Contents

1. Introduction

1.1 Diagnostic theory

A fundamental difficulty within healthcare is that two patients never look the same. Furthermore, even if only one single patient is being studied, the medical information tends to vary over time. As a matter of fact, information normally varies between several measurements on the same individual even if these are taken over a short period of time. This is explained by the variability between patients, within patients and measurement error, respectively. Consequently, interpretation of uncertain medical information is a reality which health care professionals have to deal with.

The obstacle caused by this dilemma is memorably expressed by Bertrand Russel as, ”The main task of modern philosophy is to teach man to live without certainty and yet not be paralyzed by hesitation” [1]. Another telling description is, “Medical care is often said to be the art of making decisions without adequate information.” [2].

A correct diagnosis is crucial for the subsequent care of a patient. Since diagnostic tests play an important role in the diagnostic work-up and since test results often are afflicted with uncertainty, interpretation of diagnostic test results is one situation where the clinician has to deal with the dilemma described above.

In the declaration of Helsinki, which presents ethical principles for medical research it is pointed out that most prophylactic, diagnostic and therapeutic procedures involve risk and burdens [3]. Moreover it is highlighted that, ”The primary purpose of medical research involving human subjects is to improve prophylactic, diagnostic and therapeutic procedures and the understanding of the aetiology and pathogenesis of disease.”

Thus it is rather obvious that it is important to develop diagnostic theory in order to make the use of diagnostic information as certain as possible. An erroneous diagnosis could lead to dramatic consequences in terms of unjustified treatment, lack of treatment or even mistreatment.

It is difficult to give an unambiguous definition of ”diagnostic theory”, but one possible suggestion of a general definition is “theory aiming at improving diagnostic procedures”. This general definition allows a wide range of issues. Without claiming to be an exhaustive overview, the list below contains some examples of important issues and examples of topics and questions related to diagnostic theory. For further reading, the comprehensive book “The evidence Base of Clinical Diagnosis” [4] is recommendable, and the references given for each topic in the list below is recommended as an introduction.

- Analytical procedures and quality control [5].
    - Standardization of collection and handling of test material
    - Standardization and calibration of measurement procedures and technical equipment

- Evaluation of a diagnostic tests discriminatory ability [6-8]
    - Optimal choice of diagnostic threshold
    - Test accuracy, e.g. sensitivity, specificity.

- The contribution of a single test in the diagnostic process [9]
    - Does the test contribute with information beyond already available information?

- Study design [10-11]
    - Choice of study population
    - Choice of spectrum - clinical setting
    - Are the results transferable?
    - Choice of the reference test

- Evaluation of cost/benefit [12]
    - Do the clinical benefits outweigh the costs and burdens?

- Use of informatics and computer support [13]
    - Computer based decision systems
    - The use of digital journals

- Meta analysis and review [14]
    - Which studies are possible to pool?
    - How to perform analysis of pooled data

- Decision theory [2]
    - How to reason as a clinician
    - Understanding of fundamental concepts among clinicians
    - The implementation of new diagnostic procedures, possibly computer assisted

Bearing in mind the wide spectra of topics it is natural that the theoretical area is of interest for a range of different specialists, e.g. physicians, epidemiologists, engineers, statisticians, psychologists to mention some of the professionals concerned by the topics. A theoretical arena for this cross-sectional discipline is “Medical Informatics”.

According to Van Bemmel and Musen the work being done by researchers within the field of Medical Informatics could be described as, ”We develop and assess methods and systems for the acquisition, processing, and interpretation of patient data with the help of knowledge that is obtained in scientific research” [13].

In summary, the diagnostic work-up is a crucial procedure, and it is essential to develop theory regarding how to make as certain diagnosis as possible, even if available information is uncertain. This includes a wide range of questions and a wide range of specialists is needed.

1.2 The development of diagnostic theory

The theoretical development of diagnostic research lags behind that of therapeutic research, a fact that was noted by Cochrane already in 1972 [15]. Consequently, a large number of studies are reported with a relatively low scientifical quality [16-21]. Encouragingly, there has been a theoretical development during the last decade, originating in a number of articles dealing with a spectra of methodological aspects [10-11,22-26], some of which are synthesized into a book [4].

Furthermore, a standard for how to report diagnostic studies has been successively developed [27-29]. This standard, abbreviated "STARD" is an analogue to the standard available for how to report parallel group randomized therapeutic trials, i.e. the so called CONSORT-statement [30-31].

The suggested standard STARD focuses on design issues and univariate evaluation of tests. Consequently multivariate approaches are to a great extent neglected. According to a suggestion by Moons et al. univariate evaluation of a test should be defined as "test research" while "diagnostic research" should be characterized by multivariate evaluation [32]. The purpose of such a multivariate approach is to evaluate the additional information of a test beyond other pre-existing information [ibid]. This is due to the multivariate nature of a diagnostic work-up which most often inlcudes interpretation of joint information. Pre-existing information could for instance be medical history, physical examination, gender, preceding test, etc.

Multivariate analyses also play an important role in diagnostic theory due to the fact that classical measures of accuracy, e.g. sensitivity, specificity and predicting values, actually could vary across subpopulations [33-36]. If the specificity and sensitivity depend on some factors, e.g. gender, race and smoking habits, it may be important to adjust for these factors when a test result is interpreted. In other words, a test result may be interpreted differently for patients with different characteristics, e.g. males vs. females, see also section 1.5 below.

Thus, diagnostic issues have been highlighted over the last decade and even a standard has been suggested. There is, however, a discussion regarding the distinction between "test research" and "diagnostic research" – a debate which illustrates the lack of precise definitions. Furthermore, a suggested definition of "diagnostic theory" is based on a multivariate perspective.

1.3 Briefly about laboratory medicine and reference values

Reference values are essential for being able to interpret laboratory variables. Normally such reference intervals are population based, i.e. constructed by using a sample of reference individuals randomly chosen from the intended population. Such reference intervals are intended to serve a general purpose, meaning that the given interval should be possible to use as a reference regardless of suspected target disorder, if any.

Laboratory variables are often a part of a diagnosis and reference values are claimed to be the most frequently used tool in the diagnostic work-up [37]. The development of reference values, e.g. a 95% reference interval, includes similar methodological considerations as when a diagnostic test is developed. For instance, how to define a population, sampling procedures, standardization of measurements, are all common methodological issues. Interestingly enough, available guidelines addressing methodological issues for the development of reference values were published within the field of clinical chemistry already in the late 80´s [38-44]. Methodological issues and also recommended statistical analyses are found in a comprehensive book by Harris and Boyd [45].

As earlier mentioned, it is recognized within diagnostic theory that characteristics of diagnostic accuracy, e.g. sensitivity and specificity could vary across subpopulations. Regarding reference values, variation in specificity and sensitivity, even between individuals, was described by Harris already in 1974 [46].

Harris showed that the specificity and sensitivity for a specific individual depends on the individual's steady state value and the within-individual variability.

The steady state value is the value which observations vary around if several measurements are performed within a time where no steady state changes are expected, or in other statistical terminology: the expected value of the individual. If the variation in specificity is large, dividing reference values in more homogenous subpopulations is recommended [46]. Criteria for when such division in two different subpopulations is adequate were described by Harris and Boyd 1990 [47]. In the beginning of the new millennium some new criteria have been suggested for when reference values should be "partitioned" [48-52].

In summary, the variability between individuals implying varying specificity and sensitivity of reference values was described a long time ago. There are also suggestions for when and how to adjust for this variability, e.g. criteria for when division of reference values in subpopulations is beneficial are suggested. If the between individual variance is relatively high even though appropriate partitioning has been applied, reference intervals may be of limited value anyway. In such a situation a possible solution may be to use the individual as his/her own reference, which naturally demands access to historical data on an individual level.

1.4 Diagnostic research vs. test research vs. reference values

During the renaissance of diagnostic theory a great spectra of topics have been covered. As mentioned, there is no established explicit definition of diagnostic theory, which makes it difficult to discuss how it differs, if it differs, from the theory regarding reference values. The most apparent difference is that diagnostic theory focuses on a specific condition, e.g. a disease. In the guidelines STARD mentioned above, "diagnostic accuracy" is defined as, "the ability of a test to identify a condition of interest". Noteworthy is the terminology "condition of interest" implying that a test also could be used for discrimination between other conditions than the classical "healthy" vs. "diseased". The terminology also reveals the focus on evaluation of a single test. Such univariate evaluation should, according to Moons et al., be referred to as "test research", while "diagnostic research" should emphasize the estimation of post-test probability of disease, preferably in a multivariate model [32].

The difference is that a test could be shown to have a good discriminatory ability when it is evaluated in a univariate manner, but still be redundant in the diagnostic work-up. This redundancy occurs if the test is closely interrelated with already known diagnostic information, e.g. medical history, gender and preceding tests. In such a case the post-test probability will not be significantly different from the pre-test probability, given that a multivariate logistic regression model, including existing information, is used for estimating the probabilities.

Regarding reference values these are most often intended to describe the most common values, e.g. the central 95% of the distribution, in a healthy population. However, "healthy" has no clear-cut definition and a reference population could in some cases be constituted by a population of patients with a specific relevant disease [39,5].

Generally, the aim with reference values is to support interpretation of laboratory results in clinical practice [53], which also imply that it in some cases could be valuable with references in a population with a specific health condition.

Reference values should not be interpreted as medical decision limits [38]. Noteworthy, is that there is a risk that reference values, intended as descriptive, are still used as diagnostic limits - a phenomenon that has been described as, ”the diagnosis of non-disease” [4 (chapter2)].

Thus, reference values are based on a statistical distribution most common in a healthy population and are intended to be descriptive and not to be used as clinical decision limits. Test research is suggested to be an evaluation of the ability of a test to discriminate between clinical categories in an optimal way [5], most often performed in a univariate manner. Finally, diagnostic research is suggested to include a multivariate perspective of the diagnostic work-up.

This multivariate approach suggests the use of a multivariate logistic regression model to estimate the probability of target disorder. In such a model it is reasonable to include variables that significantly affect the probability. If a laboratory variable, interpreted by using reference intervals, is shown to be significant in such a model, it may be questioned if the reference interval in that case could be considered as diagnostic decision limits, and if so, maybe these should be adjusted in order to optimize discrimination? In other words, if a reference interval is shown to significantly affect the probability of a target disorder its role may be transformed from just being referential to also being conclusive and a form of medical decision limits. This argument makes the concepts somewhat blurred.

1.5 Test accuracy

As discussed, a test is used for discrimination between different conditions, e.g. pregnant/non-pregnant, inclusion/exclusion for further tests or healthy/diseased, however, for the sake of simplicity, the conditions healthy/diseased will be used in forthcoming examples and illustrations.

