Review

Yesim Ozarda*

Establishing and using reference intervals

Referans aralıklarının oluşturulması ve kullanılması

https://doi.org/10.1515/tjb-2017-0299
Received January 17, 2020; accepted January 18, 2020; previously published online February 11, 2020

Abstract: Reference intervals (RIs) and clinical decision limits (CDLs) are fundamental tools used by healthcare and laboratory professionals to interpret patient laboratory test results. The traditional method for establishing RIs, known as the direct approach, is based on collecting samples from members of a preselected reference population, making the measurements and then determining the intervals. For challenging groups such as pediatric and geriatric age groups, indirect methods are appointed for the derivation of RIs in the EP28-A3c guideline. However, there has been an increasing demand to use the indirect methods of deriving RIs by the use of routine laboratory data stored in the laboratory information system. International Federation of Clinical Chemistry (IFCC), Committee on Reference Intervals and Decision Limits (C-RIDL) is currently working on the study for the comparison of the conventional (direct) and alternative (indirect) approaches for the determination of reference intervals. As a matter of fact that, the process of developing RIs is often beyond the capabilities of an individual laboratory due to the complex, expensive and time-consuming process to develop them. Therefore, a laboratory can alternatively transfer and verify RIs established by an external source (i.e. manufacturers' package inserts, publications). IFCC, C-RIDL has focused primarily on RIs and has performed multicenter studies to obtain common RIs in recent years. However, as the broader responsibility of the Committee, from its name, includes "decision limits", the C-RIDL also emphasizes the importance of the correct use of both RIs and CDLs and to encourage laboratories to specify the appropriate information to clinicians as needed.

Keywords: Reference intervals; EP28-A3c guideline; Direct reference intervals; Multicenter studies; Indirect methods; IFCC; C-RIDL; Clinical decision limits; Verification of reference intervals.

Öz: Referans aralıkları (RI) ve klinik karar sınırları (CDL), sağlık ve laboratuvar profesyonelleri tarafından hasta laboratuvar test sonuçlarını yorumlamak için kullanılan temel araçlardır. Doğrudan yaklaşım olarak bilinen RI'lerin oluşturulmasına yönelik geleneksel yöntem, önceden seçilmiş bir referans popülasyonunun üyelerinden örneklerin toplanması, ölçümlerin yapılması ve daha sonra aralıkların belirlenmesine dayanmaktadır. Pediatrik ve geriatrik yaş grupları gibi zorlu gruplar için, EP28-A3c kılavuzunda RI'nin türetilmesi için dolaylı yöntemler atanır. Bununla birlikte, laboratuvar bilgi sisteminde depolanan rutin laboratuvar verilerinin kullanılmasıyla UR'nin türetilmesi için dolaylı yöntemlerin kullanılması yönünde artan bir talep vardır. Uluslararası Klinik Kimya Federasyonu (IFCC), Referans Aralıkları ve Karar Sınırları Komitesi (C-RIDL) şu anda referans aralıklarının belirlenmesi için geleneksel (doğrudan) ve alternatif (dolaylı) yaklaşımların karşılaştırılması için çalışma üzerinde çalışmaktadır. Nitekim, RI geliştirme süreci genellikle bunları geliştirmek için karmaşık, pahalı ve zaman alıcı bir süreç nedeniyle bireysel bir laboratuvarın yeteneklerinin ötesindedir. Bu nedenle, bir laboratuvar alternatif olarak harici bir kaynak (yani üreticilerin paket ekleri, yayınları) tarafından kurulan RI'leri aktarabilir ve doğrulayabilir. IFCC, C-RIDL öncelikle RI'lere odaklanmıştır ve son yıllarda ortak RI elde etmek için çok merkezli çalışmalar yapmıştır. Bununla birlikte, komitenin daha geniş sorumluluğu adına "karar sınırları" içerdiğinden, C-RIDL aynı zamanda hem RI'lerin, hem de CDL'lerin doğru kullanımının önemini vurgulamakta ve laboratuvarları gerektiğinde klinisyenlere uygun bilgileri belirtmeye teşvik etmek için teşvik etmektedir.

***Corresponding author: Yesim Ozarda,** Uludag University, School of Medicine, Department of Clinical Biochemistry, Bursa, Turkey; and IFCC/Committee on Reference Intervals and Decision Limits (C-RIDL), Chair, Bursa, Turkey, e-mail: yesim@uludag.edu.tr

Anahtar Kelimeler: Referans aralıkları; EEP28-A3c kılavuzu; Doğrudan referans aralıkları; Çok merkezli çalışmalar; Dolaylı yöntemler; IFCC; C-RIDL; Klinik karar sınırları; Referans aralıklarının doğrulanması.

Introduction

In clinical laboratory medicine, patient test results are often interpreted by comparison to reference intervals (RIs), which are usually defined as the central 95% of laboratory test results obtained from a healthy reference population. Therefore, accurate RIs of laboratory analyses are an integral part of the process of correct interpretation of clinical laboratory test results [1].

