Review

Rainer Haeckel* and Werner Wosniok, Thomas Streichert, Members of the Section Guide Limits of the DGKL

Review of potentials and limitations of indirect approaches for estimating reference limits/intervals of quantitative procedures in laboratory medicine

<https://doi.org/10.1515/labmed-2020-0131>

Received November 4, 2020; accepted March 1, 2021;
published online March 29, 2021

Abstract: Reference intervals (RIs) can be determined by direct and indirect procedures. Both approaches identify a reference population from which the RIs are defined. The crucial difference between direct and indirect methods is that direct methods select particular individuals after individual anamnesis and medical examination have confirmed the absence of pathological conditions. These individuals form a reference subpopulation. Indirect methods select a reference subpopulation in which the individuals are not identified. They isolate a reference population from a mixed population of patients with pathological and non-pathological conditions by statistical reasoning.

At present, the direct procedure internationally recommended is the “gold standard”. It has, however, the disadvantage of high expenses which cannot easily be afforded by most medical laboratories. Therefore, laboratories adopt RIs established by direct methods from external sources requiring a high responsibility for transference problems which are usually neglected by most laboratories. These difficulties can be overcome by indirect procedures which can easily be performed by most laboratories without causing economic problems.

The present review focuses on indirect approaches. Various procedures are presented with their benefits and limitations. Preliminary simulation studies indicate that more recently developed concepts are superior to older approaches.

Keywords: data mining; indirect reference limits; reference intervals.

Introduction

Laboratory results released to requesters require interpretative support usually established by reference limits (RLs), which must not be confused with other interpretative guide limits [1–3]. Contrary to reference intervals/limits (unimodal consideration), clinical decision limits consider both the distribution of measurement values from “healthy” persons and patients (multimodal consideration). Because clinical decision limits are designed for particular disease states, too many limits would be required. “This stand-off is probably why reference limits, despite their difficulties, have remained popular (i.e. because they avoid the plethora of disease definitions by concentrating on the apparently simpler problem of health)” [4].

Critical values (also called “panic values”) are not classified as guide limits. Critical values trigger accelerated transmission of diagnostic findings. These limits are individually negotiated between clinicians and laboratories and are therefore extremely variable [3]. Forensic limits (as e.g. for ethanol to assess the fitness to drive) were also not considered as well as decision limits used in the fields of occupational or environmental medicine [3].

The present review is focused on reference intervals (RIs) which are solely referred to non-diseased (“healthy”, “normal”) subjects with regard to the measurand of interest. The terms “healthy” or “normal” should not be used anymore because health and normality are relative conditions lacking a universal definition [5]. There is often a

*Corresponding author: Rainer Haeckel, Bremer Zentrum für Laboratoriumsmedizin, Klinikum Bremen Mitte, 28305 Bremen, Germany, Phone: +49 412 273446, E-mail: rainer.haeckel@gmx.net, <https://orcid.org/0000-0001-8671-962X>

Werner Wosniok, Institut für Statistik, Universität Bremen, Bremen, Germany

Thomas Streichert, Institut für Klinische Chemie, Uniklinik Köln, Köln, Germany. <https://orcid.org/0000-0002-6588-720X>

gradual transition from “health” to disease. In the present review, the term pathological values are used for values from diseased subjects, and non-pathological values for values from non-diseased subjects. Lower and upper reference limits describe the central 95% reference interval of measured values obtained from samples of a “healthy” reference group (non-pathological with respect of the measurand). As a convention, 2.5% of the values are below and above the reference limits. Other percentages (as e.g. 99% for cardiac troponins) have been used. However, it was suggested to consider such limits as clinical decision limits and restrict the 95% solely for reference intervals [6, 7].

The best reference limits for an individual are derived from her or his own prior data (subject-based reference limits). Because these are often not available, population-based reference limits are widely used [8]. Ideally, reference limits “should be determined on patients with some of the signs or symptoms of the disease being considered but do not in fact have that disease because they form the ‘control group’ from which a patient with the disease must be distinguished. This ideal is difficult to achieve” [9]. Therefore, several approaches have been developed which are usually divided in direct and indirect methods.

The present review is based on several reviews [10–12] and updated according to more recent developments and reports.

Direct vs. indirect methods

Defining reference intervals via direct sampling (Box 1) involves collection of specimens from selected members of the reference population for the purpose of establishing reference limits. Direct selection of reference individuals concurs with the concept of reference values as recommended by IFCC [13, 14]. Their major disadvantages are the problems and costs of obtaining a representative group of reference individuals. These practical problems have led to the search for simpler and less expensive approaches such as indirect methods. This approach is based on the observation that most analysis results produced in the clinical laboratory seem to be from non-diseased subjects. The great majority of pathological results tend to be on a single side of the RIs for most tests.

The crucial difference between direct and indirect methods is that direct methods select particular individuals.

Box 1: Methods for establishing reference limits (modified from ref. [11]).

Direct methods: *a priori* or *a posteriori* selection of reference individuals from a probably non-diseased subpopulation by predetermined criteria. These individuals form a reference subpopulation. *A priori* selection means application of criteria before the collection of the samples, and *a posteriori* selection is the application of criteria after the collection of the samples.

Indirect methods: selection of the results from a mixed population (containing diseased and non-diseased subjects, such as it is found in routine medical laboratory databases) to get the results of a probably non-diseased subpopulation. The selection is performed by statistical tools resolving the distribution in at least two distributions.

These individuals form a reference subpopulation. Indirect methods select a reference subpopulation in which the individuals are not identified.

The present ‘gold standard’ for determining reference limits is the direct approach using a preselected group of non-diseased subjects [13, 14]. Whereas *a priori* selection is the preferred way of the original IFCC recommendation [13], *a posteriori* selection may have advantages [9, 15]. Kairisto et al. [15] argued that patients hospitalized with chest pain but later proved not to have suffered from myocardial infarction would be ideal reference subjects for cardiac biomarkers. The population used for the development of RIs should be as close as possible to the patient group being served, with the exception of the disease being tested for [16].

Analytical and biological preconditions

Before RIs of a particular subpopulation can be established it must be cleared whether the subpopulation is homogeneous or consists of several subgroups which require stratification. Rules for stratification have already been outlined [17, 18].

If reference limits established under various conditions are to be compared with each other, pre-analytical (pre-examination), analytical (examination) and biological

preconditions [19] must be considered. The preconditions become particularly important if RIs are transferred from external sources to intra-laboratory limits (transference problems). Then, the transferability must be carefully checked [17, 20]. Boyd [21] pointed out that transferability remains an elusive goal.

Pre-analytical conditions

Usually, RIs can only be established for a particular specimen, e.g. for arterial, venous or capillary serum, plasma or whole blood. The stability of measurands between time of sampling and time of analysis must be guaranteed. Thus, the prevention of glycolysis may become crucial for establishing glucose RIs. For further details see ref. [22]. Endogenous interferents (e.g. bilirubinaemia, haemoglobinaemia, lipaemic turbidity) may cause exclusion criteria related to serum or plasma samples [23–25].

Analytical procedures (for determining values from reference subjects)

It is obvious that the analytical results must be comparable if reference limits from different analytical procedures are applied. Analytical hindrances include the lack of established reference measurements systems for many quantities, lack of traceability of field methods to the reference systems and lot-to-lot variability in reagents and calibrators [21] as well as possible influences caused by different analysers. In the present review, it is assumed that analytical bias between intra-laboratory RIs can be neglected for the purpose of comparison. Reference limits cannot be compared with each other unless bias between different methods can be considered somehow. Furthermore, it is essential that the analytical procedure is stable during the time period of sample collection. This is especially important, if this time period expands over several years. Automatic programs should include a check for drift effects (an example is given in Fig. 5).

Biological preconditions

Before a RI of a particular subpopulation is established, it must be cleared whether biological variables are known which may influence the RI. If these variables cause relevant differences, the RI must be determined for each biological variable separately.

Biological variables which can influence reference limits either depend on the individual or on the subpopulation. Individual dependent variables are age, sex, body mass index, alcohol and drug consumption, cigarette smoking, exercise, food preferences, menstrual cycle, ABO blood groups, etc. Subpopulation-dependent variables are ethnicity, regional and seasonal effects. Ethnical differences must be considered if global reference limits for all parts of the world are recommended. Thus, Turks have distinctly low concentrations of HDL-cholesterol associated with elevated hepatic lipase activity and high concentrations of fasting triglycerides [26]. Some of these factors are of physiological nature starting at birth and include weaning, the active toddler, puberty, pregnancy, menopause and aging [10]. “Physiology describes our life’s journey, and is only when we are familiar with that journey that we can appreciate a pathological departure” [10].

The most common variables are age and sex differences. Many measurands are age dependent and direct RIs are often determined in an age range between 18 and 60 years with a majority around 40 years, and indirect RI in an age range between 18 and 90 with a majority around 60 years. Newer statistical tools of approaches for the indirect determination of RIs have implemented the automatic stratification for age and sex [27].