To evaluate the ability of a test to discriminate between healthy and diseased, there must be a possibility to separate these two categories. The reference used for judging the “true” condition of the patient is called “gold standard”. Obviously, it is desirable to use a gold standard which is as certain as possible. The ”accuracy” of a test is often characterized by; sensitivity, specificity and likelihood ratio [54], defined as:

- Sensitivity: the probability of a positive test (T^+) in the population of diseased (D^+), i.e. diseased according to the gold standard. In symbols: $P(T^+ \mid D^+)$
- Specificity: the probability of a negative test (T^-) in the population of healthy individuals (D^-), symbolized $P(T^- \mid D^-)$
- Likelihood ratio for a positive test: $\frac{P(T^+ \mid D^+)}{P(T^+ \mid D^-)} = \frac{sensitivity}{1 - specificity}$
- Likelihood ratio for a negative test: $\frac{P(T^- \mid D^+)}{P(T^- \mid D^-)} = \frac{1 - sensitivity}{specificity}$

The characteristics described above are all based on the status of the patient, i.e. if the patient is healthy or diseased.

These measurements may be interesting in research studies, when a single test is evaluated or when different tests are being compared. However, in clinical practice, the true condition of the patient is seldom known, which makes these characteristics inapplicable. In clinical practice, it is much more relevant to make the calculation the other way around, i.e. to consider the probability of the condition, given the test result.

This could be done with the following characteristics:

- Positive predicting value: $P(D^+ \mid T^+)$, that is the probability of disease given a positive test
- Negative predicting value: $P(D^- \mid T^-)$, that is the probability of health given a negative test

The characteristics of accuracy described above depend on the choice of threshold for positive and negative tests, respectively. To study possible choices of threshold a so called ROC-curve is a common analysis [55]. Alternative analyses also include cost/benefit-considerations [12].

It is noteworthy that the definitions above are on a population level and do not include any patient- specific information at all, except the condition. If for instance, the specificity is 95% it implies that 95 out of 100 individuals, randomly chosen from the population of healthy individuals, are expected to receive a negative test.

However, on an individual level the specificity actually varies. A healthy individual whose steady state is close to the threshold has a higher probability of a false positive result than e.g. a healthy individual with a steady state around average or below. Furthermore, even if two healthy individuals should have equal steady states, they may differ in within-individual variance. Consequently, the individual with the highest within-individual variance will have the highest probability of a false positive result. Thus, the specificity and sensitivity varies between individuals (assuming varying steady states or within-individual variances).

The variation in specificity may be more homogenous in certain subpopulations than the variation in the overall population, e.g. the specificity could vary less among women than in the total population. If such homogenous subpopulations exist it may be possible to harmonize the specificity by dividing the population in subpopulation and for each subpopulation choose a threshold corresponding to the required specificity.

However, even if the specificity is harmonized the sensitivity may differ. If, for instance, the total population is divided into females and males due to lower between-individual variability among females than in the total population, it implies that the between-individual variance is greater among males. Thus, if thresholds are adjusted to give the same specificity, the sensitivity will be higher for females than compared to males.

It is also worth remembering that characteristics of accuracy may vary between total populations. For instance, the sensitivity could be better among patients with a disease in an advanced stage than patients in an earlier stage [8]. According to Moons et al. evaluation of a test in terms of the characteristics defined above is only useful in a limited number of situations. [32]. Instead, Moons et al suggest that a test should be evaluated in a multivariable manner, analyzing if the test contributes to the diagnostic process beyond what is already known.

Regarding reference values, it is common to produce a 95% reference interval based on data from a healthy population. In terms of the characteristics above, such an interval has a specificity of 95%.

It is rather common to notice the terminology “false positive” and “false negative” corresponding to a healthy individual with a positive test and a diseased individual with a negative test, respectively. The probability of a false positive is equal to 1-specificity and the probability of a false negative equal 1-sensitivity.

The risk of receiving a false positive test increases with the increased number of independent tests being performed. For instance, if a healthy individual undergoes 12 independent tests, the probability of receiving at least one false positive is greater than 50% [2]. This calculation assumes that the tests are independent. It is also worth remembering that the probability of receiving at least one positive test among the twelve tests also depends on the individual specificity, e.g. an individual with a steady state close to threshold or a high intra-individual variance for each of the different tests, is more or less guaranteed a false positive when these tests are carried out. In contrary, an individual with steady states close to average or smallest possible within-individual variances will have a low probability of receiving a false positive.

1.6 Objectives

Overall aim

The overall aim was to describe, exemplify and possibly develop the theory for reference values and diagnostic tests, especially focusing on the variability between individuals.

Study specific aims

Study I: Existing criteria for when to partition reference values are valid only for two subpopulations. The aim was to find more generally valid criteria which were also applicable for several Gaussian subpopulations.

Study II: The aim was to study a possible relationship between accommodation capability and subjective symptoms. A secondary aim was to suggest reference values possibly based on a bimodal model discriminating between individuals with vs. without symptoms. The population was composed of invited children between the ages of 6 – 10.

Study III: The aim was to evaluate and further develop a procedure used for the diagnosis of food-hypersensitivity. The diagnostic method used is considered as the gold standard for diagnosing food-hypersensitivity and includes a technique where the individual acts as its own control. The population was composed by patients with subjective symptoms and their corresponding symptom protocol.

Study IV: The aim was to discuss how interactions between test results and other variables, e.g. patient characteristics, could be taken into account.

Study V: The aim was to investigate if a computer based support system and a multivariate prevalence function could improve allocation of life support level at a dispatch centre. The study population consisted of patients calling the centre due to acute chest pain.

2. Theory and a summary of each study

2.1 Reference values and partitioning

2.1.1 General theory

Reference values are claimed to be the most widely used tool for medical decision making [37]. To be able to compare an observed value with a relevant reference is essential for interpretation. A pioneer study, where observed values are described by using probability distributions was presented 1951 [56]. To define “normal” by using a reference interval was suggested 1969 [57].

The rationale for using an interval instead of a single point as a reference is related to the natural variation between individuals [58]. The further development included a discussion regarding the reference population, i.e. a comparison between healthy individuals vs. hospitalized patients, suggested by Pryce [59]. This was one issue among several essential methodological issues, which urged a series of articles suggesting a standard for reference values [38-43]. This standard covers essential methodological issues, e.g. choice of population, sampling, standardization of measurements and procedures and statistical recommendations.

However, even if the choice of population is adequate, the sampling is performed correctly, procedures are highly standardized and statistical procedure is correct, this does not guarantee that the received reference values will be useful in practice. This is due to the variability seen among individuals even in a well-defined population, as discussed by Harris 1974 [46]. Harris illustrates that if the within-individual-variance is low relatively to the between-individual-variance, the sensitivity of a reference interval may be very limited. Consequently the clinical usefulness can be questioned.

The main problem is that a large variance between individuals results in a wide reference interval, which will make individual changes, possibly due to illness, difficult to detect. In other words, the sensitivity will be low. A possible solution to this problem, suggested by Harris, is to divide the population in more homogenous subpopulations or to use the individual as its own reference. Naturally, the latter alternative demands individual historical data.

To develop reference values for subpopulations is according to Harris useful if the standard deviation in the subpopulations is markedly lower than in the total population, i.e. at least 37% lower. This may be fulfilled if there is a great difference in population means between the different subpopulations. Another situation where a division may increase the sensitivity, at least for one of the subpopulations, is when the between-individual variance is markedly different between the subpopulations being considered. These ideas were discussed in more detail in a proceeding paper [60].

Moreover, an analysis and formalization of criteria for when it is appropriate to partition reference values in different subpopulations was suggested by Harris and Boyd in 1990 [47]. In this comprehensive study, it was emphasized that a statistical significance between subpopulations is not in itself a reason for partitioning.

Since even a small difference could be statistically significant if the sample sizes are large enough, it is clear that partitioning criteria is not equivalent with statistical significance. Partitioning should be based on a difference of a relevant magnitude between the subpopulations. A statistical significance thus does not guarantee that a division in subpopulations will be successful.

As a matter of fact, Harris and Boyd illustrated that the differences must be highly statistically significant before partitioning is worth consideration. As criteria they suggested to partition either if the ratio between larger/smaller standard deviation is at least 1.5 or if a classical z-function (difference in means standardized by standard deviations) is greater than a threshold.

The z-function is in formulas described as:

$$z = (\bar{x}_1 - \bar{x}_2)(N/2)^{1/2} / s_g$$

, where $\bar{x}_i$ represents the sample means for i=1,2, and s_g the common standard deviation and N the number of subjects in each subgroup.

The suggested threshold was a function of sample sizes, which was a clever trick to guarantee a difference of a relevant magnitude regardless of sample sizes.

For evaluation of the suggested criteria, a special case including two Gaussian subpopulations was used as an example. Unfortunately a mix between two normal distributions is non-normal in its distribution shape, which makes mathematical calculations rather complex. The combined distribution has the following cumulative distribution function:

$$C(x) = pF_1(x) + (1-p)F_2(x)$$

, where p is the first subpopulations fraction of the total population – also called prevalence, and F_i is the cumulative normal distribution function in subpopulation i, i.e.:

$$F_i(x) = \int_{-\infty}^{x} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

, where μ_i and σ_i represent the population mean and standard deviation in subpopulation i, $i=1,2$, respectively.

Since, the combined probability distribution was obtained to be of a rather complex mathematical nature demanding tedious work, Harris and Boyd instead decided to use simulations for evaluating the partitioning criteria. In these simulations it was investigated if one combined reference interval would imply that the fraction of the distribution in each subpopulation below or above the combined reference limits would deviate greatly from the nominal 2.5%. Simulations showed that the suggested criteria for partitioning corresponded to partitioning when these fractions where greater than 4% or lower than 1%, approximately.

In a large project, common reference values for the Nordic countries have been suggested [61]. In relation to this project new criteria have been suggested by Lahti et al. [48]. The suggested criteria are based on the difference between reference limits in different subpopulations, e.g. the difference between upper reference limits.

Thus the procedure demands that reference values, e.g. the upper reference value, is calculated separately in each subpopulation, and thereafter to calculate the difference between these values.

If such a difference is small it implies that pooling the reference limits into one single reference limit would not change diagnostic properties for the subpopulations that much. More specifically, Lahti et al. suggested to partition if the ratio between larger/smaller standard deviation is at least 1.5 or if the difference between two reference values, e.g. upper reference limits, divided by the smallest standard deviation is at least 0.75. These criteria were found by, in contrary to the work by Harris and Boyd, using exact calculations.