Grasbeck and his colleagues published the initial paper entitled "Normal Values and statistics" in the mid 20th century, [2]. In subsequent years, it was realized that the terminology of "normal values was not adequate and even partially incorrect". Therefore, in 1969, the concept of the reference value(s) was launched by Saris and Grasbeck [3] in a session devoted to normal values during a Congress of Clinical Laboratory Medicine and the term of reference values are well accepted instead of normal values [4]. From 1987 to 1991, the IFCC published a series of six papers, in which it was recommended that each laboratory follow defined procedures to produce its own RIs [5–10]. Interest has been renewed in the topic as a result of the following regulatory initiatives in the last two decades: according to the European Directive 98/79 on in vitro diagnostic medical devices, diagnostic kit manufacturers are obliged to supply their clients with appropriate RIs for use with their assay platforms and reagents [11], and the International Organization for Standardization 15189 standard for clinical laboratory accreditation states that each laboratory should periodically re-evaluate its own RIs [12]. The guideline entitled "Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory" EP28-A3c provides the necessary steps mainly for the selection of reference individuals, pre-analytical and analytical considerations, analysis of reference values for a RI establishment study, transference and verification of the RIs in theoretically [13]. However, in the present-day era of evidence-based medicine, there is still a big gap between theory and practice with respect to the application of RIs as decision-making tools, despite the mandatory requirements.

The recommended process for defining a RI is the so-called "direct" approach, where subjects representing the reference population are selected and sampled and the specimen analyzed for this purpose [13]. An alternative approach is the "indirect" approach where results from specimens are collected for routine purposes, which have been collected for screening, diagnostic or monitoring purposes and are used to determine the reference intervals. The IFCC, C-RIDL has recently published a review of the strengths and weaknesses of the indirect approach to discussing the potential advantages and disadvantages of indirect approach compared with direct methods [14]. Additionally, it has been possible to derive "common" or "harmonized" RIs on a national level from multicenter studies. Over the last decade, the C-RIDL has focused on direct common RIs; it has developed guidelines on conducting such studies and has conducted multicenter studies to derive common RIs at a national level [15–18].

RIs describe the typical distribution of results seen in a healthy reference population while CDLs are associated with a significantly higher risk of adverse clinical outcomes or are diagnostic for the presence of a specific disease. However, as the two concepts are sometimes confused, the principles for describing RIs and CDLs should be known well and kept separate [19].

Under optimal conditions, a laboratory should perform its own RI study to establish RIs specific for its method and the local population. However, the process of developing RIs is often beyond the capabilities of an individual laboratory due to the complex, expensive and time-consuming process to develop them. Therefore, a laboratory can alternatively verify RIs established by an external source [20]. Although procedures for verifying RIs in the literature and guidelines are clear in theory, gaps remain for the implementation of these procedures in routine clinical laboratories [21].

The aim of the review is to describe various aspects of establishing and using RIs together with a detailed evaluation of the most recent studies and publications of IFCC, C-RIDL.

Establishing direct reference intervals

The recommended protocol for setting a direct reference interval is to perform a direct reference interval study according to standard published procedures [13]. RIs are derived from reference distribution, usually of 95% interval, and describe a specific population. The concept of RIs is now well established and the classical cascade is defined from reference individuals, a reference population, a reference sample group, reference values, reference distribution, reference limits and RIs. The reference individuals form the reference sample group for measurement of the values from the reference population. Through statistical analysis of the distribution of the obtained values, the reference limits are calculated. These limits then define RI [2].

Selection of reference individuals

Health is a relative condition lacking a universal definition. The designation of good health and determination of normality for a candidate reference individual may involve a variety of examinations, such as a history and physical and/or certain clinical laboratory tests. The inclusion, exclusion and partitioning criteria can be implemented appropriately through a well-designed questionnaire [13]. Exclusion criteria are features which prevent the individual from being included in the reference sample. Although some criteria, such as alcohol, tobacco and some environmental factors, may be potential exclusion criteria, amounts of consumption of alcohol and tobacco can be recorded in detail on the sample questionnaire and the effects are evaluated statistically, primarily using multiple regression analysis [18]. Written informed consent from participants is needed from each reference individual who agrees to participate in the study. The consent form should state clearly that laboratory personnel is allowed to obtain specimens, and use the associated laboratory values and questionnaire information for the determination of RIs [13].

There are two main sampling approaches to include the reference individuals in the direct RI studies: (1) the priori approach. This refers to RI studies where the inclusion of reference individuals is determined before samples are collected. This is usually based on a questionnaire and/or clinical examination. Typically all results from these samples are included in the data with the exception of removing samples considered to be outliers based on a statistical procedure. (2) The posteriori approach. This refers to direct RI studies where the definition of reference individuals includes information applied after the samples have been collected or after the measurements have been performed.

