$

(ii) the probability that the test value of y_p corresponding to x_p lies above c_d (analytical probability) for a given x_p: $P(y_p \geq c_d/x_p)$.

If x lies above c_d and y below c_d , the probability of a false negative decision becomes 1- $P(y_p \geq c_d/x_p)$. The population probability P_p is determined on the basis of a Gaussian distribution. From this distribution, the probability of occurrence of each possible glucose value within the study population is calculated by numerical integration. The population probability is calculated using a statistical program.

The discordance rate *P* is obtained by multiplying and summing up both probabilities. The probabilities can be calculated using a statistical program.

Results of the proposed test are shown in Fig.5 for the comparison of an Elite XL with an EBIO glucometer. The corrsponding data have already been shown on the difference and the *x/y* plot. The horizontal axis represent mg/dl fasting plasma glucose concentrations, the vertical axis means the probability $P_p(x)$, that x (comparison measurement procedure) occurs in the population studied. The blue curve shows the distribution of the population values measured by the comparison measurement procedure (EBIO). The height of the curve shows the probability (in percent) of occurence of each individual glucose value, rounded to integers. The area below the curve sums to 100 %. The decision limit is located at 6,1 mmol/L (≡110 mg/dl) on the abscisse.

The red hull curve left of the decision limit (6,1 mmol/L (≡110 mg/dl)) describe the probability that a value from that position is erroneously classified by the test measurement procedure as lying right of the decision limit (false positive classification). The hull curve is the global probability of a false positive measurement by the test measurement procedure. It is identical with the area under the hull curve in percent of the area under the population curve.

Similarily, the green curve right of the decision limit indicate the probabilities for false negative discordances and the remaining area quantifies the probabilities of correct decisions.

Conclusions:

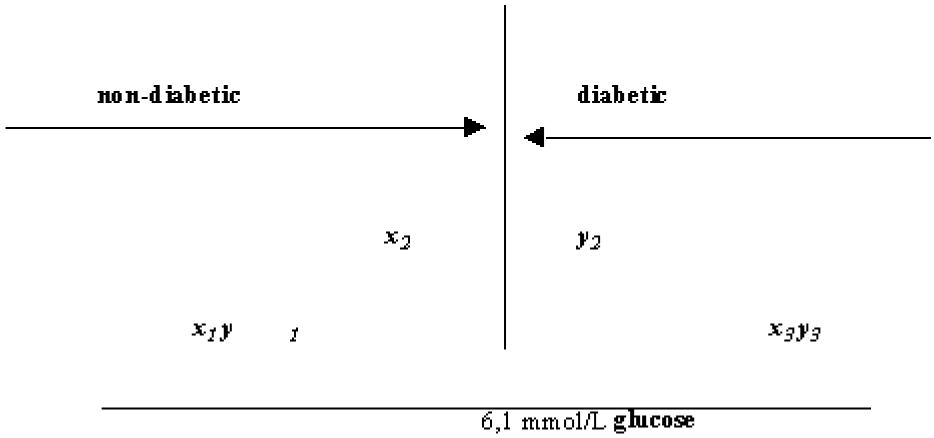
1. The responsibility for the validation of laboratory measurement procedures belongs to the institution which has developed a procedure, in most cases the manufacturer. 2. New procedures should be validated by several laboratories, in the future probably by accredited laboratories, preferably by external laboratories (that means the ECCLS concept is still valid).

3. The end-user must be aware of the validation results. If these validation data are not available or incomplete or insufficient, the laboratory itself is responsible for the validation before the new procedure is applied routineously.

4. The end-user should perform a short validation including at least one type of uncertainty, agreement with assigned values of control materials and a measurement procedure comparison study if applicable.

5. Measurement procedure comparison studies require two steps: the definition of analytical (dis)agreement, and an estimation of diagnostic relevance, e.g. of the rate of discordant classifications.