Continuous presentation of RIs over age to avoid “jumps” between discrete age groups have been suggested [7, 28, 29]. An example is shown in Figure 1. The RIs for a specific year of age in Figure 1 can be taken from the table of RIs which was used to produce Figure 1. A specific case has been reported by Palm et al. [30] for NT pro BNP concentrations of children where a double logarithmic treatment led to a straight line from which back-transformation to a desired age can be easily performed.

An extensive list of various biological variables was recently published by Özcürümez and Haeckel [19] for adults (above 18 years). Furthermore, regional effects have been described: e.g. thrombocytes in one local area [32], and several examples of differences between countries [33].

Less considered are diurnal variations, such as circadian rhythms. An example is shown in Figure 2. Leucocytes [34] are lower at the early morning, and higher during the afternoon with a mean difference of about 1×10^9 g/L, erythrocytes and haemoglobin are higher during the morning than during the afternoon [35]. Further examples are listed in ref. [36–38]. Measurands with circadian rhythms require careful selection of reference subjects according to defined time windows, both with direct and indirect procedures. If the time frame was not defined with a direct procedure, it may be assumed that the samples were taken between 7 and 10 am. Reference limits determined

Figure 1: Age dependent reference limits of the catalytic activity concentration of alanine aminotransferase (ALT, U/L) determined with the truncated minimum chi-square (TMC) method in a large data set ($n=177,912$) from outpatient provided by a university hospital. The RLs calculated in 10 years intervals are connected by a cubic smoothing spline, the lower line for lower RLs (2.5th percentile) and the upper line for upper RLs (97.5th percentile). The green symbols (crosses) and lines represent men, the red symbols (circles) and lines represent women with vertical 95% confidence limits (taken from ref. [31]).

Figure 2: Diurnal variation of the uric acid concentration in heparinised plasma (ambulant patients, $n=80,099$). Daily observations are calculated during one week and the medians from Monday to Friday are presented. The yellow areas represent the approximate 95% confidence limits based on the median absolute deviation (taken from ref. [27]).

under these conditions can only be used for comparison with indirectly estimated reference limits, if reference samples are also taken between 7 and 10 am.

A well-known variable influencing reference limits of corpuscular blood components and measurands which are protein bound is the posture [19]. Direct sampling usually is performed of a well-defined subpopulation of “healthy” subjects, relatively young hospital employees, being fully active in their professional life, walking around and only sitting for sample taking, but hardly comparable to a clinical subpopulation. Many hospitalised patients may be in a horizontal position when blood samples are taken.

RLs indirectly established of hospitalised patients (inpatients, secondary or tertiary services) may differ from those determined of outpatients (ambulant patients) [7, 39, 40]. Thus, for uric acid [27] and cardiac troponin [6], RLs of ambulant and hospitalised patients are similar up to an age of 49 years, but above 50 years, hospitalised patients have higher upper RLs than ambulant patients. The reason for this difference has been suggested to be due to a decreased glomerular filtration rate with older hospitalised patients. Outpatients may be encountered either in hospitals or in private laboratories primarily serving practitioners (primary health care service). In outpatients, the great majority of results can be expected to stem from non-pathological subjects [39]. Several authors [8, 15, 40, 41] have proposed to derive “health-associated” RLs rather from primary health care services than from hospitalised patients because they are better comparable with RLs determined according to IFCC recommendations [13, 14] from a selected subpopulation of young and “healthy” subjects. However, it can be questioned whether young and “healthy” subjects may be an adequate subpopulation to derive RLs for diseased people [15]. Separate reference intervals may be collected for ambulatory and hospitalised patients because of the postural differences (sitting or recumbent position) [19].

Most indirect approaches exclude subjects to reduce the prevalence of pathological values e.g. by excluding patients from intensity care units, and gynaecology [25]. Özarda and Aslan furthermore excluded patients from oncology, endocrinology, hepatology and nephrology units [26]. To avoid a possible bias from repeated measurements (binding effect), many authors included only the first available test result for each patient, thus forming a database of test results “on admission” [23, 36, 39–41] assuming that persons with repeated measurements have a higher chance of being diseased. Some authors prefer the last result assuming that the patient has recovered from his disease or “solo” samples assuming that the physician did not request a repeat testing because the result was more or less non-pathological.

If biological variables cause different RIs of subgroups, the requirement for stratification depends on the medical relevance of the observed differences between different RIs. Several tests have been proposed for this purpose:

- Harris and Boyd [42] recommended to separate RIs when the ratio of the standard deviation (larger over smaller) between the subgroups exceeds 1.5, or when the *z*-statistics between the two subgroup distributions exceeds 3.
- Lahti et al. [43] proposed partitioning when more than 4.1% of a subgroup falls outside the RIs.
- Equivalence test: see below.

The value of stratification becomes evident if one considers the index of individuality (II) [12, 44]. II is defined as the intra-individual variability/inter-individual variability. If II is below 0.6, RIs lose their utility [44]. Thus, stratification can increase the index of individuality to values which become diagnostic usefulness [12] as exemplified by Fraser [45]. Because of physiological, respective biological factors influencing RIs, each estimation of RIs must be carefully scrutinized. A procedural example is given below.

Statistical procedures for determining reference limits and their assumptions

General considerations

The statistical procedures for direct models are well established and their limitations are known. However, with indirect models the procedures for establishing reference limits are still open for debate. Indirect approaches need further validation, yet, and their clinical relevance is usually judged by comparison with direct approaches. For direct and indirect approaches, biological preconditions as mentioned in the previous section are often neglected but must be considered. But there are many other aspects that must further be taken into account as outlined in the subsequent sections.

The distribution of values from non-diseased subjects

The only situation in which no assumption on the distribution of values from non-diseased subjects must be made is the direct RI estimation using non-parametric quantile estimation. In all other cases, at least an assumption about the distribution of values from non-diseased subjects is needed.

A widespread assumption is that laboratory data follows a normal (Gaussian) distribution (ND) [13, 14].

Sometimes, the central limit theorem is used to justify this assumption. This theorem is certainly extremely important in statistics, but it does not universally say that data have a normal distribution. It states (in simple words) that a sum of infinitely many stochastically independent random variables, with none of them dominating the others, has a normal distribution. These four conditions have never been shown to hold for laboratory data. One can imagine many other processes that might generate laboratory data and which would lead to other statistical distributions for observed data. Johnson et al. [46] described many examples of processes generating random data and the resulting distributions. There are two simple reasons why a ND can only be an approximation to the distribution of laboratory data. The first reason is that laboratory data never have negative values, while the ND always assigns a probability >0 to the occurrence of negative values. This probability may be small if the standard deviation of the data is small compared to its mean value, which means that the ND may be an acceptable approximation in these cases. The second reason is that a ND is symmetric, while laboratory data, again for the reason of taking only non-negative values, has no symmetric distribution. In order to be symmetric, a distribution on non-negative numbers must either be bounded above (has a maximum achievable value) or has an infinite mean. Both conditions are not realistic for laboratory data. Another reason for the asymmetry of the laboratory data is the presence of the analytical error, which usually increases with the value itself.

The problems resulting from the assumption of normally distributed data is frequently acknowledged in the laboratory community by the recommendation not to consider the data itself, but a transformation of it as normally distributed. Typical transformations are the Box–Cox and the Manly transformation. Both need a transformation parameter that has to be specified by the user or estimated from the data. If such specification seems not feasible, and in the absence of better information, a simple general approach is to assume a logarithmic normal distribution (LND) for non-pathological laboratory data [47]. This avoids the need of dealing with positive probabilities for negative values and has the convenient feature that after taking logarithms the data has a ND, for which many statistical procedures are readily available. The LND is always skewed, though the degree of skewness may be small, which seems to have caused some irritation. As an example, the distribution of plasma sodium may, for practical purposes, be described equally well by a ND or a LND [46, 47]. Here, the difference between both distributions is so small that statistical tests may not be able to distinguish between them. Due to random fluctuations, the ND might even show a

better fit than a LND. However, such similarity of ND and LND is only a numerical coincidence. It does not prove the general validity of the normal distribution assumption, because the principal arguments against a ND are still valid, and plausibility argues in favour of the LND. Beyond the approximation aspect there is no point in considering a ND as distribution of laboratory data.

If a LND is found an inappropriate description of data from non-diseased subjects, another skewed distribution should be considered. Such a distribution must always be asymmetric, for the reasons given above. In order to extend the set of candidate distributions for laboratory data, the Box–Cox transformation of a ND may be used. By varying its transformation parameter λ , it offers a large number of distribution shapes including ND and LND. All transformations with λ not equal to one produce skewed distributions. Other candidate distributions that have been employed as assumptions for data from non-diseased persons are the gamma distribution and distributions described by Gram–Charlier series (see Box 2). In the course of indirect estimation, the basic assumption made about the data from non-diseased persons should be checked.