Pre-analytical and analytical aspects

The pre-analytical considerations involve biological (i.e. sampling time in relation to biological rhythms, fasting or non-fasting and physical activity) and methodological factors (i.e. sample collection techniques, type of additives, with or without tourniquet and sampling equipment, specimen handling, transportation, time and speed of centrifugation, and storage conditions). For reproducibility and standardization, it is essential that the pre-analytical aspects are accurately defined and described as the preanalytical phase is known to have the highest errors in the total test process [22]. Some pre-analytical factors will affect results and so should be considered when performing RI studies, reviewing the literature, or when applying the intervals to patient results. Examples include serum versus heparin plasma when measuring potassium or total protein; time of day for collection for serum cortisol or testosterone; sample handling, such as the time until centrifugation for potassium measurement; and common interferences, such as hemolysis for potassium, CK, AST and LDH [23].

Analytical aspects include the analytical variability of the method used for the measurement, equipment/instrumentation, reagents, calibration standards, and calculation methods. Different commercial methods may be used in a trueness-based approach to the reference measurement system providing results traceable to the system and thus, comparable results can be produced in clinical laboratories. When performing a RI study, the reference measurement systems and standard reference materials are of great importance to ensure the traceability of the test results in comparisons [23].

Statistical evaluation to establish direct reference intervals

Calculation of RIs includes parametric and nonparametric calculation methods, detection of outliers, partitioning, and confidence intervals. In the parametric calculation method, the most suitable transformation method must be selected (e.g. logarithmic, Box-Cox power or some other function) and testing is then applied to establish whether the transformed reference values conform to Gaussian distribution [9]. Box-Cox power transformation often has been used to transform data to a Gaussian distribution for parametric computation of RIs [24]. However, this transformation occasionally fails. Therefore, the non-parametric method based on the determination of the 2.5 and 97.5 percentiles following sorting of the data, has been recommended for general use. Although the EP28-A3c recommends the non-parametric calculation method, the parametric calculation method may have an advantage over the non-parametric method in allowing identification and exclusion of extreme values during RI computation [25].

Whichever method is used in the calculation of the RIs, detection, and exclusion of the outliers are very important to obtain reliable RIs. A simple but effective method for the detection of outliers is a visual inspection of the data. The method proposed by Dixon (D: the absolute value of the difference between the suspected outlier and the next or proceeding value, R: the entire range of the observations) [26]. If the D/R ratio is more than 1/3, the outlier is discarded. However, this method is not very sensitive when there is more than one outlier. The Tukey method is a more sophisticated method, which includes Box-Cox transformation of the data to obtain Gaussian distribution followed by identification of the outliers in interquartile ranges (IQR: Q3–Q1; Q1: lower quartile, Q3: upper quartile). At levels of <Q1 – 1.5 IQR and/or >Q3 + 1.5 IQR, the outliers are discarded [27]. The latent abnormal value exclusion (LAVE) method proposed by Ichihara and Boyd [25] is a secondary exclusion method to exclude possibly abnormal results hidden within the reference values. This method is an iterative approach for the derivation of multiple reference RIs simultaneously when no exclusion of values has been made in the initial computation of the RIs. The algorithm then uses those initial values of RIs to judge the abnormality of each individual's record by counting the number of abnormal results in tests other than the one for which the RI is being determined [28].

Stratification of RIs by age and gender is the minimum pre-requisite and other means include race, ethnicity, body mass index or nutritional habits. The most widely-used partitioning method is that of Harris and Boyd, in which the means and standard deviations of the subgroups are considered as a separate different standard deviation that may produce different limits [29]. However, this method is only appropriate for analytes with a Gaussian distribution with subclasses, where the values are of similar size and standard deviation. A similar method was proposed by Lahti et al. allowing the estimation specifically of the percentage of subjects in a subclass outside the RIs of the entire population in any situation [30]. More recently, Ichihara and Boyd recommended a partitioning method on the basis of the magnitude of the standard deviations of test results named standard deviation ratio (SDR) [25]. An SDR greater than 0.3 can be regarded as a guide for the consideration of partitioning reference values. This method is based on two or three-level nested analysis of variance.

It is clearly recommended that at least 120 subjects are required to calculate the confidence intervals (CIs) of the lower and upper RIs in the guideline [13]. The CI is a range of values including the true percentile (e.g. the 2.5th percentile of the population) with a specified probability, usually of 90% or 95%, as the "confidence level" of the interval. Horn and Pesce proposed a "robust method" that based on the transformation of the original data according to Box and Cox followed by a "robust" algorithm giving different weights to the data, depending upon their distance from the mean [31]. However, a robust method with such a small number of reference subjects (e.g. n = 20) can lead to uncertainty of calculated reference limits revealed by the width of its CIs. In the EP28-A3c guideline, non-parametric CIs are given from the observed values corresponding to certain rank numbers from Reed et al. [32]. When there is a small n, to calculate the 90% CIs around the limits, it is recommended to use "the bootstrap method" which is a "resampling" method and creates a "pseudosample" from the data. The RI is derived from each pseudosample and the process is repeated many times (100–2000) yielding a distribution of lower and upper RIs [33]. From this distribution, 5th and the 95th quantiles may be used to determine the 90% CI for each limit.

