SYSMEX EDUCATIONAL ENHANCEMENT AND DEVELOPMENT | APRIL 2015

SEED HAEMATOLOGY

Medical statistics – your support when interpreting results

The importance of statistical investigations

Modern medicine is often based on statistical investigations concerning the frequency of an illness, the probability of success of a therapy method, investigations as to which laboratory values are 'normal', and many others. The investigation outcomes have implications for the medical statements made in that context. Also the informative value of a laboratory parameter or test result is derived from the statistical conditions that characterize that test or which were used to define particular cut-off limits that will be referred to for decision making. Therefore, to know such statistical conditions is of importance in the interpretation of laboratory results. What does it mean for the patient, for example, when a particular value is measured?

Sampling specimens and the inherent fluctuation range

A diagnostic measurement is taken with the intention of acquiring knowledge of the conditions prevailing in the patient (concentrations of substances or their change, reaction times, local position of parts of the anatomy, morphology of tissue, etc.). So, in many cases samples are taken in order to get information about the patient, for example blood or urine specimens. A portion of each of these samples is in turn set aside and used for the respective test.

When one considers that a defined component (for example the white blood cell (WBC) concentration) is not identical everywhere, but is instead distributed randomly throughout the body, this demonstrates that a measurement result of this type can only ever represent an approximation of the actual patient status. If, for instance, an average of 10 WBC/µL is present in a patient's urine, it is entirely possible that a concentration of 12 WBC/µL may be found locally, for example, at the point of sampling. This may be balanced elsewhere by a correspondingly lower concentration. Therefore, statistically speaking, this results in a potential range of fluctuation due to the sampling. Its impact is determined by the concentration itself, because in the case of a low concentration of the analysed component, even a minor absolute fluctuation can result in a drastic change in percentage terms. In the example mentioned above (10 vs. 12 WBC/µL), the statistical fluctuation would result in an error of 20%!

In contrast, when measuring white blood cells in a normal blood sample, even a ten times higher absolute fluctuation would make no significant difference, because the actual cell concentration is much higher (around 4–10 x 10^3/µL).

Coefficient of variation

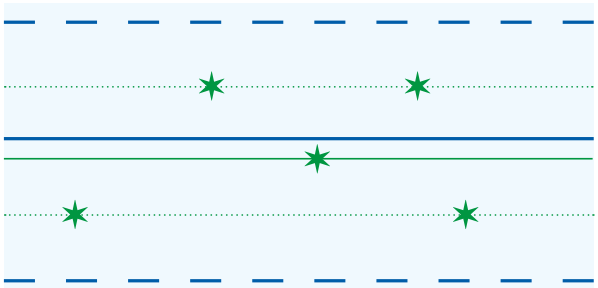
The above mentioned fluctuations are purely statistical and are independent of the method used. They are caused by the removal of a sample from a whole (e.g. from all of the blood in a patient or all of the blood in a tube) and the random distribution of the measured attribute throughout the whole. These fluctuations produce a statistical **coefficient of variation (CV)**, which indicates the fluctuation range of the measurement values around the 'true' value and is purely dependent on the concentration of the analyte.

This is reflected, for example, in the Rümke table (see Tab. 1), which shows how many cells need to be measured in order to obtain a certain reliability of the measurement value. If measurements in the laboratory indicate an alleged lower coefficient of variation it must be noted that this CV relates to the mean of those measurements, and not to the 'true' value (see Fig. 1).

In addition to this statistical coefficient of variation there are other sources contributing to the total range of fluctuation, including fluctuating pipetting volumes, subjective interpretation that is non-conforming to a standard (possibly due to changing personnel), etc. The measurement of a parameter therefore always takes place in the context of a conflict between precision and practicability.

The question of clinical relevance will ultimately be a decisive factor – how precisely must I define a parameter to be able to responsibly make a clinical decision?

Automation can therefore be of help here in two ways: At high concentrations (e.g. RBC in whole blood, around 5 x 10^6/µL), the statistical CV is kept small by counting a particularly large number of cells. At the same time, other sources of error are minimised because the sampling and its evaluation always take place in the same fashion.