Assumptions on the distribution of data from diseased persons are usually not required. However, indirect methods using general mixture decomposition techniques do require specifying the type of all distribution components in the data. This means that the user has to specify distribution types also for the pathological components in the data, though there is no interest in analysing these. Typically, the same distribution types are used for the non-pathological and the pathological components of the data, though this is not a necessary feature of mixture decomposition methods. Methods based on truncated estimation (truncated minimum chi-square, TMC and truncated maximum likelihood, TML) need no assumptions about the distribution of pathological data.

Sample size

The accuracy of any RL estimation is limited by the sample size. Lower bounds for this accuracy follow from the asymptotic distribution of quantiles [48]. Figure 3 shows an example for these bounds. Statistical procedures requiring the estimation of distribution parameters from the data enlarge the confidence intervals shown there. Imprecision in the data, like operating with rounded data, has the same disadvantageous effect.

Figure 3: Confidence limits (90%) of the upper reference limit in relation to the number of observations in the case of a log-normal distribution of thyrotropin values.

The broken red line indicates at $n=120$ (taken from ref. [53]).

Rounding

Rounding reduces the information provided by the data. It changes the character of the data from exact (continuous) representation to an interval representation. Considering the plasma sodium example again: a typical reported value of 140 mmol/L does not mean the value 140.000 ..., but instead “somewhere in the interval [139.5, 140.5)”, where the left limit is included in the interval and the right one not. Such an interval obviously provides much less information than an exact number. This loss due to rounded reporting is not incorporated in the accuracy calculation of the previous section, accuracy becomes worse by rounding. For indirect methods, rounding may be harmful in another way, because they inspect the empirical data distribution for deviations from the assumed distribution of values from non-diseased subjects. A deviation due to rounding may generate wrong conclusions, particularly if the rounding strategy generates jumps in the distribution. This is the case for the “rounding to the even digit” strategy [49]. Strong rounding, which means that only few distinct values are left (e.g. creatinine in conventional units with only one digit, see ref. [24]), is similarly disadvantageous for indirect methods inspecting the distribution shape.

Direct procedures for determining reference limits

The procedure recommended by IFCC [13] and CLSI [14] is well established and has become the present “gold standard”. Reference limits are calculated from values determined from at least 120 reference subjects for parametric and 200 for non-parametric interval determinations [14]. Further details have been reviewed recently [41]. In cases where it may be difficult to get enough volunteers, robust approaches have been described for smaller sample sizes [50].

With the IFCC concept, it is mandatory to choose individuals who are as “healthy” as possible. So far, no well accepted definition of the term health is available. For each study, the “health” criteria must be established [13, 14, 51]. Different countries may have their own standards of “healthiness” and criteria for secondary exclusion criteria may differ from country to country [51, 52]. LAVE is an iterative optimization method for refining reference individuals by excluding subjects possessing abnormal values in related analytes [18].

The conventional procedure is to compare patients with “healthy” individuals. As outlined above, the health associated values are usually determined from a small series of selected subjects, the criteria for “healthy” being often subjective and arbitrary. In many cases the subjects are not selected randomly if they are “healthy” subjects of hospital employee’s or of blood banks. Further critical comments have been published recently [53].

Indirect approaches for determining reference limits (data mining)

Several models have been developed to estimate indirect reference limits via data mining. Data mining is the process of using previously generated data to identify new information. Data mining cannot only be used for setting RIs but also for many other purposes, since a treasure trove of information is hidden in medical laboratory data pools [11].

General understanding of the indirect estimation problem

There are three general understandings of indirect estimation. The first understanding considers available data as a combination of data from non-diseased subjects plus some data from diseased subjects lying essentially outside the interval containing data from non-diseased persons. Therefore, indirect estimation is considered as an outlier removal problem. Originally, outliers are extremely small

or large values which lie outside the distribution of interest (the distribution of values from non-diseased patients in the case of indirect estimation) as a consequence of gross measurement errors. Outliers are few in number and can under some assumptions be removed from the outer ends of the data by setting some thresholds and removing values that exceed these. The inner part is considered the complete distribution of interest. The loss of a few data points is assumed to have no relevant effect on parameters estimated from the data left. Therefore, usual methods to estimate means, standard deviations and quantiles are applied to the outlier-free dataset.

The second understanding of indirect estimation assumes that in routine laboratory data the distribution of values from non-diseased persons overlaps the distribution of values from diseased persons to some extent. This seems more realistic, because routine data contains values from non-diseased persons, from persons with fully developed diseases and also from persons still developing a disease or recovering from it. Some values of the latter persons will lie within the margins of the interval containing values from non-diseased subjects. The assumption of laboratory data being a mixture is supported by the fact that routine data rarely exhibits a distribution consisting of two or three isolated sub-distributions that could be identified as the distributions of values from diseased and non-diseased persons, respectively. Therefore, indirect estimation is considered as a problem of estimating RIs from a mixture of distributions. This problem cannot satisfactorily be tackled by outlier removal methods, because in a genuine mixture there is no way of setting thresholds which separate the distribution of values from non-diseased subjects from the distribution(s) of values from diseased persons. Each separated presumed value distribution of non-diseased individuals will either contain also values from diseased persons, or the separated distribution is only a subset of all values from non-diseased persons. This means that when considering routine laboratory data as partial overlap of data from diseased and non-diseased persons, there is only a subinterval of all data which contains exclusively data from non-diseased persons.

Indirect methods that use truncated estimation (like TML and TMC below) require such an interval. They use this interval for RI estimation and also for an internal check of the distributional assumption made for non-diseased data. This subinterval must be large enough to allow reliable parameter estimation [54]. Absence of such an interval, which would be detected by the distribution check mentioned, would make RI estimation impossible. Parameter estimation from only a subinterval of all data requires methods that are tailored for this purpose, as well as the detection of this subinterval, the truncation interval (TI).

The third understanding of indirect methods assumes that routine laboratory data are a general mixture (a weighted sum) of distributions, where one of these represents the non-pathological data, while the others represent different subpopulations of pathological data. The corresponding mixture decomposition techniques do not require the existence of an interval that contains values only from non-diseased persons. Mixture deconvolution means to describe the complete dataset by a sum of weighted basis distributions as good as possible. There is no need for truncated estimation, because each mixture component may in principle contribute to the occurrence probability at each point in the data range. Consequently there is no chance of checking distributional assumptions regarding the distribution of non-pathological data.

Principles of indirect RL estimation

Early approaches of indirect RL estimation have been developed before the widespread availability of computers. These approaches had to be executable with paper, pencil and tables of logarithms and statistical distributions. Therefore, some simplifications were made that are no more necessary today. The basic components of these early methods are still visible.

Most of the early approaches transform the distribution of the empirical data such that the majority of the transformed data appears as a straight line if the raw data comes essentially from a ND. Slope and intercept of this line lead to mean and standard deviation of the assumed normal distribution. Systematic deviation from a straight line indicates the presence of additional distributions (or a wrong distribution assumption). The location of the straight line and the selection of one line for RL estimation, if there are several, are done by visual inspection.

Many approaches which follow the first understanding of the indirect estimation problem start with the elimination of outliers. Data points that are not consistent with the assumption for non-pathological data are removed. One of the most frequently used outlier detection methods is the Tukey method as described by Horn et al. [55]. Solberg and Lahti found the Tukey method to be relatively insensitive for the detection of outliers [56]. Farrell et al. called exclusion of outliers as one of several “pre-processing steps to help ‘clean up’ the data” [12]. Another outlier detection method, employed by Katayev [57, 58] consists of applying Chauvenet’s criterion. All methods use assumptions about the character of the outlier-free data. In some cases, only symmetry is assumed, other approaches use a strong assumption like a ND. Outlier detection methods remove without being able to check their validity. This may generate unwanted results like extracting a

nearly normally distributed subset from data that is not normally distributed. So far, no generally accepted recommendation for detecting and eliminating outliers is available.

Whereas the statistics required for the determination of RIs by direct methods are relatively simple, indirect methods require more complicated statistical procedures. In approaches which use a transformation to a ND (e.g. Hoffmann [59], Bhattacharya [60]), the calculation of the RI is trivial after the parameters are determined: $\text{mean} \pm 1.96 \text{ SD}$.

Some indirect approaches are listed in Box 2. In the following, we focus on actually more frequently applied approaches.

Box 2: Indirect methods to establish reference limits.

