Clinica Chimica Acta 432 (2014) 108–118

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/clinchim

Statistical considerations for harmonization of the global multicenter study on reference values

Kiyoshi Ichihara *

Department of Laboratory Sciences, Faculty of Health Sciences, Yamaguchi University Graduate School of Medicine, Ube, Japan

ARTICLE INFO

Article history:
Received 30 August 2013
Received in revised form 25 October 2013
Accepted 17 January 2014
Available online 8 February 2014

Keywords:
Multicenter study
Reference interval
Panel of sera
Modified Box–Cox transformation formula
Latent abnormal values exclusion method
Reduced major-axis regression

ABSTRACT

The global multicenter study on reference values coordinated by the Committee on Reference Intervals and Decision Limits (C-RIDL) of the IFCC was launched in December 2011, targeting 45 commonly tested analytes with the following objectives: 1) to derive reference intervals (RIs) country by country using a common protocol, and 2) to explore regionality/ethnicity of reference values by aligning test results among the countries. To achieve these objectives, it is crucial to harmonize 1) the protocol for recruitment and sampling, 2) statistical procedures for deriving the RI, and 3) test results through measurement of a panel of sera in common.

For harmonized recruitment, very lenient inclusion/exclusion criteria were adopted in view of differences in interpretation of what constitutes healthiness by different cultures and investigators. This policy may require secondary exclusion of individuals according to the standard of each country at the time of deriving RIs. An iterative optimization procedure, called the latent abnormal values exclusion (LAVE) method, can be applied to automate the process of refining the choice of reference individuals.

For global comparison of reference values, test results must be harmonized, based on the among-country, pairwise linear relationships of test values for the panel. Traceability of reference values can be ensured based on values assigned indirectly to the panel through collaborative measurement of certified reference materials. The validity of the adopted strategies is discussed in this article, based on interim results obtained to date from five countries. Special considerations are made for dissociation of RIs by parametric and nonparametric methods and between-country difference in the effect of body mass index on reference values.

1. Introduction

The standardization of laboratory test results has been improved a great deal by concerted efforts of international and national societies of clinical chemistry and laboratory medicine. However, the reference intervals (RIs) for the laboratory tests, which have an essential role in interpreting the test results, remain largely discordant among clinical laboratories. This situation reflects the difficulty of conducting a primary study to derive RIs in a harmonized way. In fact, the RIs depend on which people are included or excluded as reference individuals and what method is used for computing the RIs from the values obtained on samples from these individuals, as well as on what steps were taken for standardization of the analytes.

A previous guideline for "Defining, establishing, and verifying reference intervals in the clinical laboratories" (C28-A3) issued by CLSI/IFCC [1] was based on a series of discussions made by the expert panel of IFCC on reference intervals [2–7]. It has played a very important role in giving a scientific ground to tackle issues concerning RI by defining technical

Abbreviations: Alb, albumin; Alc, alcohol consumption; ALT, alanine aminotransferase; ALP, alkaline phosphatase; AMY, amylase; AST, aspartate aminotransferase; BMI, body mass index; C-RIDL, Committee on Reference Intervals and Decision Limits; C3, complement component 3; C4, complement component 4; Ca, total calcium; CA125, carbohydrate antigen 125; CK, creatine kinase; Cl, chloride; CLSI, Clinical and Laboratory Standards Institute; CV(b), coefficient of variation of slope "b"; CVA, analytical CV; CVI, within-individual CV; DR, Deming regression; CRE, creatinine; CRM, certified reference material; Fe, iron; GGT, gamma-glutamyltransferase; Glu, glucose; GM, grand mean; HBV, hepatitis B virus; HCV, hepatitis C virus; HDL-C, HDL cholesterol; HIV, human immunodeficiency virus; IgA, immunoglobulin A; IgG, immunoglobulin G; IgM, immunoglobulin M; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; IP, inorganic phosphate; K, potassium; LAVE, latent abnormal value exclusion; LDL-C, LDL cholesterol; LDH, lactate dehydrogenase; LL, lower limit; MLE, maximum likelihood estimation; MRA, multiple regression analysis; Na, sodium; PTH, intact parathyroid hormone; RMP, reference measurement procedure; RI, reference interval; SD, standard deviation; SD_{BI}, biological variation expressed in SD; SD_{RI}, SD_{BI} corresponding to 1/4 of reference interval; SD_{SRC}, s source of variation expressed in SD; SD_{TB}, combined between-individual and within-individual variations expressed in SD; SDR, standard deviation ratio; SDG, between-individual SD; SD_I, within-individual SD; SD_{RI}, SD comprising RI or $\sqrt{}$(SD_I^2 + SD_G^2); SDR_{LL}, SD ratio of LL; SDR_{UL}, SD ratio of UL; SE, standard error; stdβ, standard partial regression coefficient; SRM, Standard Reference Material; SV, source of variation; TCho, total cholesterol; TBil, total bilirubin; Tf, transferrin; TG, triglyceride; TP, protein, total; TTR, transthyretin (prealbumin); UA, uric acid; UL, upper limit; UN, urea nitrogen.