According to this exact evaluation, criteria corresponds to partitioning if proportions of the distribution above/below combined reference limits are greater than 4.1% or lower than 0.9% in any subpopulation. Another important finding in the study is that the partitioning criteria are valid only if the prevalence of each subpopulation is 50%, i.e. that each subpopulation constitutes half the total population. In a subsequent study by Lahti et al. criteria were redefined and a table with critical values, i.e. thresholds for partitioning, for various values of prevalence was presented [49].

Importantly, it was also pointed out that the earlier suggested criteria by Harris and Boyd also are restricted to situations when the prevalence is 50%. In another study by Lahti et al. non-parametrical alternatives of partitioning criteria are suggested [52]. In a study by Ichihara and Kawai multivariate analyses were used for partitioning considerations [51]. For an overview and comparison of existing methods, see the review by Lahti [50]. In this review, it is pointed out that criteria suggested by Harris and Boyd, are still the dominating criteria in guidelines, even though it is limited to only two subpopulations and fails to account for a prevalence different from 50%.

2.1.2 Study I

Aim: Existing criteria for partitioning of reference values are restricted to consider only two subpopulations, e.g. male vs. female. However, it is rather common with factors that divide the total population into more than two subpopulations.

The aim with this study was to develop criteria for situations when partitioning of several Gaussian subpopulations is considered. The developed procedure should take account to prevalences. A secondary aim was to provide a tailor-made computer program as support.

Theoretical idea: The suggested criteria by Harris and Boyd and also by Lahti et al. do all include some kind of measure, e.g. ratio between standard deviations, z-function or difference between reference limits, and to partition if these measures are greater than a threshold. The threshold is chosen in a way which guarantees that the proportions of the distribution outside combined reference limits should in each subpopulation not deviate markedly from the nominal 2.5%. For instance, proportions between 1-4% could be regarded as accepted, with such “proportions criteria” [52].

The new suggested procedure does not include such a measure. Instead, these proportions are calculated directly. Naturally, this demands that the combined reference interval must be calculated. Once the combined reference interval is obtained it is simple to calculate the proportions outside combined values in each subpopulation.

To be able to obtain the combined interval, the following values must be found: $C^{-1}(0.975)$ and $C^{-1}(0.025)$, assuming that a 95% interval is desired.

These values are difficult to find mathematically, but easy to find with an equation solver algorithm. The calculation of the cumulative probability function for a normal distribution is elementary statistical theory and thus it is simple to also calculate the combined cumulative probability function C(x) for any given value of x. Thus, with some computer power it is a straightforward procedure to simply test different values of x and iteratively find the values of x which fulfils the equations: $C^{-1}(0.975)$ and $C^{-1}(0.025)$, i.e. the upper and lower combined reference limits.

Results: A suggested algorithm was shown to be successful and it was possible to quickly and easily find the combined reference interval and thereby also proportions outside combined reference values in each subpopulation. The advantage with this procedure is that it is easy to generalize. In general, the combined cumulative probability distribution is

$$C(x) = \sum_{i=1}^{k} p_i F_i(x)$$

, where p_i correspond to prevalence of subpopulation *i* and $F_i(x)$ is the cumulative probability function in subpopulation *i*.

In the same manner as in the two-sample case it is possible to identify the combined reference interval by finding the values of x, which satisfy the equations: C(x)=0.975 and C(x)=0.025. Once the combined reference interval is obtained it is easy to, in each subpopulation, calculate the proportion of the distribution outside each combined reference value, i.e. upper and lower value.

Values that deviate notably from the nominal 2.5% implicate partitioning. A pilot version of a computer program is developed. An advantage with this method is that prevalences are taken into account and that reference intervals are automatically being calculated in each subpopulation and for the total population as well.

Summary: A procedure for considering partitioning of several Gaussian subpopulations has been suggested. The procedure takes into account prevalences and beyond a combined reference interval it automatically calculates reference intervals for each subpopulation as well. The procedure results in an output illustrating, all possible intervals, and proportions of the distributions in each subpopulation, which is below/above combined reference limits. The procedure is supported by a computer program and is thereby simple to use.

2.2 Reference values vs. medical decision limits

2.2.1 General theory – limits and their rationale

As discussed previously reference values found in clinical chemistry serves a general descriptive purpose, and is usually presented as a 95% reference interval, i.e. an interval where you are expected to find 95% of all observed values within the population. The calculation is based either parametrically assuming a specific probability distribution or non-parametrically, i.e. without any assumptions about underlying distribution [42]. When discussing references, the word "normal" is sometimes used, and this may lead to misunderstandings.

The word "normal" is ambiguous and six different meanings are described by Sacket et al.[62]. The term "Normal" could for instance refer to a Gaussian-shaped distribution of a studied variable. Furthermore, the default choice of a 95% reference interval has also been used for describing "normality". This may be regarded as inadequate since it is an arbitrary choice and since there is really no justification for why 95% of the population should be normal and the remaining abnormal. As a consequence, this could, according to Sacket et al. lead to the phenomenon "the diagnosis of non-diseased". Thus, there is a risk that reference values intended to be descriptive are over-interpreted and also used as decision limits, even though standards and literature emphasize the distinction between reference value and clinical decision limits [38,63-64].

Finally, Sacket et al. describes three other alternative definitions of normal. Firstly, the threshold for "normal" could be based on studies of risk factors, where the threshold is set based on an increased risk of future complications, e.g. cardiovascular ones.

A pure descriptive limit (reference value) and a limit based on risk evaluation could naturally differ. For instance, a Nordic reference interval for total cholesterol level includes values above 6 mmol/L [65], while "Läkemedelsverket" (Swedish food and drug administration), suggest as a target for treatment to be below 5 mmol/L or even lower for patients with increased risk [66].

Secondly, "normal" could be used in the sense that a diagnostic test was negative. Finally, it is also possible to use a therapeutic definition where the threshold is chosen based on an analysis identifying values which are associated with benefits of treatment.

In summary, there are several possible rationales for choosing a limit, ranging from simply a statistical description to cost/benefit-justification.

2.2.2 Accommodation references and diagnosis – a brief introduction

Accommodation could be defined as the ability of the eye to focus objects at various distances. The reference values used in practice today actually originate from a study published as early as 1912 [67]. Based on data from this study a reference curve accounting for age were formalized in a study by Hofstetter 1950 [68]. This reference curve illustrates the successive decreasing accommodation by age. However, the original data contains only a few observations on children below the age of 10 years. Based on an extrapolation to lower ages based on the reference curve, children are often assumed to have a good accommodation. This assumption may be invalid due to a misleading extrapolation. In a rather recent study by Sterner et al. it was shown that the accommodation amplitude of school children is not as good as expected [69].

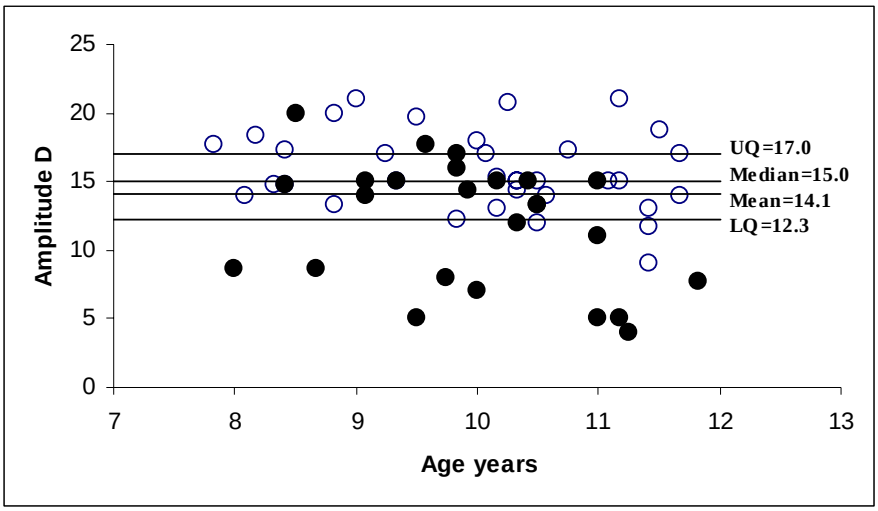
There is no established clear-cut definition of accommodation insufficiency (AI) and in earlier studies around ten different definitions are being used [70]. The term “insufficiency” is unambiguous and the underlying reason for regarding some values as insufficient is unclear. Moreover, the rationale for the choice of limit in terms of decreased abilities or symptoms may also be questioned. Finally, a recent encouraging study shows that low accommodative ability, could be improved by a simple harmless and efficient treatment, which increases accommodative ability and decreases symptoms [71].

2.2.3 Study II

Aim: To study a possible relationship between subjective symptoms at near work, e.g. write and read, and accommodation ability and to suggest reference values for school children.

Method: Children from a randomly chosen junior level school from ages 6-10 were invited to participate. This cohort was examined at two occasions with 1.8 years in between. The first examination included 72 children whereof 59 also took part in the second examination. Subjective symptoms at near work were studied by using interviews and accommodation was measured according to established methods. Regarding reference values the idea was to take a possible relation to subjective symptoms into account.

Results: It was found that subjective symptoms were related to lower accommodation, as illustrated in the following figure, illustrating binocular accommodation amplitude, age and presence of symptoms (black dots) or not (white dots):



The suggested reference values took this relationship into account. A bimodal model was used and a ROC-table complemented with positive and negative predictive values was presented (AA=amplitude of accommodation, L=left eye, R=right eye, B=binocular):

	Reference value	Proportion of children with AA less than or equal to reference value	Sensitivity	False pos.	Positive predictive value	Negative predictive value
AA (R)	7.0	20%	0.44	0.03	0.92	0.70
	8.0	29%	0.60	0.06	0.88	0.76
	9.0	32%	0.60	0.11	0.79	0.75
AA (L)	7.0	19%	0.40	0.03	0.91	0.69
	8.0	25%	0.52	0.06	0.87	0.73
	9.0	37%	0.60	0.21	0.68	0.73
AA (B)	11.0	19%	0.40	0.03	0.91	0.69
	12.0	24%	0.44	0.09	0.79	0.69
	13.0	29%	0.44	0.18	0.65	0.67

In this situation positive predictive value corresponds to the probability that an individual with a value lower than the reference limit has symptoms. Correspondingly, negative predictive value is the probability that an individual with a value above the suggested reference value manages near work without receiving symptoms. For instance, given the reference value of 11 D in binocular amplitude of accommodation, a prevalence of 19% means that around every fifth child will be “positive”.