1. Pryce [61], Becktel [62], Kairisto et al. [63]: resolution in two different distributions with one common mode
2. Hoffmann: Probit model [59], modified by Neumann [64], Tsay et al. [65] and Katayev et al. [57, 58]
3. Hoffmann et al. [66, 67]: QQ-plot
4. Bhattacharya [60]: resolution in normally distributed subgroups, modified by Baadenhuijsen et al. including Box–Cox transformation [68] and by Naus et al. [69]
5. Martin et al. [70]: Gram–Charlier series
6. Concordet et al. [71] and Benaglia et al. [72]: general mixture models not using truncated estimation
7. TML (truncated maximum likelihood) [25], modified by Zierk et al. [73]: mixture model for continuous data using truncated estimation without specification of the pathological distribution type
8. TMC (truncated minimum chi-square) [31]: mixture model for discretized data using truncated estimation without specification of the pathological distribution type

Hoffmann model

An early example of a graphical approach for indirect estimation is given by Hoffmann [59], who proposed to display the empirically distribution function of data vs. the ordered data value on probability paper. Probability paper has values of the inverse standard normal cumulative density function on the vertical axis. This gives, for data from a single ND, points randomly scattered around a straight line, its slope and intercept provide mean and standard deviation of the underlying ND.

The common problem of the Hoffmann procedure and its variations described below is that besides neglecting heteroscedasticity and dependency of points they start from a wrong assumption: if the data is a mixture of two distributions, at least one of them a ND, then the points belonging to the ND do not lie on a straight line in a Hoffmann plot. This can easily be seen if the expected positions of a mixture of two normal distributions are plotted on probability paper. The theoretical reason is that probability paper straightens a single cumulative normal distribution function $F(x)$, but the mixture $p_1F(x) + p_2G(x)$ of a cumulative normal and another distribution function $G(x)$ is not a cumulative normal distribution.

Katayev et al. [57, 58] claim to have adopted the Hoffmann approach. In fact, they did not, but used a plot of the sorted data (the empirical quantiles) on the vertical axis against the cumulative standard normal distribution function on the horizontal axis. Different from Hoffmann [59] they do not use a transformation of the ND distribution function. Katayev et al. [57] fit a regression line to the seemingly straight part of the relation, where they determine the straight part by an outlier detection method. However, the seemingly straight part is not even straight if the data consists of only a single normal distribution. The curve has the curvature of a normal probability distribution function, which is nowhere zero. Also, their way of determining a “linear” portion of the data implies a systematic bias towards a too large slope, as can be seen from their Figure 2 in ref. [57]. Holmes and Buhr [74] discuss the numerical size of errors that result from the application of the Katayev approach.

Newer modifications replacing the visual estimation by programmable statistical concepts have been developed, which however, did not solve the above mentioned difficulties.

Hoffmann et al. [66, 67] use a transformation approach related to Hoffmann [59]. Instead of plotting observed quantiles against the transformed cumulated probability they plot observed quantiles (vertical axis) against expected positions of a standard ND (horizontal axis). This plot is known as QQ-plot. This approach also provides approximate solutions for the reasons outlined for the Hoffmann approach.

Bhattacharya model

The Bhattacharya model [60] does not transform the axes of a plot in order to transform normally distributed components of a mixture into straight lines, but transforms the bins of a histogram presentation of the data. If f_i is the percentage in bin i of an equidistant histogram of the data, then $\Delta \log(f_i) = \log(f_{i+1}) - \log(f_i)$ on the vertical axis is plotted

against the bin midpoints on the horizontal axis. In this presentation, a single ND appears as points scattering around a straight line with negative slope, and the parameters of the ND follow from slope and intercept of the line. The approach was introduced as a simple graphical method in which dependency and heteroscedasticity of the points are ignored.

Mixture deconvolution

In some sense, the Bhattacharya approach is a mixture deconvolution approach, because it assumes the data to be a mixture of NDs. However, in the Bhattacharya approach the user might decide to not identify the parameters of each component, but only the one that is suspected to represent the distribution of non-pathological data. General mixture deconvolution, however, requires the determination of all components.

The approach of Concordet [69] is a mixture deconvolution approach generalizing the Bhattacharya approach by considering a mixture of two Box–Cox transformed NDs, one for the non-pathological component, the other for the pathological component. Pathological values are assumed to lie on one side of the non-pathological component. The parameters of all mixture components are automatically estimated by an expectation maximization (EM) algorithm (another generalization of the Bhattacharya approach).

The Concordet approach is limited to the situation of having pathological data only on one side of the non-pathological data, and the pathological data must be describable by a Box–Cox transformed ND. These limitations are relaxed by general mixture deconvolution methods, which allow an arbitrary large number of components to describe the total data.

In general mixture deconvolution approaches, the distribution type of each component as well as the number of components must be specified by the user. This means that even if the distribution of pathological values is usually not of interest, effort is needed to model it.

Available R packages (e.g. flexmix, mixmod, mixdist, mixtools [72]) offer a large range of deconvolution methods together with a large choice of component distribution types beyond ND and LND, including even non-parametric distributions. Some packages also provide a suggestion for the number of components obtained by a bootstrapping technique. Weights and parameters of the component distributions are chosen by a numerical method, which minimizes the distance between the total data distribution and the mixture. Examples for the application of mixture deconvolution for indirect RI determination have been given by Holmes and Buhr [74].

Interpreting the result of a mixture deconvolution can be complicated. In simple cases, the largest component of the mixture can be interpreted as the component describing the non-pathological data. However, deconvolution is done without the requirement that there is an interval in the data range which contains only non-pathological data. Therefore the decomposition result does not indicate which the non-pathological component is, and no test can be made whether the assumption about the non-pathological distribution type is adequate. If this assumption was wrong, the deconvolution will provide a result in which non-pathological data are described not by a single component, but by a sum of components and the user has to figure out which of the components describe the non-pathological data.

TML (truncated maximum likelihood) model

This method is based on the maximum likelihood estimation of the parameters of a power normal distribution for a truncated data set. This method was developed by Arzideh et al. [23]. It is assumed that the main part of the data consists of values from non-diseased subjects, and, within an unknown interval $[T_1, T_2]$, all values are from non-diseased persons, and the number of values from diseased subjects is negligible. The parameters of the power normal distribution (μ , σ , λ) are estimated using the maximum likelihood method for the truncated data. The 2.5th and 97.5th percentiles of the estimated distribution establish the RLs. The optimization algorithm for choice of truncation points are described elsewhere [25].

An example of applying this algorithm on γ -glutamyl transpeptidase data is shown in Figure 4. The applied semi-parametric model on the data ensures not only a parametric estimation of the value distribution from non-diseased subjects, and thereafter RLs, but also a non-parametric smoothed density function for values from diseased subjects. The intersection point(s) of the estimated density function for the values from non-diseased subjects with the estimated density function(s) for values from diseased subjects theoretically provides the decision limit (theoretical decision limit [3]) with the maximum diagnostic efficiency in discriminating “health” and disease [12].

An automatic programme (Reference Limit Estimator, RLE) on the Excel platform is available on the home page of the DGKL [75] which includes e.g. sex stratification and detecting drift effects during the time of data collection (Figure 5). This is especially useful, if the data are collected during several years. Drift effects are also tested for significant deviations (Figure 5).

Figure 4: Estimation of the reference interval (RI) for γ -glutamyl transpeptidase with truncated maximum likelihood, TML (taken from ref. [25]).

In total, 66,789 data from male outpatients were measured with a Roche Cobas 8000. Green and red curves display the estimated distributions for non-pathological and pathological values, correspondingly, and blue curve displays a kernel density function for the whole data. The estimated 97.5-percentile of the green curve is given (75.8 U/L).

Excel suffers from a number of limitations, especially the security settings (the RLE-tool is based on macros/VBA) are a major obstacle since exploitation of malicious macros is one of the top ways that organisations around the world are compromised by today [76]. In consequence the working group of the DGKL developed pure “R”-scripts where the statistical steps are the same as already published [75]. New features are the options to estimate the RIs of a number of

Figure 5: Example of a long-term drift.

Crosses represent calculated monthly medians, the red line is the overall median. The dashed blue line is the fitted smooth curve of monthly medians with their confidence limits (dotted blue lines). Dashed red lines (limit of the grey zone) indicate the calculated permissible uncertainty of the overall median through Equations (3)–(12) in ref. [75]. Differences of monthly medians set in grey zone can be explained by computed measurement uncertainty. The permissible uncertainty is quantified by the permissible analytical standard deviation derived from the empirical biological variation [77]. The Figure is taken from ref. [75].

measurands in a single run (useful e.g. to estimate the limits for a complete blood count with all of its measurands) and to calculate TML-based continuous RIs for age. The “R” apps are available from the authors upon request.

TMC (truncated minimum chi-square) model

The basic assumptions of the TMC approach are similar to those of the TML approach:

- Measured values are statistically independent (only one observation per individual).
- Values from Non-diseased persons follow a PN distribution (PND).
- The data set is assumed to contain a subset $[T_1, T_2]$ that consists essentially values from non-diseased persons. “Essentially” means that values from diseased subjects may exist, but their presence does not cause rejection of a goodness-of-fit test hypothesizing a PND.
- No assumption is made about the value distribution of diseased individuals.