Establishing indirect reference intervals

Data mining, or "big data", is the process of using previously generated data to identify new information. Routine pathology databases often contain many thousands or millions of results from many 100s or 1000s of patients, which can be used in establishing RIs. Using the data for the goal of determining population RIs by indirect techniques is one example of data mining [34]. In addition to setting RIs, data in pathology databases can be used for internal quality control [35], external quality assessment [36], reference interval validation [37] and determining biological variation data [38–40]. By definition, the population will be derived from one or more routine pathology databases. Before starting any statistical analysis, some basic considerations are necessary to consider which results from the data set should be included.

Source of the data

As the aim is to produce "health-associated" RIs, then results from outpatients are clearly preferred, particularly from those in a primary care setting. The high frequency of inflammation, recumbency, intravenous fluids, medications, and dietary changes, in addition to the disease(s) leading to the admission, makes inpatient samples less desirable.

Number of subjects

There is no prescription for the number of samples required; however, "more is better" to produce robust results. However, to provide a starting point for further work in this area, 1000 subjects may be considered a small number and above 10,000 as a large number, and in populations that are poorly represented in a database (e.g. extremes of age), smaller numbers may still provide useful information. It is recommended to use a minimum of 400 reference subjects for each partition for a statistically reliable reference interval calculation [25].

Stability over time

Before using any data set, it is important to ensure that the analytical method and the population have been stable over the period of data collection [41]. The first assessment is a historical review, i.e. has the method been changed or the population serviced changed during the period of data collection. This can be further assessed by reviewing medians and other percentiles over the time period of data collection, together with the assessment of QC and EQA results.

Partitioning

Partitioning can be performed by plotting medians and selected centiles (e.g. 10th and 90th) for males and females against patient age. The need for partitioning can be assessed by several objective criteria. Harris and Boyd [29] recommend a separate Rl when the ratio of standard deviation (larger over smaller) between the subgroups exceed 1.5, or when the Z-statistics between the two subgroup distribution exceeds 3.

Exclusions

Data sets can be "biochemically filtered" to reduce the frequency of results from subjects where there is a higher likelihood of disease affecting the result. A recommended approach is to exclude specific subgroups (e.g. emergency departments or intensive care units). An additional recommended approach is to limit results to a single result per patient. As a diseased patient is more likely to be retested than a non-diseased patient, failure to do this is likely to lead to overrepresentation of results from unwell subjects. When selecting a single result, the last result of a patient during a "healthcare episode" is preferred as it is most likely to represent a return towards health [14].

Statistical techniques

The Bhattacharya method is a graphical method for identifying a Gaussian distribution in the midst of other data [42]. It was originally developed in the precomputer era using manual paper-based systems. Computer-based versions have been developed in Java and Microsoft Excel. The Bhattacharya method is user-dependent, requiring selection of bin size for the data, the bin location and the number of bins included in the analysis. The Hoffmann technique also identifies a homeostatically regulated population subset of test results in a data set that is assumed to follow Gaussian or near-Gaussian distribution [43]. Like the Bhattacharya method, it was developed in the precomputer era for paper-based systems. More recently, this method has been used in a computerized form [44, 45]. A limitation of the Hoffmann procedure is that it is influenced by the presence of a secondary population of significant size [46] although filtering of the data can reduce this effect. A more sophisticated procedure than those of Hoffmann and Bhattacharya was developed by Arzideh et al. [41]. In this process, a smoothed kernel density function was estimated for the distribution of the total mixed data of the sample group (combined data of non-diseased and diseased subjects). It was assumed that the "central" part of the distribution of all data represents the non-diseased population. Standard parametric (mean and standard deviation) or non-parametric statistics (percentiles), such as those used in direct RI studies, can also be used for indirect studies. This will involve outlier removal, either before or after transformation, followed by calculation of the mean and SD or median and relevant percentiles [14].

Comparison of the direct and indirect approaches

Important benefits of the indirect approach, relative to the direct approach, include that it is faster and cheaper as the indirect approach is based on data that have already been generated as part of routine care, thus excluding the resource-intensive components, i.e. patient identification, recruiting, specimen collection and measurement, of the direct approach. Indirect methods also use the same preanalytical and analytical techniques used for patient management and can provide very large numbers for assessment [47]. However, there are risks and difficulties associated with indirect approaches. The most important risk is the question as to whether the presence of diseased individuals influences the RIs. Furthermore, there is still no consensus on the best statistical method to derive indirect RIs. IFCC, C-RIDL is currently working on the study for the comparison of different statistical techniques for establishing indirect reference intervals with the existing direct methods to close this deficit in the literature.

Using reference intervals from external source

RIs in most clinical laboratories remain out of date and incomplete due to the complex process of their establishment [48]. Therefore, instead of developing RIs directly from an apparently healthy population, most laboratories receive RIs for clinical use from various sources (e.g. manufacturers' package inserts, publications, textbooks, multicenter studies, published national or international expert panel recommendations, guidelines, local expert groups or data mining of existing data). However, several differences can exist between the sample collection procedures and laboratory operations of the laboratory originated RI study and the local laboratory receiving the RI. Therefore, it is of critical importance for a local laboratory to address the following question prior to receiving RIs from an external source: "Is this RI suitable for my laboratory's collection processes, method, and population?" [20]. The EP28A3c guideline provides recommendations for transferring and verifying RIs established by external sources for a local laboratory [13]. This approach is advantageous for many laboratories as it does not require extensive recruitment of healthy reference individuals and is thus time and cost-efficient.