The positive predictive value is around 90%, i.e. around nine out of ten children with an amplitude of accommodation lower than 11 D (binocular) will have symptoms at near work. Suggesting reference limits 8 D monocular and 11 D binocular indicate a prevalence of around 25%. This is much greater than the 2.5% prevalence received if the standard calculation: mean±2sd had been used, (assuming normal distribution). Such a high prevalence is still suggested due to the high probability of subjective symptoms at near work and the possibility of using a simple effective treatment.

Conclusion: Subjective symptoms at near work were in this population associated with lower accommodation. Preliminary reference values associated with risk of symptom was suggested. The suggested reference values imply a prevalence around 25%, which is motivated by the high probability of subjective symptoms and the possibility of using an effective treatment that the suggested reference values include. However, more research is needed in order to confirm appropriate reference values, especially if accommodation amplitude should be measured routinely or even screened among school children from the age of 8 years.

2.3 Diagnosis using the patient as its own reference

2.3.1 General theory

The variance found in a sample could be divided in analytical variance and variance within and between- individuals [72]. Statistical analyses could be used for separating these sources of variability [73-75]. As earlier discussed the clinical usefulness of reference values may be very limited if the variance between individuals are much greater than the variance found within an individual. If this ratio of variances, i.e. between/within individual variance, is great even after relevant partitioning and standardization of measurement procedures, a remaining possibility is to use historical data from the individual as reference, i.e. to use the individual as its own reference[46]. Appropriate analysis for following an individual over time includes time series analysis and surveillance theory [60,76-78].

Another situation when it is favorable to use an individual as its own reference is when the observed data is subjective, e.g. when the medical information is subjectively estimated by the patient using a rating scale. When subjective symptoms, described and rated by the patient, is being studied, there is beyond the variance described above, also a variation regarding how different patients actually interpret the subjective scale being used [79]. To be able to analyze symptoms on such a lower level of data, e.g. ordinal data, non-parametrical analysis is recommended [80]. Naturally, it is desirable with high reliability, both within and between observers, regarding the interpretation of data [81-83].

If the same observer measures the same individual repeatedly with low variance among data, the within-observer reliability is high. If different observers measure the same individual with low variance in data, the between-observer reliability is high.

Regarding diagnostic tests, where some kind of measurement equipment is used, methods for evaluating the precision of measurement is well described within the field of clinical chemistry [84]. It is claimed that evaluation of reliability traditionally have been more concerned by researchers of mental disorders than in other medical specialties [75].

In summary, when subjective symptoms are being studied it is an advantage if the individual can act as his/her own reference. Furthermore, high reliability is desired.

2.3.2 A brief introduction to the diagnosis of food hypersensitivity

In a survey as high a proportion as 20% of the population claims to be hypersensitive to certain foods [85]. To some degree medical attention regarding problems with food may affect the likelihood that people associate symptoms with something they have eaten [85]. However, the prevalence of confirmed food hypersensitivity is as low as 1-3% [87-89]. The gold standard for establishing food hypersensitivity is double-blind, placebo-controlled food challenges (DBPCFC), where differences in reactions are being studied after provocations with the suspected food and placebo [90].

When the symptoms are objective e.g. urticaria, one provocation with food and one placebo is regarded as sufficient, while it is recommended to use 3+3 or 3+2 provocations if the symptoms are subjective, e.g. abdominal pain [ibid]. Existing guidelines do not contain any information about the interpretation of subjective symptoms and further standardization is being asked for [92-93].

In contrast to objective symptoms it is common that subjective symptoms also occur on placebo. In such a case, a common strategy is to regard the DBPCFC as negative or to classify it as a failure and a non-interpretable challenge [93-94]. However, such conduct is actually inadequate.

Regarding patients with subjective symptoms, the symptom profile varies between patients when there is no formalized and evidence-based description over the magnitude and frequency of their symptoms. This lack of statistical description of the profile of symptoms and the subjective nature of the symptoms make it difficult to establish common diagnostic thresholds, and it is therefore preferable if each patient could act as his/her own control. The DBPCFC technique makes this possible.

The basic idea with using a placebo is to study if reactions seen on food provocation are beyond reactions seen on placebo within patient, regardless of the magnitude of placebo reactions.

2.3.3 Study III

Aim: When double-blind, placebo-controlled, food challenges are being used for diagnosing food hypersensitivity some patients suffer from subjective symptoms. The aim was to evaluate and further develop a strategy for how these subjective symptoms could be interpreted.

Method: Existing protocols from DBPCFC including at least four provocations, received from consecutive patients were reevaluated according to a pre-defined strategy, in total, 32 such protocols were included. In contrary to original diagnoses a challenge with reactions on placebo could still be positive assuming that symptoms on active provocations were beyond symptoms on placebos. For each protocol all included provocations were ranked after the magnitude of symptoms, and thereafter the blinded observer, suggested discrimination between active provocations (containing the suspected food) and placebos. For instance, if the patient had received five provocations (3 with food and 2 placebo or vice versa), the observer identified the two provocations with mildest symptoms and the two with the worst symptoms, suggesting that these were placebos and actives, respectively. Finally, the fifth provocation was judged and if the symptoms were similar to the two mildest it was suggested as placebo, otherwise as active.

This ranking approach is similar to the classical non-parametric test, Mann-Whitney U´s test [80]. For a DBPCFC including five or six provocations the challenge was judged as positive if at least five provocations were identified correctly. If the DBPCFC only included four provocations a challenge was regarded as positive only if all provocations were identified correctly.

Results: Since earlier diagnostic approach regarded challenges with reactions on placebo as negative, the new approach induced more positive challenges. Among the original diagnoses 21.9% were positive, while the new strategy gave 34.4% positive. All protocols were judged by three independent observers who sent the result (positive or negative) of each protocol to an administrator. The between-observer reliability was high, two of the observers judged all protocols equally, while the third had a different opinion for one single protocol among the 32 in total.

Conclusion: How the interpretation of the individual reference values, i.e. the placebo reactions, is being done, affects the results. A pre-defined strategy based on a ranking approach gave a high inter-observer-reliability.

2.4 Analysis of subpopulations

2.4.1 General theory – accuracy is not static

The accuracy of a diagnostic test in terms of the classical measures, e.g. sensitivity and specificity, varies across studies due to spectrum or selection bias [95-101]. Furthermore, it has been shown that accuracy could also vary across subpopulations even within a study population [33-36].

If, for instance, the specificity in the study population is 80%, but 70% among females and 90% among males (assuming equally many females/males), this information may be important to highlight in order to make it possible for the clinician to take individual characteristics in to account during the diagnostic work-up.

As discussed previously, Harris demonstrated already in -74 that specificity and sensitivity vary on an individual level, and that it may be beneficial to divide the population in more homogenous subpopulations [46]. Interesting to note is that Harris is discussing varying diagnostic accuracy on an individual level, while the findings within diagnostic theory is on a subpopulation level, quoting Moons et al. “Note that there is a true sensitivity, specificity and LR for each homogenous subgroup” [36]. This is a slightly different perspective than the individual perspective described by Harris. Naturally, a single individual can be regarded as the smallest possible subpopulation and in that sense there is a true sensitivity and specificity for that specific patient/subpopulation.

In summary, the variability in accuracy is discussed both within diagnostic theory and within the theory of reference values as well. How to account for this is thus a shared issue.

2.4.2 Study IV

Aim: The aim was to discuss two different alternatives to account for variation in diagnostic properties between subpopulations. Approaches found in diagnostic theory and for reference values are being compared.

Theory: As described previously there are criteria available for when it is appropriate to divide reference values in subpopulations, i.e. so called partitioning. These criteria are specially designed for partitioning due to a factor which divides the population in a number of possible subpopulations, e.g. gender, smoking, race. Adjustments of reference values due to a continuous variable i.e. a covariate could be done by using regression analysis [6].

Regarding factors and covariates which may affect the diagnostic accuracy in terms of sensitivity and specificity a similar approach has been suggested [34-36]. The approach starts with dividing the individuals into two groups, one with the diseased and one with the healthy ones. Within the ”diseased group”, logistic regression is used to estimate the probability of a positive diagnosis, i.e. sensitivity. Similarly, in the ”healthy group” logistic regression is used for estimating the probability of a negative diagnosis, i.e. the specificity.

By using the logistic regression models potential influence of different factors and covariates on sensitivity and specificity can be studied. It is natural to include information already known from earlier steps in the diagnostic process, e.g. gender, age, symptoms, and laboratory variables. The analyses demand that values from the diagnostic test are dichotomized into positive or negative based on a threshold. In a study by Moons et al. several factors significantly affected the sensitivity, but not always the specificity [36].

The approach using logistic regression may seem straightforward and intuitively appealing since it directly gives an overview of variables affecting the diagnostic accuracy, divided in sensitivity and specificity. However, a disadvantage is the dichotomization of data which lowers the power, i.e. decreases the probability of receiving a statistical significance. This becomes especially troublesome if one of the groups is small, e.g. that only a few individuals are diseased. It is well-known that multivariate logistic regression is a large sample procedure, demanding at least ten positive and ten negative observations per factor being analyzed [102]. Furthermore, the logistic regression does not explain why there is a difference between subpopulations, i.e. if it is due to a difference between population means or standard deviations. Under some circumstances, two subpopulations could actually have the same diagnostic character, e.g. the same specificity, even if there are differences between population means and standard deviations as well. This could occur if the standard deviation is lowest in the population with the population mean closest to the diagnostic threshold.

Thus, there may be situations where the diagnostic information could be more harmonized between subpopulations, but which are impossible to detect if the analysis described above is used.

If the studied diagnostic test is based on a continuous variable, potential relationships with other variables can be studied without dichotomization, i.e. by using F-test and standard multivariate regression. Regarding reference values it is worth noting that existing criteria for partitioning actually is divided in criterion associated with a difference in standard deviations and one associated with a difference in population means.

Regarding a general evaluation of a test and its potential value in the diagnostic work-up, it is suggested to use multivariate logistic regression [32]. A possible alternative would be to use pre-selection of variables, based on a univariate analysis and with a liberal choice of p-value, e.g. $p<0.2$ [103-104].

The idea with using a multivariate approach is to see if the test actually adds information beyond other already known characteristics. The use of multivariate logistic regression models is rather straightforward and known procedures. However, an important question to address is how to model the possible relationship between the test result and other explanatory variables in the model. For instance, assume that the presence of a disease is the target variable and that it is already known that symptom ”S”, laboratory variable ”L”, and gender “G” are predicting factors for the presence of disease. Further assume that the diagnostic test “T” includes a continuous variable, which has been dichotomized as positive/negative. For evaluation of the added value of T beyond S, L, and G a multivariate logistic regression model is applicable. However, it may be important to include interactions between T and the other explanatory variables. The disadvantage is that the number of parameters to estimate is rapidly increasing if interactions are included.