Green: mean measurement value with fluctuation range
✶: individual measurement results
Blue: true concentration in the patient with statistical fluctuation range

Fig. 1 *True value (blue bold line) with statistical coefficient of variation/fluctuation range (blue dashed lines). Experimental mean (green line) with fluctuation range of the measurement values (green dotted lines). The experimental CV here appears to be lower than what is statistically possible. However, the measurement values are merely coincidentally found in a more narrowly defined area around the mean. At least in one direction they are spread over almost the entire statistical fluctuation range. Additional measurements would confirm the statistical CV and move the mean still closer to the true value.*

Coefficient of variation (CV)

The coefficient of variation indicates the uncertainty of a measurement value at a certain concentration in relation to this concentration. It is a measure for the imprecision of a method and is calculated by dividing the standard deviation by the mean. The result is reported as a percentage. In contrast to the standard deviation, which provides an absolute value for the range of fluctuation, the CV makes it possible to see how drastic the impact of this uncertainty is at the given concentration.

Example: A standard deviation of 2 at a concentration of 2 cells/µL means a large coefficient of variation of 100%. However, the same standard deviation at 200 cells/µL is significantly less problematic as the corresponding CV is only 1%.

Confidence intervals

Once a result has been obtained, the question arises as to which statement and decision it allows. In order to decide whether a certain value is pathological, a **reference interval** is referred to that reflects the distribution of the measurement values in a healthy population. In order to determine this interval, measurement values are collected from healthy test subjects (in some cases also from individuals with illnesses that do not influence the measurement value). Of these values, the **95% confidence interval**, meaning the central range in which 95% of the measurement values can be found, is usually used as a reference interval.

One should bear in mind that a value outside of the reference range does not necessarily mean that the test subject is ill. Rather, 5% of the population show values that are above or below this interval. The question is: does the test subject belong to this group? Conversely, a measurement value within the reference area does not necessarily mean that the patient is healthy.

a	n = 100	n = 500	n = 1000
1	0–4	0–1	0–1
2	0–6	0–3	0–2
3	0–9	1–5	2–5
4	1–10	2–7	2–6
5	1–12	3–8	3–7
6	2–13	4–9	4–8
7	2–14	4–10	5–9
8	3–16	5–11	6–10
9	4–17	6–12	7–11
10	4–18	7–13	8–13

Tab. 1 *Excerpt of a table published by Rümke concerning the 95% confidence intervals of a percentage rate of cells with a given total number of counted cells. a = found percentage rate of a population, e.g. in the differentiation; n = counted cells; X – Y = 95% confidence interval of the result. This means that the result will be found within this range in 95% of cases, below it in 2.5% of cases and above it in 2.5% of cases. As an example, for the manual differentiation of 100 WBC with the eosinophil population found at a rate of 3%, this means that the true value in 95% of measurements can lie between 0 and 9. Even with a differentiation of 1,000 leucocytes, the possibility cannot be excluded that 5% rather than 3% eosinophils are truly present. The table can be correspondingly continued for higher percentage rates (e.g. relevant for neutrophils), but also for larger total numbers of counted cells (like with automatic analysers).*

Comparison between tests and to reference methods

The actual diagnostic question is often reduced to contrasts such as 'ill' or 'not ill' or 'therapy is effective' vs. 'therapy is ineffectual'. The decision is thereby made on the basis of certain limit values, so-called 'cut-offs', or decision limits. In order to determine the quality of a test, it is compared with established tests, for example, the respective 'gold standard', meaning the ostensibly best available method, or so-called 'reference methods'; methods that have been explicitly defined as standards for the specific parameters.

	Old test positive	Old test negative
New test positive	A	B
New test negative	C	D

Tab. 2 *Example of a fourfold table for a general comparison of two methods.*

The various tests can then be compared on the basis of a so-called 'fourfold table'. Within the scope of a pure comparison of methods, the so-called 'concordance', meaning the agreement of two test results, can be determined. In the process it should be considered that the result of the old method is not necessarily correct.

The situation is different when a comparison is made with a **reference method**. Its result is then correct by definition.

Reference method

It is an analytical method that is recognized, for example, by the relevant professional associations as the most reliable method, meaning that the value determined with it is regarded as 'true'. However, the term is often used incorrectly to express the most widely used method. This can cause problems, because even widely used methods can supply false results and can therefore not simply be regarded as delivering 'true' results. This has consequences for both the conclusions of the method comparison and for the evaluation methods to be used. In such cases it would be better to use a phrase like 'selected method of comparison'.