These assumptions express that the data set is considered as a mixture of values produced by at least two distributions, namely the distribution of values from non-diseased patients and the distribution of values from diseased patients. Mixture components may overlap to some extent, but it is assumed that an interval exists which contains only non-pathological data. A difference to general mixture deconvolution is that no distributional assumption is made for the non-pathological data.

TML and TMC do not require isolating the full value distribution from non-diseased persons, as outlier removal based methods do. These weaker but more appropriate assumptions require particular methods for both TML and TMC for estimating RIs from truncated data.

The TMC method was recently described in detail [31]. It treats laboratory data generally as rounded, not continuous data. The data are represented by a histogram, and this histogram is modelled. This allows dealing with data of the form “<DL”, where DL is the detection limit, without the need of replacing this interval by some artificial number. The method first identifies an interval containing essentially non-diseased patients, the truncation interval $[T_1, T_2]$, by fitting a PND to a series of candidate intervals. Each fit is accompanied by a goodness-of-fit test in the truncation interval and several plausibility checks. Goodness of fit and plausibility checks are combined to an assessment criterion. The TI with the best assessment criterion provides the final RI estimate.

Fitting a PND to a TI is performed by an iterative minimum chi-square approach for truncated estimation [54].

Those parameter estimates which minimise the well-known chi-square distance between observed bin frequencies and the frequencies predicted by the PND are the optimal estimates. Only bins in the truncation interval are used for the estimation. As with TML there is no need for transforming data to normality. Components of the estimation process are shown in Figure 6. The asymptotic properties of minimum chi-square estimation are the same as those of maximum likelihood estimation used in TML.

The chi-square approach was chosen because it uses an illustrative optimisation criterion, and is easy to formulate for the present problem of truncated estimation. RIs are calculated directly from the PND using the estimated PND parameters.

A script for performing the TMC analysis, written in the R programming language, can be requested from the authors or is available on the home page of the DGKL [75]. The

Figure 6: Histogram for alanine aminotransferase data in plasma (females, 70–79 years, $n=11,056$).

Grey bins indicate the truncation interval, white bins lie outside the truncation interval. The blue PN distribution (PND) probability density curve is fitted by the truncated minimum chi-square (TMC) approach. Solid red and green rectangles indicate the differences between observed and expected counts which contribute to the χ^2 criterion. Red rectangles indicate bins in which the expected count is larger than the observed. These rectangles contribute to the χ^2 criterion inside and outside the truncation interval. Bins outside the truncation interval with expected count smaller than observed, marked by green hatched rectangles, do not contribute to the χ^2 criterion. The vertical dashed blue lines indicate the 2.5 and 97.5% RILs. Details of the calculation are given in ref. [31]. Observed bin values are white coloured areas + green areas or white coloured areas without red areas.

script calculates a marker for inconsistent rounding in the data. For long-term data sets, it also detects drifts during the time of data sampling. Drift effects are also tested for significant deviations. If the user supplies data on the patients' sex and age together with an age grouping, the analysis is automatically stratified by the resulting sex/age groups. If more than four age groups are defined, a spline function is used to compute a continuous relation between patient age and the RLs. The spline function provides numerical RL predictions for all ages in the age interval covered by the data. Their graphical presentation has no artificial "jumps" between age groups. Typically, 10 years intervals are used for adults. Using five years intervals usually leads to very similar spline functions, but confidence intervals of estimated RLs were slightly larger, as expected. The script also allows stratification e.g. in outpatients (ambulant patients) and hospitalised patients and detection of daytime variation (if the sample collection time or the arrival time of the samples in the laboratory is available). Furthermore, several other features are automatically estimated, as e.g. the prevalence of the non-pathological and pathological data. The upper prevalence (uPrev) is the ratio of all values above the mode ($n_{>\text{mode}}$) minus all estimated non-pathological values above the mode ($n_{\text{non-pathological}>\text{mode}}$) over all values of the particular subpopulation (n_{all} , number of all values):

$$\text{uPrev} = [(n_{>\text{mode}} - n_{\text{non-pathological}>\text{mode}})]/n_{\text{all}}$$

As mentioned above, each estimated RI must be scrutinized as exemplified for the TMC approach in Box 3. The strategy described in Box 3 considers the most important biological variables influencing RLs. Other variables (e.g. obesity, smoking habits, medication etc.) are neglected for practical reasons. First of all, it is difficult for most laboratories to receive this information. Furthermore, too many stratified RLs may confuse the requesters of laboratory test results and, therefore, probably are of no benefit for the diagnostic efficiency of RLs.

Confidence intervals and equivalence limits

Confidence intervals can be used for considering the relevance of the difference between two reference intervals only if both intervals were determined by the same number of reference samples. The determination of confidence intervals and examples are given in ref. [27, 31]. If the numbers differ, equivalence limits have been proposed [77]: the permissible difference pD at the lower reference

limit (IRL) is defined as ($\text{ps}_{\text{A,IRL}}$ = permissible analytical standard deviation at IRL)

$$\text{pD}_1 = \leq \pm \text{ps}_{\text{A,IRL}} \times 1.28$$

and the permissible difference at the upper reference limit (uRL) as

$$\text{pD}_2 = \leq \pm \text{ps}_{\text{A,uRL}} \times 1.28$$

Details of calculating pD are given in ref. [77]. For many measurands, pD is automatically calculated by a script which is gratuitously provided on the home page of the DGKL [75].

Henny et al. [78] pointed out that reference limits should be evaluated according to their confidence limits and stated that "it is generally accepted that the confidence interval for each reference limit be <0.2 times the width of the reference interval concerned" [78]. The confidence interval (CI) strongly depends on the number of contributing reference values and on the distribution pattern [3]. With a log-normal distribution as with all other skewed distributions, a much wider CI for the upper reference limit must be accepted. In Figure 3, the sample of $n=120$, which is the minimum of the IFCC recommendation [14], is highlighted.

In the original IFCC recommendation [13], a 0.90 CI was proposed. However, a 0.95 CI may be more appropriate. In Table 1, 0.90 and 0.95 CIs are calculated for

Table 1: Confidence intervals (CI) depending on the number of observations (n) for thyrotropin, TSH (taken from ref. [53]).

n	Lower CL	Higher CL	Confidence range/reference range (permissible limit <0.20)
Normal distribution, 90%			
120	3.46	4.14	0.21
1,000	3.68	3.92	0.07
5,000	3.75	3.85	0.03
Normal distribution, 95%			
120	3.4	4.2	0.24
1,000	3.66	3.94	0.08
5,000	3.74	3.86	0.04
Log-normal distribution, 90%			
120	3.01	4.58	0.48
1,000	3.52	4.07	0.17
5,000	3.67	3.92	0.08
Log-normal distribution, 95%			
120	2.86	4.73	0.57
1,000	3.47	4.12	0.20
5,000	3.65	3.94	0.09

The reference interval chosen was 0.5–3.8 U/L. The CI was calculated at the upper reference limit with a lower and an upper confidence limit (CL). Reference range (RR) and confidence range (CR) = upper limit – lower limit. Permissible limit of CR/RR=0.20.

Box 3: Assessment of reference limits estimated by the TMC approach (if possible after stratification of ambulant and hospitalised patients after excluding patients from particular wards, e.g. from gynaecology and intensity care units. Primary health care laboratories may use unselected subpopulations).

some examples with different numbers of observations. The ratio confidence range/reference range exceeds 0.2 in all cases if $n=120$. Under the assumption of a normal distribution the ratio is only slightly above 0.2, but with a log-normal distribution, the ratio is unacceptably high. If the number of observations is $\geq 1,000$, the ratio is well below 0.2 even at the 0.95 CI (Table 1).

The direct approach is usually applied in external laboratories and the reference limits are then transferred to individual laboratories. According to IFCC/CLSI recommendations [13, 14], transference shall be examined because

the conditions under which the external RLs were established more or less deviate from the internal conditions concerning pre-analytical conditions, analytical procedures and population characteristics. Therefore, bias and imprecision components should be added to the above mentioned confidence limits. However, these components to be added are unknown and the authors are not aware that any corresponding study has been performed. If a theoretical bias of +5% (or +10%) is assumed [79], the average rate of false positive results increases by about 6.2% (12.8%) for 70 measurands listed in the RiliBÄK [80]. This effect varies

between +1% (e.g. C-reactive antigen) and +50% for (plasma sodium for a +5% bias). The calculation of false positive results has been shown in ref. [81].

Verification of reference intervals established by indirect methods

Whereas analytical procedures are well standardized and quality assessed, the establishing of reference limits is less standardized. The key question is, “Is this reference interval suited for my collection process, my method, and my population?” [20]. Transference and verification of reference intervals are required.