A possible alternative would be to use different dichotomizations of T, i.e. different thresholds for T, based on the relationship with other variables. If for instance, T differs in population mean between males and females, and if disease has the same additive effect of T regardless of gender, it would be possible to achieve the same diagnostic properties of T by adjusting the threshold by gender. This is analogue to the approach found for reference values, i.e. partitioning of reference values.

As previously described “test research” is suggested to be defined as univariate evaluation of a test, while “diagnostic research” is characterized by a multivariate evaluation, where the added value of the diagnostic test is in focus. However, it would be reasonable to include analyses of interactions between the test result and other relevant factors, in the concept ”test research”. Such an analysis could evaluate the possibility to harmonize the properties of a diagnostic test.

The aim with the analysis is to prepare the evaluation in a multivariate model, i.e. diagnostic research, and to be able to evaluate the test in a multivariate fashion under the correct conditions, i.e. under the correct assumptions, e.g. with adjusted thresholds or with included interaction factors.

Discussion – summary: The potential relationship between the diagnostic accuracy of a test and other factors are discussed in diagnostic theory and within the theory of reference values as well. Section 23 in the STARD-statement recommends analysis of accuracy across subpopulations [31]. However, there are no advices given regarding how this analysis should be performed and adjusted for. In this article two possibilities have been presented, firstly the possibility to use a multivariate model including interactions, secondly to adjust thresholds for the test in order to harmonize the accuracy across subpopulations. It is recommended to develop further guidelines regarding this issue.

2.5 Individually adjusted diagnostic information and computer support

2.5.1 Logistic regression – prevalence function

By using the classical and well established analysis: multivariate logistic regression, the probability of target disorder could be estimated, given relevant diagnostic information in terms of one or several explanatory variables. This probability could be regarded as a comprehensive index summarizing the existing medical information on a subpopulation or even on a patient level. This index could be used when making medical decisions, e.g. if the probability of disease is high, i.e. greater than a threshold, treatment may be implied, otherwise not [2]. The choice of threshold should be based on decision theory and evaluation of cost/benefits [2,12]. The explanatory variables could either be factors, i.e. variables that could take a limited number of values, e.g. gender, race, positive/negative test, etc or covariates, i.e. continuous variables.

Since the multivariate logistic regression model estimates the probability of disease, a suitable name of such a function is "prevalence function". Due to the mathematical complexity of such a prevalence function, it is favorable to use some computer assistance, when implemented in practice. A computer based decision support system is defined by a system where medical knowledge and patient data are used for generating support for medical decisions [13].

2.5.2 Briefly about priorities at the dispatch centre

The dispatchers at the dispatch centre constantly make medical decisions, e.g. prioritize the patient's need for help and ambulance transportation. In a study performed in Göteborg all consecutive patients complaining about chest pain, were studied during a three month period.

The dispatchers were instructed to interview these patients and address questions according to a standardized questionnaire. Depending on the answers, the dispatcher estimated the risk of an acute myocardial infarction or other life threatening conditions. If the risk was judged to be low the patient was allocated basic life support, whereas a high risk implied allocation of advanced life support, including the use of a specially equipped ambulance.

All patients were followed up and their final diagnoses were recorded. After the final diagnosis it was also judged by a clinician, whether their condition had been life-threatening or not. Furthermore, it was recorded if the patient survived or not. It was shown that a relationship between the estimated risk and final diagnosis and conditions existed [105], even though some patients with an acute myocardial infarction had received basic life support [106].

2.5.3 Study V

Aim: To evaluate if a computer based decision support system could be useful for the emergency medical system for identifying patients with acute myocardial infarction or life threatening conditions and thereby improve allocation of life support level.

Method: The study included data from 503 consecutive patients who called the dispatch centre due to chest pain. The data was used for estimating a prevalence function, i.e. a multivariate logistic regression model, which estimates the probability of the target condition. The primary target condition was the presence of acute myocardial infarction (AMI). The prevalence function estimated the probability of AMI for each individual patient.

A retrospective allocation was done based on the prevalence function, i.e. if the probability of AMI was high, i.e. above a threshold, advanced life support was allocated otherwise basic life support was allocated. No formalized evaluation of costs and benefits was done. However, to be able to evaluate the use of the prevalence function, the model was restricted to using advanced life support just as frequently as in the original data. This was done by choosing a threshold which gave the same distribution between basic and advanced life support as the original allocations.

Furthermore, the number of patients with AMI who were allocated basic life support by using the prevalence function was compared with the same number found among the original allocations. Thus, it was possible to compare the medical risk (AMI patient allocated basic life support) given the same economical budget. The following table illustrates a comparison between the original allocations (dispatchers' allocation) and the allocations according to the model based on the prevalence function (model allocation). The distribution of allocations to basic life support (BLS) vs. advanced life support (ALS) by diagnosis of AMI is compared:

			Level of life support Dispatchers allocation		**Level of life support Model allocation**	
			BLS	**ALS**	**BLS**	**ALS**
AMI	**No** N=398 (79%)	n	107	291	114	284
		%	26.9	73.1	28.6	71.4
	Yes N=105 (21%)	n	15	90	8	97
		%	14.3	**85.7**	7.6	**92.4**
Total	N=503	n	122	381	122	381
p-value=0.167, regarding sensitivity (McNemar)						

According to the table above, 85.7% of all AMI-patients were allocated advanced life support by the dispatchers. If the allocations had been based on the prevalence function 92.4% of all AMI-patients would have been allocated advanced life support. Or, in other words, among the 105 patients with AMI, the dispatchers called for BLS for 15 of those patients as compared with 8 allocated to BLS by the model. Observe that the number of ALS and BLS is exactly the same in the allocation based on the model as in the original allocations. The analysis was also repeated but with the presence of life threatening conditions as target variable.

Results: The comparison between allocations made by using the prevalence function vs. original allocation made by the dispatchers was made with the restriction that the advanced life support was allocated as frequently. This was achieved by adjusting the threshold for the probability of target disorder implying allocation of advanced life support.

15 patients with AMI where allocated basic life support by the dispatchers. The corresponding figure found when evaluating allocation based on the prevalence function was 8. Interestingly, among the patients with AMI allocated basic life support by using the prevalence function were not found among the 15 originally erroneously allocated AMI patients. Among the 15 originally erroneously allocated AMI-patients 9 died, while only one of the 8 erroneously allocated by the model died.

Conclusion: A computer based decision support system including a prevalence function could be a valuable tool for allocating the level of life support. However, the case record form used for the interview can be refined and a model based on a larger sample and confirmed in a prospective study is recommended.

3. Discussion, conclusions and future development

3.1 Discussion

3.1.1 Study I

The suggested method is a straightforward method which directly calculates the proportions outside combined reference vales for each subpopulation. The procedure is supported by a computer program which produces both partitioned and combined reference interval, and misclassification rates in each subpopulation if the combined reference interval is used. This overview is valuable for considering partitioning or not.

The main advantage with the procedure is that it is also applicable for considering partitioning of several subpopulations. Another advantage with the procedure is the automatic calculation of both separated and combined reference intervals. These calculations take account to prevalences even if the sample fractions do not reflect the prevalences. No resampling or other data management procedures are needed.

Unfortunately, there is no distinct partitioning criteria suggested and further development is needed. If for instance, two subpopulations can be combined but deviates from a third, it is not clear if the two should first be pooled and regarded as one population to be compared with the third or if all three subpopulations should be considered jointly. Perhaps it is possible to use a stepwise procedure, i.e. a clustering analysis, according to the following steps:

1. Compare all subpopulations pair wise. Pool the pair of subpopulations which are closest to each other, e.g. smallest difference between reference values.
2. Consider the two pooled subpopulations as one subpopulation and repeat the procedure, i.e. step 1.
3. Continue the “loop” until the most similar subpopulations are not similar enough for pooling, i.e. according to partitioning criteria.

A limitation of the suggested procedure is that it is only valid for Gaussian subpopulations, or subpopulations with a probability distribution possible to transform into a normal distribution.

However, it is simple to construct an analogous non-parametrical procedure, just by replacing the cumulative normal probability functions by the corresponding cumulative empirical distribution functions.

3.1.2 Study II

Existing reference values for accommodation amplitude are basically based on the classical calculation: mean ± 2 standard deviations, adjusted for age [69]. Previous studies include several different definitions of accommodation insufficiency (AI), more or less related to the existing reference values [71]. Since the diagnosis is called accommodation insufficiency, it is reasonable to wonder in when and in what situations the accommodation ability is insufficient, consequences of this insufficiency and possible treatment. In this study, it was actually shown that there is a relationship between low accommodation amplitude and the presence of subjective symptoms at near work. The suggested reference values are related to the probability of these symptoms among school children.

The suggested reference values imply 25% prevalence, i.e. children with lower values than the reference value, which may be regarded as high. This high prevalence may be partly explained by a bias induced by a higher frequency of symptoms and low accommodation among children who accepted an invitation, than it would be with a totally random sample. Another limitation of the study is that all measurements were made by the same observer and the reliability has not been investigated.

However, even though the high prevalence may be overestimated, the suggested reference values may still imply a high prevalence. Nevertheless, this could be motivated since there is an efficient and harmless treatment available. However future studies, including reliability studies and confirmative studies of reference values are needed, especially if accommodation amplitude should be measured routinely or even screened among school children.

Regarding the diagnosis of accommodation insufficiency a possible definition would be a combination of accommodation amplitude below reference value and the presence of subjective symptoms. To make the diagnosis even more certain it is possible to confirm the diagnosis further if symptoms disappear after improving the accommodation amplitude by treatment (training).

3.1.3 Study III

Double-blind, placebo-controlled, food challenges (DBPCFC) are regarded as the gold standard for diagnosing food hypersensitivity. As late as 2004 a review article presented recommendations regarding how to perform DBPCFC [90]. Strangely enough, this review does not contain any information about how reactions should be interpreted and no information about appropriate thresholds and their rationale.

Since the method is gold standard it is not only used in clinical practice, it is also used as a reference test when new diagnostic tests are being evaluated. Naturally, it is desirable that the reference test is as certain as possible. Unfortunately there is no formal evaluation of the gold standard and its precision, even though such an evaluation is actually possible to achieve. Regarding objective symptoms, it would be possible to study variability within and between-individuals and thereby develop adequate thresholds for when reactions seen on active provocation is reasonably beyond possible reactions on placebo. Such thresholds could in this case be population based.