* Department of Clinical Laboratory Sciences, Faculty of Health Sciences, Yamaguchi University Graduate School of Medicine, Minami-Kogushi 1-1-1, Ube 755-8505, Japan. Tel.: +81 836 22 2884; fax: +81 836 35 5213.
E-mail address: ichihara@yamaguchi-u.ac.jp.

0009-8981/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cca.2014.01.025

terms and clarifying what aspects must be considered in conducting a study for establishing RI. However, the guideline is not complete in describing a clear procedure for the proper selection of reference individuals. Besides, statistical methods proposed in the guideline for computing the RI and judging a need for partitioning the RI by sex and age remain controversial because of a paucity of evidence showing superiority of one method over others [8].

Despite the methodological issue for establishing RI, it is generally agreed that we need to conduct a multicenter study, ensuring a sufficient sample size to obtain statistically reproducible RIs. This collaborative approach appears very practical to establish RIs for globally standardized analytes (such as electrolytes, serum enzymes, glucose, uric acid, creatinine, etc.). Therefore, it has been proposed that common RIs should be derived for those analytes by strictly ensuring the traceability of test results among participating laboratories [9–12].

Accordingly, the Nordic countries [13], Spain [14], Australia [15], Japan [16], and IFCC groups [17] have conducted such studies for establishing regional or national RIs. However, the study protocols used have not been well harmonized, in part due to the lack of detailed descriptions in C28-A3 for conducting multicenter studies. In fact, all the studies assumed cross-standardization of the analytical systems among the collaborating laboratories with no anticipation of regional or assay platform dependent differences in test results.

On the other hand, a series of multicenter studies conducted in East and South-East Asia for derivation of common RIs, mainly targeting serum proteins (IgG, IgA, IgM, C3, C4, Tf, etc.) [18,19], were unique in that all the specimens collected in the collaborating laboratories were sent to Japan for centralized measurement by the same laboratory. Analyzer-dependent bias could not otherwise be fully eliminated despite the assumed global standardization of test results by the worldwide distribution of the certified reference material CRM470 [20]. Appreciable regional differences in test results were observed for most of those proteins which act as inflammatory markers: immunoglobulins, complement components, etc. This finding led to the third, expanded Asian study targeting additional non-standardized analytes and a wider geographical area [21,22]. The same centralized measurement scheme was again adopted to explore regionality of test values and to cope with non-standardized analytes. As a result, the study not only confirmed the regional differences in test results for inflammatory markers but also revealed regionality in HDL-C, PTH, CA125, and other analytes. These intriguing findings pointed to the importance of conducting a similar study in a global scale in order to address the question of to what extent we can share RIs, in consideration of sources of variation including regional and racial differences in test results.

With this background, the C-RIDL held a series of discussions between 2010 and 2011 on the feasibility of launching a global study and generated a protocol to be used by all participants in the study [23]. The proposed protocol can be divided into three aspects: 1) to conduct the multicenter study on reference values country by country by use of a harmonized protocol for recruitment, sampling and measurement, 2) to establish RIs for each country applying common statistical procedures, and 3) to explore sources of variation of reference values, including regionality and ethnicity, for a wide range of analytes after harmonization of test results through measurement of a common panel of sera.