Sensitivity and specificity

The use of a fourfold table helps calculate how often a positive test is correctly positive as well as how often a negative test is correctly negative. Briefly speaking, how often the test shows the correct results. Tab. 3 shows an example of a test evaluation in this way.

In this case, the non-conforming samples are either false positive (quadrant B) or false negative (quadrant C). When comparing these data with a reference method, the following characteristics can now be determined:

- How often is a positive sample correctly recognized as such? **(sensitivity)**
- How often is a negative sample correctly recognized as such? **(specificity)**

These two characteristics show a conflicting relation: the more sensitive a test is, the higher the probability of falsely defining a negative sample as positive becomes. Conversely, the same goes for a rather specific test showing a potential tendency of falsely defining a positive sample as negative. Of significance here is the application area of a test.

Sensitivity
Sensitivity indicates how often a positive sample is positively registered by a test. With low sensitivity, the test overlooks many positive samples.

Specificity
Specificity indicates how often a negative sample is also negatively registered by a test. Low specificity leads to many false alarms.

One can optimise the key figures of a test by, for example, adjusting the cut-off value. Depending on the specific question, high sensitivity or high specificity can be the desirable criterion. It is not always necessary to find an optimal equilibrium between the two characteristics. However, sensitivity and specificity have little to say about a concrete case. They rather tend to describe the societal viewpoint or that of the provider. They provide general indications of the reliability of a test.

Predictive values

The so-called **'predictive values'** can also be derived from these data. Where sensitivity and specificity pose the question of how often the condition of the patient is correctly registered, the **positive and negative predictive values** indicate how often positive or negative medical findings are actually correct.

A high positive predictive value provides security that a positive sample also accompanies an illness. However, a non-positive medical finding does not necessarily mean 'no illness'. This means that illness can still be overlooked. Conversely, a high negative predictive value makes it possible to classify negative samples as unsuspicious with certainty.

	Illness is present	Illness is not present	
New test positive	A (e.g. 67)	B (e.g. 3) (false positive result)	How often has a positive test correctly indicated the illness? (70 tests positive, 67 of these correct)
New test negative	C (e.g. 12) (false negative result)	D (e.g. 128)	How often has a negative test correctly excluded the illness? (140 tests negative, 128 of these correct)
	How often was a positive sample also recognised as such? (79 samples positive, 67 of these correctly recognised)	How often was a negative sample also recognised as such? (131 samples negative, 128 of these correctly recognised)	

Tab. 3 *Example of a fourfold table for estimating the quality of a diagnostic test. The example data results in a sensitivity of 85% (67 of 79 samples), a specificity of 98% (128 of 131 samples), a positive predictive value of 96% (67 of 70 positive tests were correct) and a negative predictive value of 91% (128 of 140 tests were correct).*

Positive predictive value

The positive predictive value indicates the probability of, for example, the presence of a certain illness when a test for it has produced a positive result.

Negative predictive value

The negative predictive value indicates the probability that in the event of a negative test result, the tested person does indeed not suffer from the illness being tested for.

Of importance here is to keep in mind the frequency (prevalence) with which an illness appears. If an illness is very rare, there will actually be a large proportion of negative samples among the samples from the general population. It is very difficult to obtain a good positive predictive value in such cases. The predictive values effectively describe the viewpoint of the treating physician or patient. They want to know what a positive or negative finding means: How probable is it that I suffer from the illness despite a negative test result? How concerned should I be about a positive test result? Some screening tests accept a high false-positive rate because they can be carried out quickly and affordably and the result can then be confirmed with an additional test carried out with a higher degree of prevalence on a reduced number of test subjects.

However, when taking the prevalence into account, the sample pool involved must also be considered. The frequency of illness among the general population can rarely be applied in a hospital laboratory: most patients are in hospital for a good reason; the probability that they do suffer from an illness is therefore significantly higher than among the general population.

As a measure of frequency, the prevalence also indicates a probability that a certain illness is present even in the absence of any other information about the patient. A good diagnostic test changes this somewhat diffuse probability with as much clarity as possible: A positive test should show a clearly higher certainty that the illness is present. If this is not the case, the test does not really contribute to the gaining of new insights.