Transference of reference intervals

Assuming the original RI study was performed using a robust methodology and statistical procedures, transferring a RI requires that certain conditions be fulfilled in order to be acceptable, prior to verifying and receiving a RI. There are two main scenarios in which RIs are transferred. First, reference values may originate from a different population/laboratory method than the receiving laboratory, and second, reference values may originate from a laboratory that shares the same laboratory method/population as the receiving laboratory. In the first instance, comparing the laboratory methods serves as an instructive early screening tool to assess the suitability of the reference values for the receiving laboratory. Laboratory methods can be compared by a method comparison study between the method used during the development of the RI and the method used by the receiving laboratory to determine the statistical validity of a RI transfer [49]. For a method comparison study, samples must be collected with an appropriate distribution of values spanning the RI, as an insufficient range may underestimate and a range too large may overestimate the strength of the correlation. The correlation between the two methods is subsequently analyzed and, if appropriate, linear regression analysis is performed to determine the slope and y-intercept values of the best-fit regression line [50]. These values are subsequently used to transfer the RI. According to the EP28-A3c guideline, the best-fit regression line should have a slope bias close to 1, a y-intercept close to 0 and a correlation coefficient (r2) close to 1 [13]. Furthermore, according to CLSI EP09-A3guidelines, the scatter and bias plots should be examined for constant scatter to ensure there are no dramatic differences between the variation at the upper and lower ends of the range of values [50].

If the preanalytical processes, the laboratory methods, and the populations are very similar to those of the laboratory where the RIs originated, the method comparison study is still recommended to confirm the comparability, although the bias between the laboratory methods is expected to be very small [50]. However, in this situation, subsequent verification using samples from healthy reference individuals may not be necessary, and the laboratory may opt to simply perform a subjective assessment by carefully inspecting the reference population demographics, geographic location, preanalytical and analytical procedures, analytical performance and the statistical methods used in the RI study. If these factors are all consistent with the receiving laboratory's population and procedures, the RI may be transferred with further verification [13].

Verification of reference intervals

Following transference, the CLSI EP28-A3c guideline recommends subsequently verifying the transferred RI. The guideline emphasizes that three approaches can be used to verify RIs: (1) a subjective assessment, (2) using a small number of reference individuals (e.g. n = 20) and (3) using a large number of reference individuals (e.g. n = 60, but fewer than 120) [13]. Using a large number of reference individuals is not generally preferred by routine laboratories, as this is nearly the same as the sample size required for a RI study. The standard approach recommended by the guideline for routine practice in laboratories is to collect and analyze samples from 20 healthy subjects per age and/or sex partition from the receiving laboratory's local population and to compare these reference values with the RI established from the larger, more robust, original study. The Dixon [26] or Tukey methods [27] should be applied to test and subsequently remove outliers, and new specimens should be obtained to replace those removed. If no more than 2 of the 20 samples (i.e. 10% of the test results) fall outside the RI, it may be received for use, at least provisionally. If 3 or 4 of the 20 samples fall outside the RI, a second set of 20 reference specimens should be obtained. If again 3 or more of the new specimens (i.e. ≥10% of the test results) or 5 or more of the original 20 fall outside the RI, the user should reexamine the analytical procedures used and consider possible differences in the biological characteristics of the two populations sampled [13]. Although this guideline appears straightforward, RI partitions, the presence of outliers and initial unsuccessful verification can further complicate the verification process [21].

Establishing common reference intervals

Although direct RIs are most commonly established using a well-defined and representative reference population, with sample analysis performed by a single laboratory, RIs can also be determined with the intention of serving a much broader population demographic and/or geographic location with sample analysis performed by a single platform or multiple platforms, termed common RIs. There are two types of common RIs. The first is objective RIs, which have many prerequisites [51] and are defined by well-conducted multicenter studies [52, 53], and the second is subjective RIs, defined by the survey(s) and guidance from a group of experts using the harmonization approach [54].

In recent years, IFCC, C-RIDL has performed a direct multicenter RI study in many countries with total recruitment of 13,386 healthy adults to determine global RIs of 25 analytes were measured chemically and 25 immunologically [18, 55]. This is an example of a well-conducted multicenter RI study, in which each laboratory acts as a central laboratory and sample analysis is performed using multiple platforms. In this type of multicenter study, it is essential to perform rigorous quality control monitoring to detect analytical deviations and use internationally accepted reference materials for standardized analytes to ensure traceability in each center. In addition to internationally accepted reference materials, the global IFCC, C-RIDL study is based on a common protocol [15] and the use of a panel of sera to harmonize measurement results [56]. The basic scheme for conducting the global study was to make test results comparable among countries based on the panel test results measured in each participating country [56]. This approach resulted in a method comparison and successful transference of the data obtained from the global study. As part of the global study, a multicenter RIs study was also performed in Turkey, including seven geographical regions, using traceable materials and panel of sera from 40 reference individuals from the global study in the central laboratory, using a single platform, as an example of studies where the measurements were performed in one center acting as the central laboratory [53]. With the lack of regional differences and the well-standardized status of test results, common RIs for Turkey have been derived from this nationwide study. Additionally, Ozarda et al. performed "cross-check testing" using at least 20 samples to compare results among the participating laboratories in Turkey as recommended in the protocol for multicenter studies [15]. Thus, common RIs were transferred from the multicenter study to each participating laboratory in Turkey using the linear regression slope and intercept [50].