The most challenging issue in generating the protocol was the specification of globally applicable inclusion/exclusion criteria for recruitment of healthy individuals because of differences in interpretation of what constitutes healthiness by different cultures and investigators. Therefore, lenient criteria were adopted by setting no upper limits to BMI and daily consumption of alcohol, for example. However, this policy may require secondary exclusion of test results at the time of computing country specific RIs according to criteria suitable for each country. The C-RIDL also discussed the statistical theory for aligning test results from one country to another and described the results of the pilot study conducted to evaluate its applicability to 45 commonly-tested analytes [24].

In this article, the rationales for the statistical procedures adopted for harmonization of the global study are discussed, mainly based on interim analyses of the data obtained from five countries: Japan, Turkey, China, the USA, and Saudi Arabia. The data size and composition of gender (males: females) were 652 (291:361), 2785 (1395:1390), 448 (212:236), 494 (217:277), and 826 (398:428), respectively.

2. Harmonization of recruitment and sampling

There are no agreed-upon criteria regarding who should be included as reference individuals. The most critical issue in recruitment is how carefully we need to exclude individuals who might have latent but highly prevalent pathological conditions such as the 'metabolic syndrome', diabetes mellitus, alcohol-related liver diseases, or iron deficiency anemia. We may adopt strict criteria in recruitment excluding anyone who is in slight excess of ideal body weight or who has a habit of smoking cigarettes or drinking alcoholic beverages. However, the number of volunteers who are qualified for the study would be reduced to a small fraction, and thus it becomes virtually impossible to conduct a primary study of reference values. In fact, we cannot completely eliminate latent diseases without performing specific tests to diagnose them.

To address this dilemma, the last Asian study conducted in 2009 may be of relevance. It targeted healthy individuals from 20 to 65 years of age, most of whom were working for the hospitals affiliated with the collaborating laboratories. The following exclusion criteria were adopted by unanimous consensus among the participating laboratories [22]: 1) Alcohol consumption $\geq$75 g of ethanol/day, 2) body mass index (BMI) $\geq$28 kg/m^2, 3) smoking >20 cigarettes/day, 4) regular drug therapy, 5) pregnant or within one year after childbirth, or 6) known carrier state for HBV, HCV, or HIV. Considering the high prevalence of the above-mentioned disorders, the criteria were far from ideal, allowing for a moderate degree of obesity compared to the Asian standard, a fairly high level of regular consumption of alcohol, etc. However, the fourth criterion, ineligibility of those individuals who are on regular medication, was regarded as a strict one which precludes inclusion of individuals with clinically-active stages of the related disorders. Furthermore, the application of the latent abnormal values exclusion (LAVE) method (see below) [8] prior to computing the RIs was expected to be a protection against inclusion of individuals with latent disorders.

In the ongoing global study, however, the above Asian criteria were regarded impractical for conducting a study in Europe and the USA, where moderate obesity (BMI > 28) is highly prevalent, and there is no specific criterion defining a healthy level of consumption of alcoholic beverages. Furthermore, with adoption of the policy setting no upper age limit, exclusion of those receiving on-going medication makes it more difficult to recruit individuals from the older age groups. Thus, if we adopt too stringent criteria for recruiting volunteers, implementation of the global study aiming at comparison of reference values will be practically impossible. As a result, the most pragmatic inclusion and exclusion criteria were adopted by consensus for the global study even allowing subjects taking medications or vitamin supplements if not conflicted with exclusion criteria including history of DM or of chronic renal or hepatic diseases [23].

It was decided that, at the time of analyzing the reference values and computing the RIs for a given country, it is necessary to exclude individuals secondarily based on standards suitable for that country. In contrast, in order to align and compare reference values among countries for evaluation of regionality and ethnicity, it would be necessary to adopt more stringent criteria in matching individuals to make appropriate comparisons. In fact, looking at the interim results obtained from the five countries, the demographic profile differs