Indirect methods assume that the distribution of values from non-pathological subjects is correctly derived from a mixed subpopulation. As an internal check, the TMC method always checks if the data in the truncation interval follows the assumed PND. A warning message is issued if this is not the case. The overall properties of an indirect method can most appropriately be tested by simulation studies with mixtures of artificial data sets which mimic the most common situations occurring in laboratory medicine [28, 69]. Another approach to test the plausibility of reference limits established indirectly is the comparison with established reference limits determined with direct methods. For this purpose, population characteristics must be considered (see under transference problems, above). Also, the numerical quality of the direct method must be taken into account. If the direct RLs were obtained from a sample of small size like $n=120$, the random fluctuation in the result is quite large, as can be seen from Figure 3.

So far, most authors have verified the indirect method applied solely by comparing the obtained RLs with those limits determined by the “gold standard”, that means either with the RLs presently applied in their own laboratories or taken from other literature sources. If the reference limits from both methods agreed, this was considered as a satisfying validation [23, 24, 82]. In many cases, however, the limits from both methods disagreed more or less. The reason for this discrepancy was often left open. Another short check was suggested by CLSI [14]. The RI is considered verified if two or less results out of 20 fall outside the RI estimated by the indirect method (95% probability). The 20 results are obtained from “healthy” subjects without the predefined condition in the reference population. This test can only provide a preliminary guess. Bolann recommended another simple tool to verify established RLs [39] by plotting the patients’ results on normal probability paper. However, this tool is only useful for a limited number of measurands.

The reasons for discrepancies between RLs can be divided in two groups: one reason may be that the direct RI chosen for comparison were taken from external sources without considering the transference problems, that means without having verified the external RL for their internal use. Another group explaining discrepancies is the fact, that biological influence factors are neglected (see Chapter 3.3). These two groups can make comparisons between direct and indirect approaches as a verification tool obsolete. Therefore, simulation studies could provide more objective data [70, 74].

Common reference intervals

If the number of reference samples is reduced critically, it may become useful that several laboratories cooperate with each other. This situation is particular important at both ends of lifetime (that means for children and elderly people). Common RLs are usually derived of multicentre (collaborative) studies. Several examples have been reported [29, 83–85].

Several criteria must be cleared before the data of several laboratories can be pooled, especially for combining the RLs from several laboratories. The prerequisites have been extensively discussed by Ceriotti [86]. Cautions in the adoption of common reference intervals has been reported [21]. Data can be combined if they stem from the same analytical platform, have the same pre-analytical conditions and similar subpopulations. The results of the various laboratories participating in the multicentre study should be based either on reference method values [86] or on strict common requirements for quality assessment [85]. It must be decided (preferably *a priori*), what difference of the RLs from an individual laboratory and the total RL after the combination of the data from several laboratories can be tolerated.

Overlapping confidence intervals of two RLs indicate that from a statistical point of view these RLs cannot be distinguished. However, confidence intervals from large data sets get extremely small so that small differences are detected which would prevent joining the RLs but which are not relevant in a clinical sense. Therefore, the consideration of confidence limits must be complemented by a concept of permissible imprecision that allows combining different RLs if their difference is clinically irrelevant. For this purpose, the concept of equivalence limits [77] can be used. In the case that RLs of more than two sources are to be combined, the difference between the mean or median of all RLs and the RLs of each single source may be checked by the equivalence

limits. If the RLs of several sources are established for several age classes, the relation between age and RLs can be described by a continuous function (e.g. a spline function), specifically for each source. For the mean function, obtained as average of the individual functions per age over all sources, equivalence limits can be calculated, giving an equivalence band around the mean age vs. RL relation. If the individual RLs for each source laboratory are within the equivalence band, common RLs are justified.

Zierk et al. [85] have established the *PEDREF* study (Next-generation paediatric reference intervals, <https://www.pedref.org/>), a network of paediatric tertiary care centres and laboratory service providers across Germany. The goal is the creation of high-quality paediatric reference intervals using a data-mining approach with accurate representation of paediatric dynamics, which requires a multi-centre approach to overcome restrictions due to a limited number of paediatric samples of all ages in single-centre analyses. Standardization of measurement methods in the participating centres has allowed the creation of common reference intervals after data-driven verification of transferability of test results between laboratories. In a multi-centre pilot study, a dataset of >350.000 paediatric alkaline phosphatase samples from seven centres was analysed, resulting in reference intervals represented using percentile charts with unprecedented accuracy and age-resolution [29]. These results have been accompanied by an editorial highlighting the importance of novel approaches to paediatric reference intervals [84]. Currently, 15 German centres are participating in the *PEDREF* study, and have provided pseudonymized laboratory test results, resulting in a comprehensive dataset containing >20,000,000 data points from >1,000,000 German children. Based on this dataset, Zierk et al. established haematology reference intervals, improving age-resolution and accuracy in comparison to previously available reference intervals [85], and reference intervals for other laboratory tests are currently being prepared.

General conclusions for the application of indirect approaches

Benefits

The benefits of indirect approaches mainly in comparison with direct procedures as indicated above are summarized in Box 4. The major advantages are their lower expenses, their lack of transference problems and their applicability

at any time and for convenient stratification strategies. Although the use of indirect methods for determining RIs have been criticised, they have been proposed as valuable tools for quality assessment and for verifying established RIs [17, 81, 87]. Farrell and Nuyen [12] pointed out that indirect approaches may be valuable if a significant proportion of the general population requires exclusion (e.g. parathormone).

Box 4: Benefits and disadvantages of direct vs. indirect methods for establishing reference limits (modified according to Jones et al. [11]).

	Direct methods	Indirect methods
Defining “health” status	May be difficult	Not required
Pre-analytical and analytical conditions	May not match routine conditions	Match routine conditions
Complexity of statistics	Low	High
Transference problems	Relevant	Not relevant
Ethical problems	Must be considered	Not relevant
Expenses (dependent on number of samples)	Considerable	Negligible
Stratifications (e.g. for age)	Difficult (to get enough subjects)	Easy
Confidence limits	Usually broader	Usually smaller
Subpopulation	Usually “healthy” and younger than patients’ groups	Closer to patients’ situation

Limitations of indirect methods

Limitations of indirect approaches may apply when one of the following is present: a large prevalence of results from hospitalized patients, a limited number of observations (especially for subject groups like paediatric or geriatric patients) or rare sample types like synovial fluid, and lack of standardization between the methods in use. This limitation can be minimized by linking laboratories that are operating similar instrumentation and methods into peer group-based operational networks (common RLs).

Jones et al. [11] concluded in their review: no RI is absolutely accurate and is only an estimation. Regardless

of the method used, once a laboratory has derived RIs, it is important that they are subjected to critical scrutiny whether they reflect one single distribution or whether partitioning is required. Indirect methods must be aware of the reason for skewness: either overlapping of different subgroups (which must be compensated by appropriate partitioning), or an incorrect assumption about the distributions involved (needs specific investigation). A genuinely skewed distribution of values in non-diseased individuals is no problem for more recent approaches.

Conclusions for the time being

If RIs estimated by intra-laboratory indirect models are compared with those from direct methods usually taken from extra-laboratory sources, it is essential to consider comprehensive transference aspects and the effect of posture for all corpuscular and protein-dependent blood components.

Direct methods are usually applied with so-called healthy subjects which can walk to the sampling station. The sampling occurs between 7 and 10 (12) am. Indirect methods should consider similar conditions if they are to be compared with direct methods, at least considering the time of sampling and the sitting position. Maybe that the concept of deriving RIs from strictly “healthy” subjects must be modified towards subpopulations which are better comparable with patients.

The time of sampling is interesting for all measurands with circadian rhythms. As long as sufficient data are not available, sampling should occur between 8 and 10 (12) am.

Indirect methods are based on resolution techniques, of which the limits have not been sufficiently investigated, yet. The limits depend on the prevalence of the pathological values and the distance between the value modes of non-diseased and the diseased subjects. Presumably, a prevalence of 25% may be tolerated. The prevalence can be kept low by excluding patients from particular senders of samples (e.g. intensive care units, gynaecological units, etc.). This goal may also be reached by including only patients with one request during a defined time period. Furthermore, emergency cases may be excluded. If they are often not identified in the data pools, only values obtained between 7 and 12 (according to internal requirements) am at workdays may be included. Outside this time window, emergency cases may occur relatively often. In many studies, only the first value was used if several values were obtained during a hospital stay.

The disadvantage of a stringent exclusion policy is the reduction of the data numbers and consequently the

enlargement of the confidence intervals. The critical data number is about 2,000. This reduction can be overcome by collaboration between laboratories using the same analytical platform (common RIs) and serving comparable subpopulations.

Any estimated RL must be critically evaluated for the influence of biological variables. Independent of the selection criteria, partition strategies have to be considered. The most important variables are age and sex which are already automatically integrated in some software programs (e.g. TMC, TML).

