CLIN. CHEM. 37/9, 1580–1582 (1991)

Statistical Criteria for Separate Reference Intervals: Race and Gender Groups in Creatine Kinase

Eugene K. Harris,[1] Edward T. Wong,[2] and S. T. Shaw, Jr.[3]

Previously published data confirming differences in creatine kinase (EC 2.7.3.2) among various race and gender subgroups in the Los Angeles area have been re-examined with use of recently proposed statistical criteria for defining separate reference intervals. Results indicate that one criterion may be too lenient, whereas another is clearly too restrictive in suggesting the need for separate intervals. Further experience with other analytes in both large and small population samples would be helpful.

The validity of separate reference intervals for creatine kinase (CK; EC 2.7.3.2) in different race and gender groups was confirmed almost a decade ago by Wong et al. (*1*) in a study of >1500 apparently healthy employees of a medical center in Los Angeles. Three categories of CK values were identified and verified by analysis of variance: (*a*) high CK, found in black men; (*b*) intermediate CK, found in black women and Hispanic, Asian, white, and other (Native American) men; and (*c*) low CK, in Hispanic, Asian, white, and other women. Mean differences among the race–gender groups placed in the intermediate CK category and in the low CK category were not statistically significant according to analysis of variance.

Recently, new statistical criteria have been proposed (*2*) for establishing separate reference intervals for different population groups. These criteria are based on the idea that, when the proportions of groups falling above or below the reference limits for the combined population differ substantially from the expected values for those limits (e.g., 2.5% on each side), then separate reference intervals are needed. Specifics of these criteria, which involve critical differences between both the means and the standard deviations of groups, are discussed below. The CK data analyzed previously (*1*) provide an opportunity to examine the application of these new criteria to more than two groups and to large numbers of reference subjects within each group.

Brief Review of CK Data

Details of the blood sampling and analytical procedures we followed were published previously (*1*), as were age, height, and weight characteristics of the different race and gender groups. As stated by Wong et al. (*1*), volunteers for this study were questioned about their participation in strenuous physical activities (e.g., running or weight-lifting) because such exertion is known to produce unusually high values of serum CK. Results for individuals participating in such activities were excluded from the database. Values from pregnant women were not excluded.

Wong et al. (*1*) noted the extreme skewness of their data on the commonly accepted measurement scale (IUB enzyme units, U/L). Conversion to logarithms greatly improved the symmetry of the overall distribution but did not entirely eliminate nongaussian skewness and kurtosis, which remained statistically significant (probably because of the statistical power of a large sample). The analysis here has been carried out with the log CK values because these were much more gaussian-like in their distribution than were the nontransformed data. Later (see Table 2), we contrast nonparametric 95% reference limits with the comparable limits based on the assumption of gaussian (normal) distributions.

Two of the >1500 results collected appeared to be statistical outliers within their race–gender groups and were deleted. One of these (6756 U/L) was from a black man, the other (3971 U/L) from an Hispanic man. By the somewhat conservative "one-third" rule suggested by Reed et al. (*3*) (i.e., labeling as an outlier an extreme result whose difference from the next higher or lower observation exceeds one-third of the range of all observations), no other CK value came close to qualifying as an outlier.

With these two deletions, the means and standard deviations of the log CK results, by group, are given in Table 1. For reference, the geometric means (antilogarithms of the means of the logarithms) are also listed.

New Statistical Criteria

The statistical criteria proposed by Harris and Boyd (*2*) for establishing separate reference intervals were concerned with only two groups and an analyte distributed in gaussian form within each group. The following two rules were suggested:

1) Assuming that the same number of individuals (N) has been sampled in each group, separate reference intervals should be established if the normal deviate test statistic,

$$z = (\bar{x}_i - \bar{x}_2)N^{1/2}/(s_i^2 + s_2^2)^{1/2} \quad (1)$$

exceeds the critical value $z^* = 3(N/120)^{1/2}$, where $\bar{x}_1$, $\bar{x}_2$, s_1, and s_2 are the means and standard deviations,

[1] Clinical Laboratories, Department of Pathology, University of Virginia Health Sciences Center, Box 168, Charlottesville, VA 22908.
[2] Department of Pathology, University of Southern California Medical Center, 1200 North State St., Los Angeles, CA 90033.
[3] Department of Pathology, The University of Chicago, 5841 South Maryland Ave., Chicago, IL 60637.
Received March 11, 1991; accepted July 8, 1991.