Representation of characteristics' results

The basic characteristics described in the previous sections indicate the general diagnostic quality of a test. It is therefore of interest to visualise these in a comprehensible manner. The results obtained from a test evaluation can be represented in the most varied ways. Instead of the predictive values, so-called 'likelihood ratios' or 'odds ratios' are given, meaning ratios of probability. Sensitivity and specificity can be represented in a so-called **'receiver operating characteristic (ROC)'** curve dependent on the defined cut-off values. The area below the curve is also indicated where appropriate: The closer this is to 1, the better the chances are that high degrees of sensitivity and specificity can be achieved simultaneously.

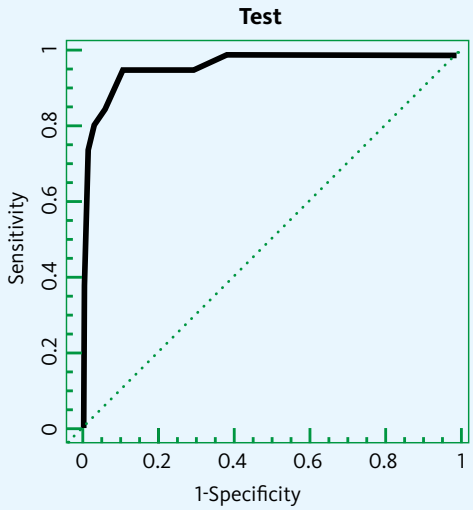


Fig. 2a *Example of a ROC curve for a test that offers a high degree of sensitivity and specificity and successfully reconciles these with one another. The area under the curve is in this case very close to the area of the diagram as a whole, meaning 1.*

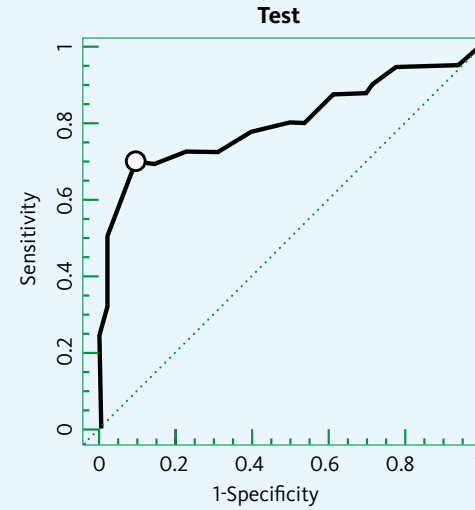


Fig. 2b *Example of a test that can only achieve a high degree of sensitivity at the expense of a significant decrease in specificity. The point offering the best combination is marked. The area under the curve is considerably smaller than in example 2a.*

ROC curve

'Receiver operating characteristic' curve: Plotting sensitivity versus '1-specificity' makes it possible to read the extent to which sensitivity and specificity can be optimised. The diagonal (angle bisector) of the axes indicates a 50 % chance of the presence of illness. Tests located close to this line are not significantly more valuable than the toss of a coin in diagnostic terms. The area under the curve (AUC) of the diagonal is 0.5. Conversely, tests with curves that project into the upper left corner and have an AUC close to 1 allow simultaneously for high degrees of sensitivity and specificity. However, for many questions (e. g. when screening) it is not necessary to find the optimum of the entire curve, as, for example, a high degree of sensitivity is sufficient.

However, these forms of representation are ultimately based on the same data and may be transferred into one another. A deeper discussion of this subject will not be part of this article, though. In face of the diversity of statistical aspects, which play a role in both the interpretation of lab values per se and in the evaluation and comparison of various tests, future contributions to this subject may address these points where appropriate.

References

[1] ***Altman DG.*** *(1991): Practical Statistics for Medical Research. London: Chapman & Hall.*

[2] ***Kirkwood BR.*** *(1988): Essentials of medical Statistics. London: Blackwell Science Ltd.*

Sysmex Europe GmbH
Bornbarch 1, 22848 Norderstedt, Germany, Phone +49 40 52726-0 · Fax +49 40 52726-100 · info@sysmex-europe.com · **www.sysmex-europe.com**