Regarding subjective reactions, the suggested ranking procedure including 3+3 or 2+3 provocations is comparable with the classical non-parametrical rank-sum test, i.e. Mann-Whitney U´s test. It has earlier been suggested to use 3+3, but we suggest the possibility to use 2+3 (or 3+2). The reason for this is that if the patient is told that there will be 3 active and 3 placebos, there are 20 possible sequences to chose between, i.e. the probability to correctly guess the sequence is one out of twenty. However, if the patient is told that there will be 3 actives and three placebos or vice versa, there will still be 20 possible sequences. This indicates that the risk of a false positive could remain the same even with one provocation less. Since this could save time, money and harm for the patient, this was appreciated and accepted as a standard at a specialist meeting, when the review paper was being structured and discussed [90].

The ranking and identification of active and placebo provocations, respectively, is a task for both the clinician and the patient. The clinician uses experience partly on a populations level, i.e. the characteristics of similar patients, while the patient uses his/her own experience of how the symptoms usually appear. In this way, both “between-patients-knowledge” and “within-patient-knowledge” is used mutually. The study illustrates that the ranking approach gave high between-observer-reliability.

The study also shows that different approaches gave different results, which indicates the need for further standardization and establishing criteria for how to interpret. There is also a need for studies considering cost/benefits and the precision of the DBPCFC.

3.1.4 Study IV

The variability of test accuracy across subpopulations is recognized and discussed in diagnostic theory and within the theory of reference values as well. To use multivariate analyses are rather well-established in epidemiological and therapeutic studies, but is now also suggested as a tool in diagnostic studies. However, there is no guideline available for how to adjust for interactions between a test and other diagnostic predicting variables, e.g. gender and age. In this debate article, two different possibilities are discussed. One possibility is to mimic the approach found for adjusting reference values/thresholds to different subpopulations, i.e. partitioning. The second possibility presented, is to include interaction effects in the multivariate logistic regression model.

In therapeutic studies it is rather common to assume that the treatment effect is additive. With a similar assumption regarding the effect of a present disease this could make analysis more powerful. Different alternatives must be tested and evaluated by using real data.

Regarding laboratory variables and corresponding reference values these are, as discussed previously, intended to be descriptive. However, if a laboratory test result, dichotomized as “positive” or “negative” is shown to be significant in a multivariate logistic regression model, there may be a reason for reevaluation of the limits, due to the fact that the variable has been shown to also be useful as a decision limit. This has never been discussed before and further studies and guidelines for how to successively build a diagnostic multivariate model must be further developed.

3.1.5 Study V

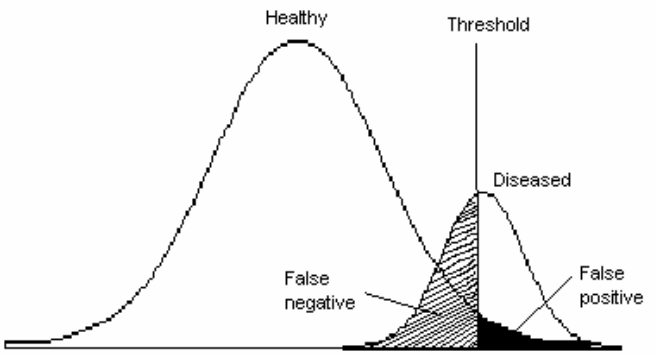
The study illustrates that a prevalence function with individually estimated probabilities of target disorder may be beneficial. Naturally, such a system requires adequate input and during the last years differences in symptom profiles due to gender have been highlighted. In this study, models including interaction factors were tested in an explorative manner, but did not improve the results.

However, in a forthcoming study it may be important to analyze these differences in more detail. It is also interesting to consider if the dispatchers with their experience could add information beyond the formal description of the patient according to a questionnaire. This experience may be taken into account by adding the judgment made by the dispatcher as a factor in the model. This was also tested in this study in an explorative manner, but did not improve results.

In the study, the numbers of advanced and basic life support allocations were fixed to be equal as the original allocations. This was done in order to make the comparison of correct medical allocations under equal economical circumstances. In an explanatory manner, it was also analysed, even though it was not included in the paper, if the economical budget could be decreased, given the same medical precision. It was found that if the threshold for when advanced life support should be allocated was increased to a degree which gave 15 AMI-patients allocated the basic life support (the same frequency of "medical error" as in the allocation made by dispatchers), the frequency of advanced life support allocations could be decreased with around 50.

3.1.6 Overall discussion

Limits and their rationale
To facilitate interpretation of medical information it is common to use some kind of limit. There are several different rationales for the choice of a limit. Within laboratory medicine and the theory considering reference values, the limits, i.e. reference values are solely descriptive in their purpose. This is reasonable due to the role as general reference regardless of potential target disorder, even though the reference values may be transformed to medical decision limits if it is shown that the laboratory variable significantly affects the probability of the target disorder. In such a situation it may be worth re-evaluating the existing limits and adjusting these in an optimal way related to the specific target disorder. When the limit is intended to be used for a specific present target disorder, i.e. a diagnostic test, the choice of limit should be based on an optimal cut-off between healthy and diseased individuals. The difference between these two populations is often illustrated by the following figure, illustrating the distribution of the results, by population:

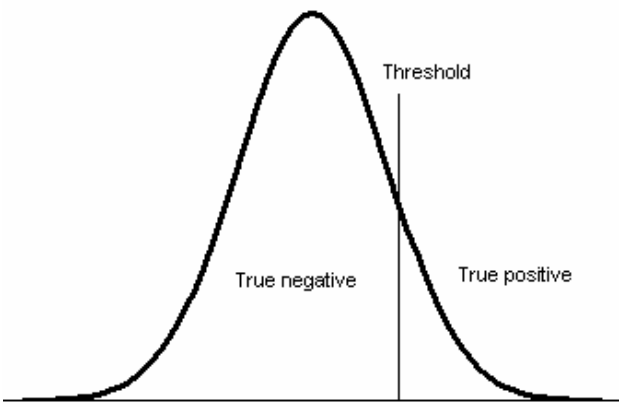


In the figure above, values above the threshold are considered as positive (assuming that the presence of disease is related to increased values), and values below as negative. It is obvious that the properties of the test, i.e. sensitivity and specificity, change if the threshold is changed, e.g. if the threshold is moved to the left, it will increase the sensitivity (more diseased are being detected) but decrease the specificity (more false positives).

This illustration often considers "healthy" vs. diseased, but naturally it could also be used for discrimination of other populations. Regarding the reference values suggested for accommodation amplitude among school children, these references considered children with subjective symptoms at near work vs. children without symptoms. In general, a diagnostic test aims at discriminating between two different health conditions, e.g. healthy/diseased; symptoms/no symptoms, etc.

It is reasonable that the explicit choice of threshold takes cost/benefits into account. The division in two subpopulations as in the model above is a model of a bimodal nature.

Regarding a risk factor a model for describing the choice of limit and consequences, may be viewed as a unimodal model according to the following figure [107]:



In this unimodal model all individuals belong to the same population and the “prevalence of positive values” is related to the choice of threshold. In this model, there are no false positive or false negative results.

However, it is actually possible to apply a bimodal perspective even when considering risk factors. Since a risk factor is used for predicting potential future target disorders, e.g. a high cholesterol level implies an increased risk of acute myocardial infarction, a bimodal model could be used for discrimination between the two possible future conditions, e.g. AMI or not.

From this point of view the bimodal model is applicable no matter if it is supposed to discriminate healthy individuals from individuals with a present disease or a predicted future disease. To be able to study the discrimination of future conditions there is naturally a need for longitudinal studies of epidemiological or therapeutical nature.

According to Jørgensen et al. the threshold used in the unimodal affects the prevalence [107]. This is true if ”prevalence” describes the proportion of individuals with a positive test value. However, the prevalence of the potential future target disorder is unaffected by the choice of threshold. Thus it is important to differentiate between prevalence meaning “proportion of positive tests” and prevalence meaning ”proportion of individuals stroked by the future target disorder”. Perhaps the bimodal illustration would make this distinction more clear, and if it is applied for a risk factor, false positives correspond to individuals with a positive test who will not be stroked by the future target disorder, while false negatives are individuals with a negative test, but still will be stroked by the future disorder.

Thus, the illustrative bimodal model is actually applicable not only for diagnostic tests, but for considering risk factors as well. The main difference is that in a diagnostic test a healthy population is compared with a population with a present disorder, while modeling risk factor means comparing healthy individuals which will remain healthy (not stroked by the disorder in the future) with healthy individuals which, in the future, will be stroked by the disorder.

If a variable used in a diagnostic test also is used as a risk factor the concepts may be confused.

For instance, lipid levels in blood, where increased values (hyperlipidemia) imply increased risk of future disorders, e.g. AMI. Hyperlipidemia could be atopic or lifestyle conditioned. To find a limit, there are four different possibilities. Firstly, a statistical description in terms of a reference interval can be developed. Secondly, a bimodal model for a present disease could be applied and an adequate threshold chosen.

The third alternative is to regard hyperlipidemia as a risk factor and apply a bimodal model discrimination between those who will be stroked by the future disorder and those who will not. Finally, it would be possible to choose a threshold based on analysis of finding a lower limit for which treatment will still be beneficial. As a matter of fact, regarding lipids there is a rather recent study implying that the lowest possible limit still beneficial to treat, is much lower than the present level [108]. Consequently, the latter approach could more or less imply prevalences (of positive tests) close to 100%, at least in the study population.

Regarding accommodation amplitude among school children the suggested reference values were related to the presence of symptoms at near work.

Since the symptoms occurred during near work, this type of work could be regarded as a kind of provocation, which makes the diagnosis of accommodation insufficiency similar to the diagnosis of food hypersensitivity. A main difference is that the symptoms occurring after eating a suspected food cannot be related to any objective measure. To be able to judge if the symptoms occurring truly are associated to the food placebo, provocations are used. This gives the possibility to use the patient as his/her own control.

Adjusting for individual characteristics
Within laboratory medicine partitioning reference values in subpopulations is a suggested solution to the problem with varying sensitivity and specificity across subpopulations. Criteria for when partitioning is appropriate are related to “proportion criteria”, i.e. partitioning is recommended if the proportion of the distribution of a subpopulation outside a combined reference value markedly differs from the nominal value (usually 2.5%).