Table 1. Means and Standard Deviations of Log CK, by Race–Gender Group[a]

Group	N	Log CK Mean	Log CK SD	Geometric mean, U/L
High CK				
Black men	195	2.231	0.270	170
Intermediate CK				
Hispanic men	117	1.996	0.237	99.0
Asian men	70	2.033	0.165	108
White men	164	1.973	0.251	93.9
Black women	317	1.975	0.224	94.3
Low CK				
Asian women	143	1.808	0.160	64.3
Hispanic women	219	1.803	0.185	63.5
White women	283	1.759	0.214	57.4

[a] The group "Other" has been omitted because the number of individuals sampled was small (11 men, 16 women).

Table 2. Reference Limits for High, Intermediate, and Low CK Categories

Category	N	Log CK Mean	Log CK SD	95% reference limits, U/L[a,b] Param.	Nonparam.
High	195	2.231	0.270	50–574	52–520
Intermediate	668	1.984	0.229	34–270	35–345
Low	645	1.785	0.195	25–147	25–145

[a] Parametric limits are the antilogarithms of the mean ± 1.96 SD. Nonparametric limits are taken from ref. *1* (Table 3).

[b] Miller et al. (*5*) recommend that in highly skewed distributions such as that for serum CK results, the estimate of the upper 97.5th percentile should be based on >400 subjects. We exceeded this number in the intermediate and low categories, but not in the high category. Although the logarithmic transformation substantially reduced skewness in all categories, the accuracy of the estimated upper limit among black men would undoubtedly be improved by a larger sample of this group.

respectively, of groups 1 and 2.

2) Regardless of the value of z, separate reference intervals should be established if the ratio of the larger to the smaller standard deviation, e.g., s_2/s_1, exceeds 1.5.

Clearly, the CK data represent a more complicated situation: three overall categories, of which two contain more than one population group (in this case, race–gender), with unequal numbers of individuals in each group. However, the rules can still be applied. First, we compute a weighted mean and standard deviation for log CK to represent the combined race–gender groups in each category of high, intermediate, and low CK. No combination is needed for the high category, which consists only of black men. For the other two categories, the combined mean is given by the formula

$$\bar{x}_c = \Sigma(n_i\bar{x}_i)/\Sigma n_i \qquad (2)$$

where n_i and $\bar{x}_i$ refer to the number of individuals and the mean result in the ith race–gender group included in the cth category. The combined variance for this category is computed from the formula

$$s_c^2 = \bar{s}_i^2 + \text{variance}\,\bar{x}_i \qquad (3)$$

where s_i^2 is the variance of the analyte in the ith race–gender group, and variance $\bar{x}_i = \Sigma n_i(\bar{x}_i - \bar{x}_c)^2/\Sigma n_i$. The combined standard deviation for the category is s_c.

Another statistical criterion has been proposed by Sinton et al. (*4*), namely, that separate intervals for two groups should not be established unless the difference between the means exceeds 25% of the combined 95% reference interval.

Results

The results of these calculations, and the parametric 95% reference limits based on them, are presented in Table 2. For comparison, nonparametric reference limits representing 2.5th and 97.5th percentiles (as in ref. *1*) are also given.

Comparing the mean log CK values of the high and intermediate categories, we calculate

$$z = \frac{2.231 - 1.984}{[(0.270)^2/195 + (0.229)^2/668)]^{1/2}} = 11.6$$

much higher than the critical value $z^* = 3\,(432/120)^{1/2} = 5.7$, where N has been replaced by the average number of individuals per category.