Clinical decision limits

It is important that RIs and CDLs are clearly distinguished [57]. The question "Is the patient healthy or not healthy?" relates to RIs that describe the typical distribution of results seen in an apparently healthy reference population. However, the other questions ("Is the patient at risk of a developing a disease, or is the patient diseased, or worsening?") are related to CDLs, where values above or below the threshold are associated with a significantly higher risk of adverse clinical outcomes or are defined as diagnostic for the presence of a specific disease [58]. In contrast to the RIs, where there are two limits (upper and lower), there is only one CDL, which is usually an upper limit. However, according to the likelihood of various clinical situations or different clinical questions, multiple low and high CDLs may be used. The most obvious example is fasting glucose, where several decision limits are defined: a level ≤3.9 mmol/L (≤70 mg/dL) for the diagnosis of hypoglycemia [≤2.2 mmol/L (≤40 mg/dL) for life-threatening hypoglycemia] [59]; 5.6–6.9 mmol/L (100–124 mg/dL) for an increased risk of diabetes or prediabetes; and ≥7.0 mmol/L (≥125 mg/dL) for diabetes mellitus [60].

RIs are focused on optimizing specificity (typically to 95%) while CDLs are also focused on optimizing sensitivity for the disease. In other clinical circumstances that place importance on both sensitivity and specificity, optimal limits may be derived from receiver operator characteristic (ROC) curves, which balance sensitivity and specificity [61]. Ideally, a Bayesian approach is also required to balance pretest probability or prevalence against sensitivity and specificity [62]. "Optimal limits" derived from ROC curves, an intermediate category of a threshold, are a compromise between specificity and sensitivity [63]. The term "critical value," also known as a critical result, panic value, or alert value, represents a pathophysiological state different from normal that poses a risk to a patient's life unless immediate action is taken [64]. Currently, the use of the term, panic value, is discouraged, because it suggests emotional stress and because it is contrary to the process of communicating information clearly [65].

Analytical quality affects the reliability of RIs. The biological variability theory suggests that the desirable bias for RI classification takes into account intraindividual and interindividual variability [$<0.25\times(\mathrm{CVi}^2+\mathrm{CVg}^2)^{1/2}$] and that it will prevent an unacceptable increase in the proportion of healthy individuals flagged as outside RIs. Analytical quality will similarly affect the application of CDLs, although the impact is defined not by the statistics of the reference population distribution but by the clinical risk definitions as well as the prevalence of the disease [66]. Increasing measurement uncertainty generally causes greater clinical uncertainty; similarly, the impact of uncorrected measurement bias will lead to clinical bias. The traceability of method calibration is vitally important for both RIs and CDLs. Neither universal CDLs (e.g. for lipids and HbA1c) nor common CDLs can be clinically reliable without traceability and analytical quality standards [67].

IFCC, C-RIDL has recently published a review to emphasize the importance of the correct use of both RIs and CDLs and to encourage laboratories to specify the appropriate information to clinicians as needed [19]. The Committee concluded that the distinction of RIs and CDLs should be clear in the laboratory report that would greatly improve the post-analytical quality of interpretation and facilitate the national and international adoption of common RIs (including harmonized RIs) and CDLs.