Regarding diagnostic tests, there are no formalized criteria for when and how adjustments by subpopulations should be made. Two possible alternatives are discussed: either to mimic the approach with partitioning test threshold by subpopulation or to include interaction effects in the diagnostic prevalence function. However, further evaluations of these possibilities are necessary to perform before any general guidelines can be suggested. In some simple situations where there is a pure additive difference between subpopulations, e.g. male and female patients, harmonization of the test accuracy may be created just by choosing different thresholds, i.e. a kind of partitioning. To use a prevalence function with an interaction effect demands rather complex statistics and computer support is more or less essential. Moreover, if computer assistance is used, it may be possible to use continuous test variables as input in the model, without dichotomization.

3.2 Conclusions

- A bimodal model can be used not only for discrimination between healthy vs. diseased but for discrimination between other subpopulations as well, e.g. between individuals with/without symptoms. As an illustration, subjective symptoms at near work among school children were studied and shown to be associated with lower accommodation amplitude. Reference values for amplitude of accommodation were suggested based on a bimodal model discriminating between children with vs. without symptoms occurring at near work.

- In situations where the variability between individuals is high compared to the variability found within an individual, it may be necessary to use the individual as his/her own reference. Another situation when it is favorable to use the individual as his/her own reference is when the diagnostic information is subjective. The diagnosis of food-hypersensitivity for patients with subjective symptoms was used as an illustration. It was shown that the strategy used for interpretation affects the diagnostic outcome. Furthermore, it was shown that a pre-defined strategy based on a ranking approach gave a high inter-observer reliability.

- To harmonize the sensitivity and specificity of reference values across subpopulations division of references in subpopulations is one suggested solution. A procedure for considering partitioning of several Gaussian subpopulations has been developed. The procedure takes into account prevalences and in addition to a combined reference interval it automatically calculates reference intervals for each subpopulation. The procedure is supported by a computer program and is thereby simple to use.

- The potential relationship between diagnostic accuracy of a test and other factors are highlighted in diagnostic theory. However, there is no advice regarding how to adjust for this relationship. Two possibilities have been presented, firstly the possibility to use a multivariate model including interactions, secondly to adjust thresholds for the test in order to harmonize the accuracy across subpopulations. It is suggested that “test research” should include such analysis.

- Diagnostic information could be individually adjusted by using a prevalence function used to estimate probability of target disorder, given patient characteristics. A computer based decision support system including such a prevalence function was shown to have potential benefits for assisting medical decisions.

3.3 Future development

Partitioning of reference values:

- The development of more precise criteria based on the suggested procedure and the discussed clustering technique is already initiated and started.
- The development of a non-parametrical version of the procedure is planned. This version should simply replace the cumulative normal distribution function with the empirical cumulative function according to the established standard.

Reference values for amplitude of accommodation

- The suggested reference values must be confirmed in larger studies.
- A larger study would make it possible to also study if any partitioning, e.g. due to gender is appropriate.
- It is important to use a relevant population and clinical spectra. That is, if accommodation amplitude should be recommended to undergo screening, a random sample of school children must be used. Furthermore, measurements must be taken by the same professionals as those expected to measure routinely if screening is launched.
- Evaluations of reliability within and between observers are desirable.

The diagnostic work-up for food-hypersensitivity

- It would be beneficial to further develop guidelines for how to interpret symptoms, both objective and subjective symptoms. A common strategy would improve the diagnostic work-up and is necessary for being able to compare research studies.
- Evaluations of the gold standard and its reliability are desirable and actually possible to perform.
- A systematic collection of symptoms in a formalized and user-friendly computer system would benefit future development and understanding of food-hypersensitivity.

Adjustments to subpopulations and analytical approaches

- There is a need for formalizing terminology, e.g. what is diagnostic theory, test theory, etc. This thesis only highlights some similarities and some differences, but the concepts are still blurred. What is, e.g. distinct definition of a diagnosis?
- It is recommended to further develop recommendations regarding statistical analysis suitable for test research and diagnostic research.
- If a laboratory variable reported as “positive” or “negative” according to reference values is shown to be a significant factor in a prevalence function, it may be worth a re-evaluation of adequate limits. This is a situation were the laboratory variable actually is a diagnostic test and therefore general reference values may not be optimal thresholds.

Individually adjusted diagnostic information and computer support

- The results found in the study must be confirmed in a prospective study. This work has already been initiated.
- It is important to also develop a user-friendly system and also highlight ethical questions.
- The prevalence function in the study used AMI as a preliminary target variable. This is thus a sub-system for patients with e.g. acute chest pain. It must be evaluated as to

how such a sub-system should interact with other sub-systems and existing overall-systemsprioritization.

- Using multivariate logistic regression is one possibility for constructing a prevalence function. However, there are several other alternatives, including machine learning methods, e.g. neural networks, Tree analysis, etc, to be tested.

Acknowledgements

My supervisor Martin Rydmark is one of the pioneers in Sweden within the field of Medical Informatics. Thanks to a book in Medical Informatics, edited by Martin and a colleague of his, I became interested in the interpretation of medical information and diagnostic tests. Martin is humble and humoristic and a chat with him always ends with a smiling face. Without advices from Martin the conferences in sunny California would never have been as fruitful.

The true Aphrodite of sciences is statistics. I am grateful to Ziad Taïb for being an excellent teacher and for introducing me to this numerical beauty.

I am deeply grateful to all co-authors who have introduced me to very interesting applications. With energy and inspiration Bertil Sterner introduced me to ophthalmology. I truly admire Ulf Bengtsson for all his care about patients with food hypersensitivity, and I want to thank him, Jenny Magnusson, Staffan Ahlstedt and Urban Gråsjö for good cooperation in all the work with developing the diagnostic work-up for these patients. Finally, I am grateful to Angela Bång and Johan Herlitz for sharing all their experience within pre-hospital medicine and patients with cardiovascular disorder - we got a really good start on forthcoming mutual research within this field.

I am grateful to my colleagues Tobias Arvemo and Lars Svensson for being good listeners and frustration resolvers. My colleague Will Jobe has frenetically helped me, even through nice summer days, with proof reading and programming tasks. Thanks Will – I am looking forward to forthcoming cooperation in research projects. Many thanks also to Irene Johansson who helped me with preparing the manuscript.

Finally and most importantly, I am grateful to my beloved wife Ulrika and our wonderful children Victoria and Wilma, for bringing my attention to more important things than research.

References

1. Russel B. A History of Western Philosophy, New York: Simon and Schuster 1945.
2. Sox HC, Blatt MA, Higgins MC, Marton KI: Medical Decision Making. Butterworth Publishers, Stoneham 1988.
3. The declaration of Helsinki, available at http://www.wma.net/e/ethicsunit/helsinki.htm
4. Knottnerus JA ed. The evidence base of clinical diagnosis. London BMJ books 2002.
5. Burtis CA, Ashwood ER (editors). Tietz textbook of Clinical Chemistry 2nd ed. Philadelphia W.B. Saunders Company 1994.
6. Altman DG. Practical Statistics for Medical Research. London:Chapman&Hall 1991.
7. Bland M. An introduction to medical statistics. 3rd ed Oxford:Oxford University Press 2000.
8. Haynes BR, Sackett DL, Guyatt GH, Tugwell P. Clinical Epidemiology: How to Do Clinical Practice Research. 3rd ed. Lippincott Williams & Wilkins 2006.
9. Knottnerus JA. Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. Med Dec Making 1992;12(2):93-108.
10. Sacket DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324(7336):539-41.
11. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324(7338):669-71.
12. Hilden J. The area under the ROC curve and its competitors. Med Dec Making 1991;11(2):95-101.
13. van Bemmel JH, Musen MA ed. Handbook of Medical Informatics. Heidelberg: Springer 1997.
14. Irwig L, Tosteson ANA, gatsonis C, . Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120(8):667-76.
15. Cochrane AL. Effectiveness and efficiency. Random reflections on health services. The Nuffield Provincial Hospitals Trusts, 1972. Reprinted: London, the Royal Society of Medicine Press Limited, 1999.
16. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA 1995;274(8):645-51.

17. Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. JAMA 1984;252(17):2418-22.
18. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Van Der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282(11):1061-6.
19. Bogardus ST, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research. The need for methodological standards. JAMA 1999;281(20):1919-26.
20. Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33(1):85-91.
21. Niggemann B, Grüber C. Unproven diagnostic procedures in IgE-mediated allergic diseases. Allergy 2004;59(8):806-8.
22. Knottnerus JA, Van Weel C., Muris JWM. Evaluation of diagnostic procedures. BMJ 2002;324(7335):477-80.
23. Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making:selective review of the cognitive literature. BMJ 2002;324(7339):729-32.
24. Winkens R, Dinant G-J. Rational, cost effective use of investigations in clinical practice. BMJ 2002;324(7340):783-5.
25. Foy R, Warner P. About time: diagnostic guidelines that help clinicians. Qual Saf Health Care 2003;12(3):205-9.
26. Gluud C, Gluud LL. Evidence based diagnostics. BMJ 2005;330(7493):724-6.
27. Bruns DE, Huth EJ, Magid E, Donalds SY. Towards a Checklist for Reporting of Studies of Diagnostic Accuracy of Medical Tests. Clin Chem 2000;46(7):893-5.
28. Bruns DE. Editorial. The STARD Initiative and the Reporting of Studies of Diagnostic Accuracy. Clin Chem 2003;49(1):19-20.
29. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. Clin Chem 2003;49(1):7-18. Also available at: http://www.consort-statement.org/stardstatement.htm
30. Moher D, Schultz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. BMC Med Res Met 2001;1:2. Also available at: http://www.biomedcentral.com/1471-2288/1/2
31. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D et al. The Revised CONSORT Statement for Reporting Randomized Trials:Explanation and Elaboration. Ann Intern Med 2001;134(8):663-94.