Manufacturers of analytical systems are obliged to provide RIs by several directives. The transference responsibility remains with the customer who often does not receive the necessary support from the manufacturers. The transference problem would disappear if laboratories would determine their own RIs. Then, the laboratory would again have sole responsibility for its RIs, and its local role in the medical decision process would regain its importance due to the professional expertise required [17].

The availability of commercially available laboratory information system with high capacity of data storage and the possible and already achieved implementation of indirect RIs scripts (as e.g. the TML approach) creates convenient chances to derive intra-laboratory RIs. This eliminates the problems with transferability, fulfils the dogma of intra-laboratory RIs and facilitates the periodic review of RIs recommended by ISO 15189 [88].

Research funding: None declared.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

References

1. Haeckel R, Wosniok W, Arzideh F. Proposed classification of various limit values (guide values) used in assisting the interpretation of quantitative laboratory test results. *Clin Chem Lab Med* 2009;47:494–7.
2. Özarda Y, Sikaris K, Streichert T, Macri J. Distinguishing reference intervals and clinical decision limits – a review by the IFCC Committee on reference intervals and clinical decision limits. *Crit Rev Clin Lab Sci* 2018;55:420–31.
3. Özcürümez MK, Haeckel R, Gurr E, Streichert T, Sack U. Determination and verification of reference interval limits in clinical chemistry. Recommendations for laboratories on behalf of the Working Group Guide Limits of the DGKL with respect to ISO Standard 15189 and the Guidelines of the German Medical Association on quality assessment in medical laboratories examinations (Rili-BAEK). *J Lab Med* 2019;43:127–33.

4. Sikaris KA. Weighing up our clinical confidence in reference limits. *Clin Chem* 2020;66:1475–6.
5. Özarda Y. Reference intervals: current status, recent developments and future considerations. *Biochem Med* 2016;26:5–16.
6. Haeckel R. The influence of age and other biological variables on the estimation of reference limits of cardiac troponin T. *Clin Chem Lab Med* 2018;56:685–7.
7. Haeckel R, Wosniok W, Torge A, Junker R. Reference limits of high-sensitive cardiac troponin T indirectly estimated by a new approach applying data mining. A special example for measurands with a relatively high percentage of values at or below the detection limit. *J Lab Med* 2021;45:87–94.
8. Kouri T, Kairisto V, Virtanen A, Uusipaikka E, Rajamäki A, Finneman H, et al. Reference intervals developed from data for hospitalized patients: method based on combination of laboratory and diagnostic data. *Clin Chem* 1994;40:2209–15.
9. Cook MG, Levell MJ, Payne RB. A method for deriving normal ranges from laboratory specimens applied to uric acid in males. *J Clin Pathol* 1970;23:778–80.
10. Sikaris KA. Physiology and its importance for reference intervals. *Clin Biochem Rev* 2014;35:3–14.
11. Jones GRD, Haeckel R, Loh TP, Sikaris K, Streichert T, Katayev A, et al. Indirect methods for reference interval determination – review and recommendations. *Clin Chem Lab Med* 2019;57:20–9.
12. Farrell CJL, Nguyen L. Indirect reference intervals: harnessing the power of stored laboratory data. *Clin Biochem Rev* 2019;40: 99–111.
13. Solberg HE. Approved recommendation (1987) on the theory of reference values. *J Clin Chem Clin Biochem* 1987;25:645–56.
14. CLSI CLSI/IFCC Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline, 3rd ed. CLSI document C28-P3. Wayne, PA: Clinical and Laboratory Standards Institute; 2008, vol 28:1–50 pp.
15. Kairisto V, Hänninen KP, Leino A, Pulkki K, Peltola O, Näni ÖV, et al. Generation of reference values for cardiac enzymes from hospital admission laboratory data. *Eur J Clin Chem Clin Biochem* 1994;27:789–96.
16. Poole S, Schroeder LF, Shah N. An unsupervised learning model to identify reference intervals from a clinical database. *J Biomed Inf* 2016;59:276–84.
17. Haeckel R, Wosniok W, Arzideh F. A plea for intra-laboratory reference limits. Part 1. General considerations and concepts for determination. *Clin Chem Lab Med* 2007;45:1033–42.
18. Ishihara I, Boyd JC. An appraisal of statistical procedures used in derivation of reference interval. *Clin Chem Lab Med* 2010;48: 1537–51.
19. Özcürümez MK, Haeckel R. Biological variables influencing the determination of reference limits. *Scand J Clin Lab Invest* 2018; 78:337–45.
20. Tate JR, Yen T, Jones GRD. Transference and validation of reference intervals. *Clin Chem* 2015;61:1012–5.
21. Boyd JC. Cautions in the adoption of common reference intervals. *Clin Chem* 2008;54:238–9.
22. Haeckel R, Wosniok W, Torge A, Junker R, Bertram A, Krebs A, et al. Age and sex dependent reference intervals for random plasma/serum glucose concentrations related to different sampling devices and determined by an indirect procedure with data mining. Urgent plea for studying the diagnostic efficiency of various concepts proposed to improve the pre-examination phase for determining blood glucose concentrations. *J Lab Med* 2021;45:95–101.
23. Arzideh F, Brandhorst G, Gurr E, Hinsch W, Hoff T, Roggenbuck L, et al. An improved indirect approach for determining reference limits from intra-laboratory data bases exemplified by concentrations of electrolytes. *J Lab Med* 2009;33:52–66.
24. Arzideh F, Wosniok W, Haeckel R. Reference limits of plasma and serum creatinine concentrations from intra-laboratory data bases of several German and Italian medical centres. Comparison between direct and indirect procedures. *Clin Chim Acta* 2010;411: 215.
25. Arzideh F, Wosniok W, Gurr E, Hinsch W, Schumann G, Weinstock N, et al. A plea for intra-laboratory reference limits. Part 2. A bimodal retrospective concept for determining reference limits from intra-laboratory databases demonstrated by catalytic activity concentrations of enzymes. *Clin Chem Lab Med* 2007;45: 1043–57.
26. Özarda Y, Aslan D. Use of total patient data for indirect estimation of reference intervals for 40 clinical chemical analytes in Turkey. *Clin Chem Lab Med* 2006;44:867–76.
27. Haeckel R, Wosniok W, Torge A, Junker R. Age and sex dependent reference intervals for uric acid estimated by the truncated minimum chi-square (TMC) approach, a new indirect method. *J Lab Med* 2020;44:157–63.
28. Asgari S, Higgins V, McCudden C, Adeli K. Continuous reference intervals for 38 biochemical markers in healthy children and adolescents: comparison to traditionally partitioned reference intervals. *Clin Biochem* 2019;73:82–9.
29. Zierk J, Arzideh F, Haeckel R, Carlo H, Fröhheld MC, Groß H, et al. Pediatric reference intervals for alkaline phosphatase. *Clin Chem Lab Med* 2017;55:102–16.
30. Palm J, Hoffmann G, Klawonn F, Tutarei D, Palm H. Continuous, complete and comparable NT-pro BNP reference ranges in healthy children. *Clin Chem Lab Med* 2020;58:1509–16.
31. Wosniok W, Haeckel R. A new estimation of reference intervals: truncated minimum chi-square (TMC) approach. *Clin Chem Lab Med* 2019;57:1933–47.
32. Biino G, Balduini CL, Casula L, Cavallo P, Vaccargiu S, Parracciani D, et al. Analysis of 12,517 inhabitants of a Sardinian geographic isolate reveals that predispositions to thrombocytopenia and thrombocytosis are inherited traits. *Haematologica* 2011;96: 96–101.
33. Balduini CL, Noris P. Platelet count and aging. *Haematologica* 2014;99:953–5.
34. Torge A, Haeckel R, Özcürümez M, Krebs A, Junker R. Diurnal variation reference intervals of leucocyte counts indirectly estimated by data mining. *J Lab Med* 2021;45:121–4.
35. Hilderink JM, Klinkenberg LJJ, Aakre KM, de Wit NCJ, Henskens YMC, van der Linden N, et al. Within-day biological variation and hour-to-hour reference change values for hematological parameters. *Clin Chem Lab Med* 2017;55:1013–24.
36. Ishihara K, Ceriotti F, Tam TH, Sueyoshi S, Poon PMK, Thong ML, et al. The Asian project for collaborative derivation of reference intervals: (1) strategy and major results of standardized analytes. *Clin Chem Lab Med* 2013;51:1429–42.
37. Sennels HP, Jørgensen HL, Hansen ALS, Goetze JP, Fahrenkrug J. Diurnal variation of hematology parameters in healthy young males: the Bispebjerg study of diurnal variation. *Scand J Clin Lab Invest* 2011;71:532–41.