References

1. Ozarda Y. Reference intervals: current status, recent developments and future considerations. Biochem Med 2016;26:5–16.
2. Gräsbeck R, Fellman J. Normal values and statistics. Scand J Clin Lab Invest 1968;21:193–5.
3. Gräsbeck R, Saris NE. Establishment and use of normal values. Scand J Clin Lab Invest 1969;10(Suppl 1):62–3.
4. Gräsbeck R. Reference values (formerly called normal values). Chronobiologia 1977;4:59–61.
5. Solberg HE. International Federation of Clinical Chemistry. Scientific committee, Clinical Section. Expert Panel on Theory of Reference Values and International Committee for Standardization in Haematology Standing Committee on Reference Values. Approved recommendation (1986) on the theory of reference values. Part 1. The concept of reference values. Clin Chim Acta 1987;165:111–8.
6. Petit Clerc C, Solberg HE. International Federation of Clinical Chemistry (IFCC). Approved recommendation on the theory of reference values. Part 2. Selection of individual for the production of reference value. Clin Chim Acta 1987;170:S1–12.
7. Solberg HE, Petit Clerc C. International Federation of Clinical Chemistry (IFCC). Approved recommendation on the theory of reference values. Part 3. Preparation of individuals and collection of the specimens for the production of reference values. Clin Chim Acta 1988;177:S3–11.
8. Solberg HE, Stamm D. International Federation of Clinical Chemistry (IFCC). Approved recommendation on the theory of reference values. Part 4. Control of analytical variation in the production, transfer and application of reference values. Clin Chim Acta 1991;202:S5–11.
9. Solberg HE. International Federation of Clinical Chemistry (IFCC). Approved recommendation on the theory of reference values. Part 5. Statistical treatment of collected reference values. Clin Chim Acta 1987;170:S13–32.
10. Dybkaer R, Solberg HE. International Federation of Clinical Chemistry (IFCC). Approved recommendation on the theory of reference values. Part 6. Presentation of observed values related to reference values. Clin Chim Acta 1987;170:S33–42.
11. Directive 98/79/EC of European Parliament and the Council of 27 October 1998 on in vitro diagnostic medical devices. Offical J Eur Commun 7 December 1998; L331/1-L331/37.
12. International Organization for Standardization. Medical laboratories – requirements for quality and competence [cited 2018 Jun 9]. ISO 15189:2012. Available from https://www.iso.org/standard/56115.html.
13. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; Approved Guideline. 3rd ed. CLSI EP28-A3. Wayne, Pennsylvania: Clinical and Laboratory Standards Institute; 2010.
14. Jones GR, 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.
15. Ozarda Y, Ichihara K, Barth JH, Klee G. Committee on Reference Intervals and Decision Limits (C-RIDL), International Federation of Clinical Chemistry. Protocol and standard operating procedures for common use in worldwide multicenter study on reference values. Clin Chem Lab Med 2013;51:1027–40.
16. Ceriotti F. Common reference intervals the IFCC position. Clin Biochem 2009;42:297.
17. Ceriotti F. Henny J, Queraltó J, Ziyu S, Özarda 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.
18. Ichihara K, Ozarda Y, Barth JH, Klee G, Qiu L, Erasmus R, et al. A global multicenter study on reference values: 1. Assessment of

methods for derivation and comparison of reference intervals. Clin Chim Acta 2017;467:70–82.