Fig.3

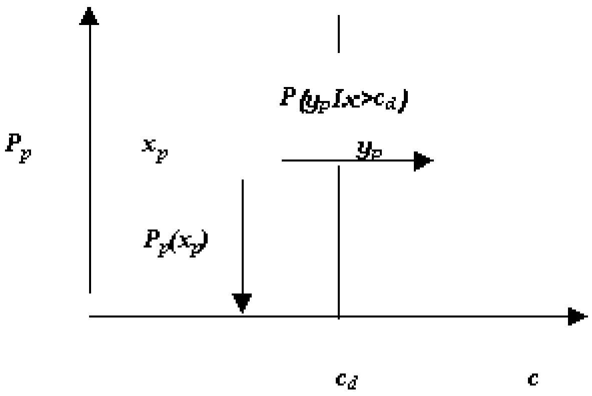


x_i = comparison measurement procedure

y_i = measurement procedure to be tested

Fig.4

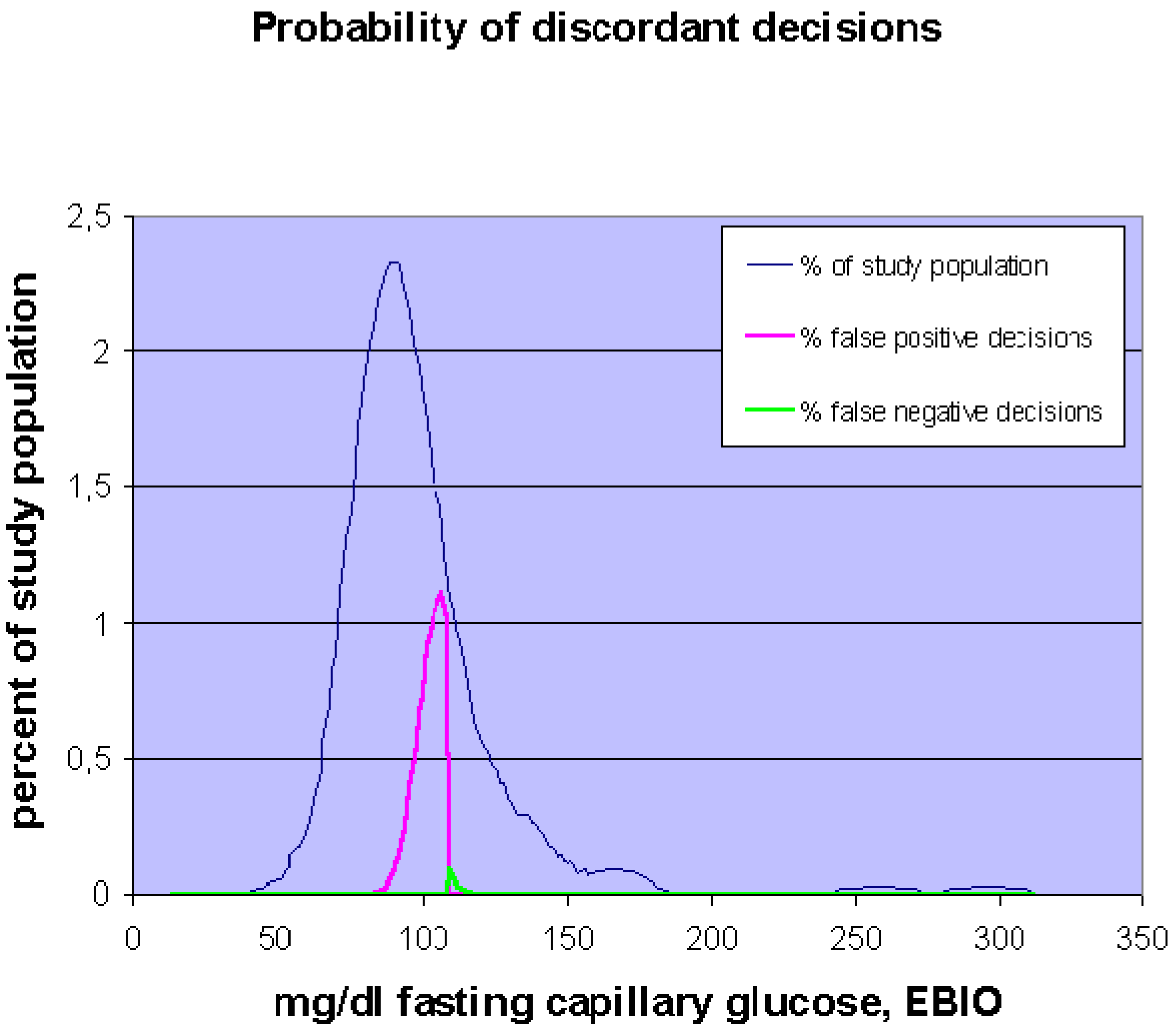
Principle of the new test for discordant classifications:



$$P(x_p < c, y_p \geq c) = \ ?\ P_p(x_p)\ .\ P(y_p I x \geq c)$$

$$P(x_p \geq c, y_p < c) = \ ?\ P_p(x_p)\ .\ 1\text{-}P(y_p I x \geq c)$$

Fig.5



<u>Citació recomanada per a aquest document:</u>

Haeckel R. Validation of measurement procedures. In vitro veritas 2001;2, art 20:<http://www.acclc.cat>