38. Sennels HP, Jörgensen HL, Hansen ALS, Fahrenkrug J. Rhythmic 24-hour variation of frequently used clinical biochemical parameters in healthy young males – the Bispebjerg study of diurnal variation. *Scand J Clin Lab Invest* 2012;72:287–95.
39. Bolann BJ. Easy verification of clinical chemical reference intervals. *Clin Chem Lab Med* 2013;51:e279–81.
40. Kallner A, Gustavsson E, Hendig E. Can age and sex related reference intervals be derived for non-healthy and non-diseased individuals from results of measurement in primary health care? *Clin Chem Lab Med* 2000;38:633–54.
41. Farrell CJL, Nguyen L, Carter AC. Data mining for age-related TSH reference intervals in adulthood. *Clin Chem Lab Med* 2017;55:e213–5.
42. Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranges. *Clin Chem* 1990;36:265–70.
43. Lahti A, Hylthoft Petersen P, Boyd JC, Fraser CG, Jörgensen N. Objective criteria for partitioning Gaussian-distributed reference values into subgroups. *Clin Chem* 2002;48:338–52.
44. Harris EK. Effects of intra- and interindividual variation on the appropriate use of normal ranges. *Clin Chem* 1974;20:1535–42.
45. Fraser CG. Inherent biological variation and reference values. *Clin Chem Lab Med* 2004;42:758–64.
46. Johnson NL, Kotz S, Balakrishnan N. Distributions in statistics: continuous univariate distributions. *Wiley Series in Probability and Mathematical Statistics*; 1994:1–784 pp.
47. Haeckel R, Wosniok W. Observed, unknown distributions of clinical chemical quantities should be considered to be log-normal: a proposal. *Clin Chem Lab Med* 2010;48:1393–6.
48. Serfling RJ. Approximation theorems of mathematical statistics. NY: John Wiley & Sons; 1980:1–371 pp.
49. ISO/IEC/IEEE 60559:2020. Information technology – Microprocessor systems – Floating-point arithmetic. Available from: www.iso.org [Accessed 5 Feb 2021].
50. Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. *Clin Chem* 1998;44:622–31.
51. Ishihara I. Statistical considerations for harmonization of the global multicenter study on reference values. *Clin Chim Acta* 2014;432:108–18.
52. Ichihara K, Özarda Y, Barth JH, Klee G, Shimizu Y, Xia L, et al. A global multicenter study on reference values: 2. Exploration of sources of variation across the countries. *Clin Chim Acta* 2016; 467:83–97.
53. Haeckel R, Wosniok W, Arzideh F, Gurr E, Streichert T. Critical comments to a recent EFLM recommendation for the review of reference intervals. *Clin Chem Lab Med* 2017;55:341–7.
54. Cohen AC. Truncated and censored samples, theory and applications. New York: Marcel Dekker; 1991:1–328 pp.
55. Horn PS, Feng L, Li Y, Pesce AJ. Effect of outliers and non-healthy individuals on reference interval estimation. *Clin Chem* 2001;47: 2137–45.
56. Solberg HE, Lahti A. Detection of outliers in reference distributions: performance of Horn's algorithm. *Clin Chem* 2005;51:2326–32.
57. Katayev A, Balciza C, Seccombe DW. Establishing reference intervals for clinical laboratory test results; is there a better way? *Am J Clin Pathol* 2010;133:175–7.
58. Katayev A, Fleming JK, Luo D, Fisher AH, Sharp TM. Reference intervals data mining. No longer a probability paper method. *Am J Clin Pathol* 2015;143:134–42.
59. Hoffmann RG. Statistics in the practice of medicine. *J Am Med Assoc* 1963;185:864–73.
60. Bhattacharya CG. A simple method of resolution of a distribution into Gaussian components. *Biometrics* 1967;23:115–35.
61. Pryce JD. Level of haemoglobin in whole blood and red blood-cells, and proposed convention for defining normality. *Lancet* 1960;2:333–6.
62. Beckett JM. Simplified estimation of normal ranges from routine laboratory data. *Clin Chim Acta* 1970;28:119–25.
63. Kairisto V, Poola A. Software for illustrative presentation of basic clinical characteristics of laboratory tests – GraphROC for windows. *Scand J Clin Lab Invest* 1995;55:43–60.
64. Neumann GJ. The determination of normal ranges from routine laboratory data. *Clin Chem* 1968;14:979–88.
65. Tsay JY, Chen IW, Maxon HR, Heminger L. A statistical method for determining normal ranges from laboratory data including values below the minimum detectable value. *Clin Chem* 1979;25: 2011–4.
66. Hoffmann G, Lichtinghagen R, Wosniok W. Simple estimation of reference intervals from routine laboratory tests. *J Lab Med* 2015; 39:389–402.
67. Klawonn F, Hoffmann G, Orth M. Quantitative laboratory results: normal or lognormal distribution? *J Lab Med* 2020;44:143–50.
68. Baadenhuijsen H, Smit JC. Indirect estimation of clinical chemistry reference intervals from total hospital patient data: application of a modified Bhattacharya procedure. *J Clin Chem Clin Biochem* 1985;23:829–39.
69. Naus AJ, Borst A, Kuppens PS. The use of patient data for the calculation of reference values for some haematological parameters. *J Clin Chem Clin Biochem* 1980;18:621–5.
70. Martin HF, Hologgitas JV, Drisoll J, Fanger H, Gudzinowicz BJ. Reference values based on populations accessible to hospitals. In: Gräsbeck R, Alström T, editors. *Reference values in laboratory medicine*. Chichester: Wiley; 1981:233–62 pp.
71. Concordet D, Geffré A, Braun JP, Trumel C. A new approach for the determination of reference intervals from hospital-based data. *Clin Chim Acta* 2009;405:43–8.
72. Benaglia T, Chauveau D, Hunter DR, Young DS. Mixtools: an R package for analyzing finite mixture models. *J Stat Software* 2009;32:1–29.
73. Zierk J, Arzideh F, Kapsner LA, Prokosch HU, Metzler M, Rauh M. Reference interval estimation from mixed distributions using truncation points and the Kolmogorow-SWmirnow distance (cosmic). *Sci Rep* 2020;10:1704.
74. Holmes DT, Buhr KA. Widespread incorrect implementation of the Hoffmann method, the correct approach, and modern alternatives. *Am J Clin Pathol* 2019;151:328–36.
75. German Society of Clinical Chemistry and Laboratory Medicine. Decision limits/guideline values. Available from: www.dgkl.de/arbeitsgruppen/entscheidungsgrenzen-richtwerte [Accessed 18 Dec 2018].
76. National Cyber Security Centre. Macro security for Microsoft Office (2019 update). Available from: <https://www.ncsc.gov.uk/>

- guidance/macro-security-for-microsoft-office [Accessed 28 Sep 2020].
77. Haeckel R, Wosniok W, Arzideh F. Equivalence limits of reference intervals for partitioning of population data. Relevant differences of reference limits. *J Lab Med* 2016;40:199–205.
 78. Henny J, Vassault A, Boursier G, Vukasovic I, Brguljan PM, Lohmander M, et al. Recommendation for the review of biological reference intervals in medical laboratories. *Clin Chem Lab Med* 2016;54:1893–900.
 79. Haeckel R, Wosniok W. A new concept to derive permissible limits for analytical imprecision and bias considering diagnostic requirements and technical state-of-the-art. *Clin Chem Lab Med* 2011;49:623–35.
 80. Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen. *Dt Ärzteblatt* 2008;105:C301–13 and *Dt Ärzteblatt* 2014;111:A1583–618. Available from: www.aerzteblatt.de/plus1308 [Accessed 28 Sep 2021].
 81. Haeckel R, Wosniok W, Gurr E. Diagnostic efficiency in models for permissible measurement uncertainty. *J Lab Med* 2017;41: 309–15.
 82. O'Halloran MW, Studley-Ruxton J, Wellby ML. A comparison of conventionally derived normal ranges with those obtained from patients' results. *Clin Chim Acta* 1970;27:35–46.
 83. Ceriotti F, Henny J, Queralto J, Ziyu S, Özrada Y, Chen B, et al. Common reference intervals for aspartate aminotransferase (AST), alanine aminotransferase (ALT) and γ -glutamyl transferase (GGT) in serum: results from an IFCC multicenter study. *Clin Chem Lab Med* 2010;48:1593–601.
 84. Metz MP, Loh TP. Describing children's changes using clinical chemistry analytes. *Clin Chem Lab Med* 2016;55:1–2.
 85. Zierk J, Hirschmann J, Toddenroth D, Arzideh F, Haeckel R, Bertram A, et al. Next-generation reference intervals for pediatric hematology. *Clin Chem Lab Med* 2019;57:1595–607.
 86. Ceriotti F. Prerequisites for use of common reference intervals. *Clin Biochem Rev* 2007;28:115–21.
 87. Oosterhuis WP, Modderman TA, Pronk C. Reference values: Bhattacharya or the method proposed by the IFCC? *Ann Clin Biochem* 1990;27:359–65.
 88. International Standard Medical Laboratories. Particular requirements for quality and competence, Geneva, Switzerland: ISO 15189-2003(E): 1–39 pp.