19. Ozarda Y, Sikaris K, Streichert T, Macri J, IFCC Committee on Reference intervals and Decision Limits (C-RIDL). Distinguishing reference intervals and clinical decision limits – a review by the IFCC Committee on Reference Intervals and Decision Limits. Crit Rev Clin Lab Sci 2018;55:420–31.
20. Tate JR, Tina Y, Jones GR. Transference and validation of reference intervals. Clin Chem 2015;8:1012–5.
21. Ozarda Y, Higgins V, Adeli K. Verification of reference intervals in routine clinical laboratories: practical challenges and recommendations. Chem Lab Med 2018;57:30–7.
22. Plebani M, Sciacovelli L, Aita A, Chiozza ML. Harmonization of pre-analytical quality indicators. Biochem Med (Zagreb) 2014;24:105–13.
23. Jones G, Barker A. Reference intervals. Clin Biochem Rev 2008;29:S93–7.
24. Box GEP, Cox DR. An analysis of transformations. JR Stat Soc 1964;B26:211–52.
25. Ichihara K, Boyd J. An appraisal of statistical procedures used in derivation of reference intervals. Clin Chem Lab Med 2010;48:1537–51.
26. Dixon WJ. Prossesing data for outliers. Biometrics 1953;9: 74–89.
27. Tukey JW. Exploratory data analysis. Reading, MA, USA: Addison-Wesley, 1977:688.
28. Klee GG, Ichihara K, Ozarda Y, Baumann NA, Straseski J, Bryant SC, et al. Reference intervals: comparison of calculation methods and evaluation of procedures for merging reference measurements from two US medical centers. Am J Clin Pathol 2018;150:545–54.
29. Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranges. Clin Chem 1990;36:265–70.
30. Lahti A, Hyltoft Peterson P, Boyd JC. Impact of subgroup prevalances on partitioning Gaussian distributed reference values. Clin Chem 2002;48:1987–99.
31. Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. Clin Chem 1998;44:622–31.
32. Reed AH, Henry RJ, Manson WB. Influence of statistical method used on the resulting estimate on normal range. Clin Chem 1971;17:275–84.
33. Efron B. The jacknife, the bootstrap, and other resampling plans. Philadelphia, PA: CBMS-NSF regional conference series in applied mathematics, 1982:92.
34. Solberg HE. Using a hospitalized population to establish reference intervals: pros and cons. Clin Chem 1994;40:2205–6.
35. Fleming JK, Katayev A. Changing the paradigm of laboratory quality control through implementation of real-time test results monitoring: for patients by patients. Clin Biochem 2015;48:508–13.
36. De Grande LA, Goossens K, Van Uytfanghe K, Stöckl D, Thienpont LM. The Empower project – a new way of assessing and monitoring test comparability and stability. Clin Chem Lab Med 2015;53:1197–204.
37. Jones GR. Validating common reference intervals in routine laboratories. Clin Chim Acta 2014;432:119–21.
38. Cembrowski GS, Tran DV, Higgins TN. The use of serial patient blood gas, electrolyte and glucose results to derive biologic variation: a new tool to assess the acceptability of intensive care unit testing. Clin Chem Lab Med 2010;48:1447–54.
39. Loh TP, Ranieri E, Metz MP. Derivation of pediatric within-individual biological variation by indirect sampling method: an LMS approach. Am J Clin Pathol 2014;142:657–63.
40. Loh TP, Metz MP. Indirect estimation of pediatric between-individual biological variation data for 22 common serum biochemistries. Am J Clin Pathol 2015;143:683–93.
41. 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.
42. Bhattacharya CG. A simple method of resolution of a distribution into Gaussian components. J Biometric Soc 1967;23:115–35.
43. Hoffmann RG. Statistics in the practice of medicine. J Am Med Assoc 1963;185:864–73.
44. Katayev A, Balciza C, Seccombe DW. Establishing reference intervals for clinical laboratory results. Is there a better way? Am J Clin Pathol 2010;133:180–6.
45. 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.
46. Grindler ME. Calculation of normal ranges by methods used for resolution of overlapping Gaussian distributions. Clin Chem 1970;16:24–8.
47. Barth JH. Reference ranges still need further clarity. Ann Clin Biochem 2009;46:1–2.
48. Ceriotti F, Hinzmann R, Panteghini M. Reference intervals: the way forward. Ann Clin Biochem 2009;46:8–17.
49. Adeli K, Higgins V, Trajcevski K, White-Al Habeeb N. The canadian laboratory initiative on pediatric reference intervals: a CALIPER white paper. Crit Rev Clin Lab Sci 2017;54:358–413.
50. CLSI document EP09-A3. Method procedure comparison and bias estimation using patient samples, approved guideline, 3rd ed. Wayne (PA): CLSI, 2013.
51. Ceriotti F. Prerequisites for use of common reference intervals. Clin Biochem Rev 2007;28:115–21.
52. Adeli K, Raizman JE, Chen Y, Higgins V, Nieuwesteeg M, Abdelhaleem M, et al. Complex biological profile of hematologic markers across pediatric, adult, and geriatric ages: establishment of robust pediatric and adult reference intervals on the basis of the Canadian Health Measures Survey. Clin Chem 2015;61:1075–86.
53. Ozarda Y, Ichihara K, Aslan D, Aybek H, Ari Z, Taneli F, et al. A multicenter nationwide reference intervals study for common biochemical analytes in Turkey using Abbott analyzers. Clin Chem Lab Med 2014;52:1823–33.
54. Tate JR, Sikaris KA, Jones GR, Yen T, Koerbin G, Ryan J, et al. Harmonising adult and paediatric reference intervals in Australia and New Zealand: an evidence-based approach for establishing. A first panel of chemistry analytes. Clin Biochem Rev 2014;35:213–35.
55. Ichihara K, Ozarda 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 2017;467:83–97.
56. Ichihara K, Ozarda Y, Klee G, Straseski J, Barth JH, Baumann N, et al. Utility of panel of sera for alignment of test results in the worldwide multicenter study on reference values. Clin Chem Lab Med 2013;51:1007–20.
57. Waise A, Price HC. The upper limit of the reference range for thyroid-stimulating hormone should not be confused with a

cut-off to define subclinical hypothyroidism. Ann Clin Biochem 2009;46:93–8.

58. Lindstedt G, Tryding N. There is difference between decision limits and reference intervals. Reference intervals are based on measurements in healthy individuals, decision limits on measurements in patients. Lakartidningen 2007;104:2076–9.
59. EMA. Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus. European Medicines Agency, 2012. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129-256.pdf [cited 2018 June 15].
60. Expert Committee on the Diagnosis Classification of Diabetes Mellitus. Report of the expert committee on the diagnosis and classification of diabetes mellitus. Diabetes Care 2003;26:S5–20.
61. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–77.
62. Kallner A. Bayes’ theorem, the ROC diagram and reference values: definition and use in clinical diagnosis. Biochem Med (Zagreb) 2018;28:010101.
63. Reach G. Which threshold to detect hypoglycemia? Value of receiver-operator curve analysis to find a compromise between sensitivity and specificity. Diabetes Care 2001;24:803–4.
64. Campbell C, Caldwell G, Coates P, Flatman R, Georgiou A, Horvath AR, et al. Consensus statement for the management and communication of high risk laboratory results. Clin Biochem Rev 2015;36:97–105.
65. White GH, Campbell CA, Horvath AR. Is this a critical, panic, alarm, urgent, or markedly abnormal result? Clin Chem 2014;60:1569–70.
66. Petersen PH, Jensen EA, Brandslund I. Analytical performance, reference values and decision limits. A need to differentiate between reference intervals and decision limits and to define analytical quality specifications. Clin Chem Lab Med 2011;50:819–31.
67. Petersen PH, Klee GG. Influence of analytical bias and imprecision on the number of false positive results using guideline-driven medical decision limits. Clin Chim Acta 2014;430:1–8.