32. Moons KGM, Biesheuvel CJ, Grobbee DE. Test Research versus Diagnostic Research. Clin Chem 2004;50(3):473-6.
33. Moons KGM, Harrell FE. Sensitivity and specificity should be deemphasized in diagnostic accuracy studies. Acad Radiol 2003;10(6):670-2.
34. Hlatky MA, Pryor DB, Harrell FE, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77(1):64-71.
35. Levy D, Labib SB, Anderson KM, Christiansen JC, Kanell WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. Circulation 1990;81:815-20.
36. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8(1):12-7.
37. Sasse E. Reference intervals and clinical decision limits. In: Kazmierczak S ed. Clinical Chemistry: Theory, analysis, and correlation. St. Louis, MO:Mosby 2003:362-78.
38. Solberg HE. Approved recommendation (1986) on the Theory of Reference Values. Part 1. The Concept of Reference Values. J Clin Chem Clin Biochem 1987;25(5):337-42.
39. Petitclerc C, Solberg HE. Approved recommendation (1987) on the Theory of Reference Values. Part 2. Selection of Individuals for the production of reference values. J Clin Chem Clin Biochem 1987;25(9):639-44.
40. Solberg HE, Petitclerc C. Approved recommendation (1988) on the Theory of Reference Values. Part 3. Preparation of Individuals and Collection of Specimens for the production of Reference Values. J Clin Chem Clin Biochem 1988;26(9):593-98.
41. Solberg HE, Stamm D. Approved recommendation on the Theory of Reference Values. Part 4. Control of Analytical Variation in the Production, Transfer and Application of Reference Values. Eur. J Clin Chem Clin Biochem 1991;29(8):531-35.
42. Solberg HE. Approved recommendation (1987) on the Theory of Reference Values. Part 5. Statistical Treatment of Collected Reference Values. Determination of Reference Limits. J Clin Chem Clin Biochem 1987;25(9):645-56.
43. Dybkaer R, Solberg HE. Approved recommendation (1987) on the Theory of Reference Values. Part 6. Presentation of Observed Values Related to Reference Values. J Clin Chem Clin Biochem 1987;25(9):657-62.
44. International Federation of Clinical Chemistry (IFCC) Expert Panel on Theory of Reference Values. The theory of reference values, statistical treatment of collected

reference values:Determination of reference limits. J Clin Chem Clin Biochem 1987;25:645-56.

45. Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York:Marcel Dekker, Inc, 1995.
46. Harris EK. Effects of Intra- and Interindividual Variation on the Appropriate Use of Normal Ranges. Clin Chem 1974;20(12):1535-1542.
47. Harris EK, Boyd JC. On Dividing Reference Data into Subgroups to Produce Separate Reference Ranges. Clin Chem 1990;36(2):265-270.
48. Lahti A, Hyltoft Petersen P, Boyd JC, Fraser CG, Jørgensen N. Objective criteria for partitioning Gaussian-distributed reference values into subgroups. Clin Chem 2002;48(2):338-52.
49. Lahti A, Hyltoft Petersen P, Boyd JC. Impact of Subgroup Prevelences on Partitioning of Gaussian-distributed Reference Values. Clin Chem 2002;48(11):1987-1999.
50. Lahti A. Partitioning biochemical reference data into subgroups: comparison of existing methods. Clin Chem Lab Med 2004:42(7):725-733.
51. Ichihara K, Kawai T. Determination of Reference Intervals for 13 Plasma Proteins Based on IFCC International Reference Preparation (CRM470) and NCCLS Proposed Guideline (C28-P,1992): A Strategy for Partitioning Reference Individuals With Validation Based on Multivariate Analysis. J Clin Lab Anal 1997;11:117-124.
52. Lahti A, Hyltoft Petersen P, Boyd JC, Rustad P, Laake P, Solberg HE. Partitioning of nongaussian distributed biochemical reference data into subgroups. Clin Chem 2004;50(5):891-900.
53. Solberg H. Using a hospitalized population to establish reference intervals: Pros and Cons. Clin Chem 1994;40(12):2205-6.
54. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986;134(6):587-94.
55. Hanley JA, McNeil BJ. The meaning and use of the area under receiver operating characteristics (ROC) curve. Radiology 1982;143:29-36.
56. Wooton I, King E, Smith J. The quantitative approach to hospital biochemistry. Br Med Bull 1951;7:307-11.
57. Gräsbeck R, Saris N-E. Establishment and use of normal values. Scand. J. Clin. Lab. Invest. 1969;24:62-63.
58. Wooton I. Individual variation. Proc Nutr Soc 1962;21:129-35.

59. Pryce J. Level of haemoglobin in whole blood and red blood cells, and proposed convention for defining normality. Lancet 1960;2:333-6.
60. Harris EK. Some Theory of Reference Values. I. Stratified (Categorized) Normal Ranges and a Method for Following an Individual's Clinical Laboratory Values. Clin Chem 1975;21(10):1457-1464.
61. Rustad P. Nordic reference Interval Project. Described at http://www.furst.no/norip/
62. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: a Basic Science for Clinical Medicine, 2nd edn. Boston: Little, Brown and Co, 1991;p 58-61.
63. Burtis CA, Ashwood ER (editors). Tietz textbook of Clinical Chemistry 2nd ed. W.B. Saunders Company 1994. p. 471.
64. Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York:Marcel Dekker, Inc, 1995, p. 21-22.
65. Information available at: http://www.furst.no/norip/
66. Information (In Swedish) available at: http://www.lakemedelsverket.se/upload/H%C3%A4lso-%20och%20sjukv%C3%A5rd/behandlingsrek/HRT2003.pdf
67. Duane A. Normal values of the accommodation at all ages. J Am Med Assoc. 1912;59:1010-3.
68. Hofstetter HW. Useful age-amplitude formula. Optom World. 1950;38:42-5.
69. Sterner B, Gellerstedt M, Sjöström A. The amplitude of accommodation in 6-10-year-old-children –not as good as expected! Ophthal. Physiol, Opt. 2004;24:1-6.
70. Cacho, P, Garcia A, Lara F and Segui M. Diagnostic signs of accommodative insufficiency. Optometry and Vision Science. 2002;79:614-620.
71. Sterner B, Abrahamsson M, Sjöström A. Accommodative facility training with a long term follow up in a sample of school aged children showing accommodative dysfunction. Documenta Ophthalmologica 1999;99:93-101.
72. Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. Crit Rev Clin Lab Sci. 1989;27(5):409-37
73. Harris EK. Distinguishing physiological variation from analytic variation. J Chron Dis. 1970;23:469.
74. Young DS, Harris EK, Cotlove E. Biological and analytical components of variation in long-term studies of serum constituents in normal subjects. IV. Results of a study designed to eliminate long-term analytical deviation. Clin Chem 1971;17:403.
75. Fleiss JL. The design and analysis of clinical experiments. John Wiley & Sons Inc. 1986.

76. Harris EK, Cooil BK, Shakarji G, Williams GZ. On the Use of Statistical Models of Within-Person Variation in Long-Term Studies of Healthy Individuals. Clin Chem 1980;26(3):383-91.
77. Frisén M, De Maré J. Optimal surveillance. Biometrika.1991; 78: 271-280
78. Garfield FM. Quality Assurance Principles for Analytic Laboratories, 2nd edn. Washington, DC: Association of Analytical Chemists, 1991.
79. Svensson, E. Analysis of systematic and random differences between paired ordinal categorical data (dissertation). Stockholm: Almqvist & Wiksell International 1993.
80. Conover W.J. Practical Nonparametric Statistics, 2ed. John Wiley & Sons 1980.
81. Koran LM. The reliability of clinical methods, data and judgements (first part). N Eng J Med 1975;25(13):642.
82. Koran LM. The reliability of clinical methods, data and judgements (second part). N Eng J Med 1975;293(14):695-701.
83. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992;304(6840):1491-4.
84. Bruns DE. Laboratory-related Outcomes in Healthcare. Clin Chem 2001;47:1547-52.
85. Young E, Stoneham MD, Petruckevitch A, Barton J, Rona R: A population study of food intolerance. Lancet 1994;343(8906):1127-1130.
86. Sampson HA. Food Allergy. Part 2: diagnosis and management. J Allergy Clin Immunol 1999;103:981-9.
87. Parker SL, Kroondl M, Coleman P. Foods perceived by adults as causing adverse reactions. J Am Diet Assoc 1993;93:40-4.
88. Bousquet J, Metcalfe D, Warner J. Food allergy. Report of the Codex Alimentarus. ACI Int 1997;9:10-21.
89. Bischoff SC, Herrman A, Manns MP: Prevalence of adverse reactions to food in patients with gastrointestinal disease. Allergy 1996;51:811-818.
90. Bindslev-Jensen C, Ballmer-Weber BK, Bengtsson U, Blanco C, Ebner C, Hourihane J, et al. Standardization of food challenges in patients with immediate reactions to foods – position paper from the European Academy of Allergology and Clinical immunology. Allergy 2004;59:690-97.
91. Bischoff SC, Herrmann A, Mayer J, Manns MP. Food allergy in patients with gastrointestinal disease. Highlights in Food Allergy 1996;32:130-142.

92. Ortolani C, Bruijnzeel-Koomen C, Bengtsson U, Bindslev-Jensen C, Bjorksten B, Host A et al. Position paper. Controversial aspects of adverse reactions to food. Allergy 1999; 54:27-45.
93. Bindslev-Jensen C. Standardisation of double-blind, placebo-controlled, food challenges. Allergy 2001;56 suppl 67:75-7.
94. Bindslev-Jensen C. Food allergy: a diagnostic challenge. Curr Probl Dermatol 1999;28:74-80.
95. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299(17):926-30.
96. Begg CB. Biases in the assessment of diagnostic tests. Statistics Med. 1987;6(4):411-23.
97. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39(1):207-15.
98. Knottnerus JA, Leffers JP. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epid 1992;45(10):1143-54.
99. Diamond GA. Selection bias and the evaluation of diagnostic tests: a metadissent. J Chron Dis 1986;39:359-60.
100. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med. 1992;117:135-140.
101. Mulherin SA, Miller CW. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med. 2002;137(7):598-602.
102. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epid 1996;49:1373-1379.
103. Hopstaken RM, Stobberingh EE, Knottnerus JA, Muris JWM, Nelemans P, Rinkens PELM, et al. Clinical items not helpful in differentiating viral from bacterial lower respiratory tract infections in general practice. J Clin Epid 2005;58:175-183.
104. van Klei VA, Kalkman CJ, Tolsma M, Rutten CLG, Moons KGM. Pre-operative detection of valvular heart disease by anaesthetists. Anaesthesia 2006;61:127-132.
105. Herlitz J., Bång A., Isaksson L., Karlsson T. Outcome for patients who called for an ambulance for chest pain in relation to the dispatcher´s initial suspicion of acute myocardial infarction. Eur. J. of Emergency Medicine, 1995;2: 75-82.

106. Herlitz J., Karlsson B.W., Richter A., Liljeqvist J-Å., Strömbom U. and Holmberg S. (1992b) Early identification of acute myocardial and prognosis in relation to mode of transport. Am. J. Emergency Medicine, 1992;10(5): 406-412.
107. Jørgensen LGM, Hyltoft Petersen P, Brandslund I. The impact of variability in the risk of disease exemplified by diagnosing diabetes mellitus based on ADA and WHO criteria as gold standard. Int J Risk Assesment and Manegement 2005;5:358-373.
108. Collins R, Peto R, Armitage J. The MRC/BHF Heart Protection Study: preliminary results. Int J Clin Pract. 2002 Jan-Feb;56(1):53-6.