Similarly, comparing the intermediate and low categories, we find

$$z = \frac{1.984 - 1.785}{[(0.229)^2/668 + (0.195)^2/645)]^{1/2}} = 17.0$$

even farther above the critical value of $z^* = 7.0$.

Clearly, the classification of these CK measurements by various race–gender groups into high, intermediate, and low categories more than satisfies the proposed statistical criteria (*2*) for establishing separate reference intervals.

Table 1 indicates that within the intermediate CK category, Asian men show a higher mean CK and a lower standard deviation than the other groups in that category. Although these tendencies are probably real, calculations show that the z-statistic comparing the mean for this group with the combined mean for the other three groups falls below 3.0, whereas the ratio of the combined standard deviation for the other groups to the standard deviation among Asian men is <1.5. Only one of the 70 Asian men sampled showed CK values outside the reference limits for the intermediate CK category, whereas three or four would have been expected. However, narrower reference limits for Asian men, on the basis of this relatively small sample, may not be warranted in practice. Additional study to include more Asian men would be useful to test this possibility further.

Applying the criterion of Sinton et al. (*4*), we note that the width of the combined 95% reference interval is

approximately equal to $4s_c$. Computing the differences between the means of each category and the combined standard deviation for each pair of categories, we obtain the following ratios of mean difference to $4s_c$: 0.24 for high vs intermediate log CK, and 0.21 for intermediate vs low log CK.

From these results, the criterion of Sinton et al. (*4*) appears too stringent, not allowing the establishment of separate reference intervals where they are clearly indicated. On the other hand, the critical z-value proposed by Harris and Boyd (*2*), $z^* = 3\ (N/120)^{1/2}$, may be too lenient, calling for separate reference intervals in some cases where the difference between two distributions may not be clinically important. Comparing the two criteria when the number of reference subjects is the same in both groups shows the critical z-value proposed by Harris and Boyd to be equivalent to a Sinton et al. ratio (e.g., p%) of only 9.5%. Writing the z^*-value in general form as $z^* = k(N/120)^{1/2}$ indicates a simple relation between k and p, namely, $k = 0.31p$.

Remarks

In theory, separate reference intervals for general medical use should not become an issue until research studies have demonstrated a physiological basis for differences between population subgroups with respect to a given analyte. Field studies would then be undertaken to determine whether these physiological grounds lead to clinically important differences that should be recognized by separate reference intervals.

In practice, however, it is more likely that analysis of existing data, guided by statistical criteria, will reveal unanticipated differences between demographic groups. Only then would research into possible physiological bases for such differences be undertaken and their clinical relevance considered. In our view, a clinically important difference exists whenever the proportion of individuals in a given subgroup of the population that falls outside (or inside) the 95% reference limits for the combined population is considerably different from the expected value of 2.5% on each side. Such differences can lead to significant discrepancies between actual and expected sensitivities and specificities. The critical value $z^* = 3\ (N/120)^{1/2}$ was based on this consideration. Nevertheless, setting any critical value involves an arbitrary judgement, and only further experience will indicate whether the factor 3 represents the best choice.

References

1. Wong ET, Cobb C, Umehara MK, et al. Heterogeneity of serum creatine kinase activity among racial and gender groups of the population. Am J Clin Pathol 1983;79:582–6.
2. Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranges. Clin Chem 1990;36:265–70.
3. Reed AH, Henry RJ, Mason WB. Influence of statistical method used on the resulting estimate of normal range. Clin Chem 1971;17:275–84.
4. Sinton TJ, Cowley DM, Bryant SJ. Reference intervals for calcium, phosphate, and alkaline phosphatase as derived on the basis of multichannel-analyzer profiles. Clin Chem 1986;32:76–9.
5. Miller WG, Chinchilli VM, Gruemer HD, Nance WE. Sampling from a skewed population distribution as exemplified by estimation of the creatine kinase upper reference limit. Clin Chem 1984;30:18–23.