Are the common reference intervals truly common? Case studies on stratifying biochemical reference data by countries using two partitioning methods

A. LAHTI

Department of Clinical Chemistry, Rikshospitalet University Hospital of Oslo, Norway

Lahti A. Are the common reference intervals truly common? Case studies on stratifying biochemical reference data by countries using two partitioning methods. Scand J Clin Lab Invest 2004; 64: 407–430.

The Harris-Boyd method, recommended for partitioning biochemical reference data into subgroups by the NCCLS, and a recently proposed new method for partitioning were compared in three case studies concerning stratification by countries (Denmark, Finland, Norway, and Sweden) of reference data collected in the Nordic Reference Interval Project (NORIP) for the enzymes alkaline phosphatase (ALP), creatine kinase (CK), and γ-glutamyl transpeptidase (GGT). The new method is based on direct estimation of the proportions of two subgroups outside the reference limits of the combined distribution, while the Harris-Boyd method uses easy-to-calculate test parameters as correlates for these proportions. The decisions on partitioning suggested by the Harris-Boyd method deviated from those obtained by using the new method for each of the three enzymes when considering pair-wise partitioning tests. The reasons for the poor performance, as it seems to be, of the Harris-Boyd method were discussed. Stratification of reference data into more than two subgroups was considered as both a theoretical problem and a practical one, using the four country-specific distributions for each enzyme as illustration. Neither the Harris-Boyd method nor the new method seems ideal to solve the partitioning problem in the case of several subgroups. The results obtained by using prevalence-adjusted values for the proportions seemed, however, to warrant the conclusion to be made that there are no major differences in terms of the partitioning criteria between the levels of each of the three enzymes in the four countries. Because these three enzymes include those two tests (CK, GGT), which in the preliminary analyses of the project data had shown largest variation between countries, the tentative conclusion was drawn that application of common reference intervals in the Nordic countries is feasible, not only for the three enzymes examined in the present study but for all of the tests involved in the NORIP project.

Key words: Alkaline phosphatase (ALP); creatine kinase (CK); gamma-glutamyl transpeptidase (GGT); partitioning into several subgroups; prevalence; reference interval; reference limit; reference value; stratification; tied observation

Ari Lahti, Department of Clinical Chemistry, Rikshospitalet University Hospital of Oslo, NO-0027 Oslo, Norway. Tel. +47 2307 1057, fax. +47 2307 1080, e-mail. Jari.lahti@rikshospitalet.no

DOI 10.10.80/00365510410006027

INTRODUCTION

The reference intervals produced by the Nordic Reference Interval Project (NORIP) were calculated by pooling the reference data obtained from the five Nordic countries. However, before those reference intervals can be considered useful for each participating country, one should examine whether the country-specific reference distributions are similar enough in terms of partitioning criteria to justify such pooling of data. If differences related to ethnicity occur, these might be expected to be more pronounced for enzymes than for non-enzymes, because as gene-expression regulated proteins, enzymes are probably more susceptible to the effects of genetic variation.

The traditional method in statistics that is used for comparisons between several groups simultaneously is the analysis of variance (ANOVA). Because this method focuses on comparisons between group means, it does not seem ideal for making decisions on partitioning biochemical reference data. When reference distributions are being compared for partitioning purposes, differences between means should not be the main concern of the investigator, because partitioning is about deciding whether or not a certain reference interval (the reference interval for the combined distribution of two or more subgroups) could be used to substitute another reference interval (a subgroup-specific reference interval). Whether such a substitution is appropriate or not should rather be evaluated by assessing how much the specificity and sensitivity of the laboratory test in question would be altered by it, i.e. by estimating to what extent that substitution would change the proportions of a subgroup distribution outside the reference limits. Those proportions should not deviate much from 2.5%, which is their value outside subgroup-specific reference limits (assuming the conventional definition of a reference interval as covering the central 95% of a reference distribution), to keep the actual and expected sensitivities and specificities close to each other.

Recently, new partitioning criteria were proposed, which are in their most general form based explicitly on estimates of the subgroup proportions outside the common reference limits [1 – 3]. In this report, three case studies, concerning the enzymes alkaline phosphatase (ALP), creatine kinase (CK), and γ-glutamyl transpeptidase (GGT), will be presented in which both these new partitioning criteria and those of Harris & Boyd [4, 5], recommended by the NCCLS (National (USA) Committee for Clinical Laboratory Standards) [6], will be applied to assess whether the reference distributions obtained from Denmark, Finland, Norway, and Sweden for these enzymes should be stratified. Stratification of reference data into several subgroups will be discussed in theoretical terms, and the usefulness of both partitioning methods to perform pair-wise comparisons and to solve multi-partitioning problems, involving more than two subgroups simultaneously, will be examined.

GENERAL CONSIDERATIONS ON PARTITIONING BIOCHEMICAL REFERENCE DATA INTO MORE THAN TWO SUBGROUPS

Partitioning into more than two subgroups is a common problem in reference interval studies. The most frequently encountered multi-partitioning task is probably to decide whether stratification should be performed with respect to age. Because age is a continuous variable, the first thing to do is to select appropriate age groups for the partitioning tests. However, such groups can be determined in many ways. One investigator could divide the age range into decades while another might consider larger groups, such as young age, middle age, and old age, defined more or less arbitrarily by the investigator himself. The number of subjects that the investigator has collected in particular age groups may also affect the division of the age range made by him, because he has to consider the statistical quality of his results.

Any pre-test division of a continuous independent variable into subgroups is to some extent arbitrary. Unfortunately, different divisions can lead to different conclusions: one division, assessed by an appropriate statistical method, might lead to a suggestion that group-specific reference intervals should be calculated for the groups defined by using that division, while the same method would perhaps not detect important differences between groups obtained by another division of the same reference data. Statistical technicalities may explain, together with possible true differences between reference data, the variability observed in the stratifications applied by different laboratories to biochemical markers showing correlation with a continuous variable, such as age.

To preclude pre-test selection of subgroups, regression-based continuous reference intervals could be considered. Applying regression analysis, stratification into subgroups becomes immaterial, and regression-based continuous reference intervals have also the advantage of making estimation of age-specific reference limits feasible at any age, through interpolation. Methods to calculate continuous confidence intervals for each continuous reference limit also exist [7]. Hence, producing continuous reference intervals seems preferable to considering calculation of group-specific reference intervals for arbitrarily selected subgroups, whenever a laboratory test shows clear-cut correlation with a continuous variable.

However, regression-based continuous reference intervals are not useful for clinicians at present because regression curves cannot be integrated into the existing laboratory information systems (LIS). Laboratories obviously cannot expect clinicians to estimate age-specific reference limits for each patient from such curves visually. Rather, the LIS should calculate these limits from the regression equation of a desired biochemical marker when given the age, gender, and possibly some other data on a patient as input parameters, and supply every test result reported by the laboratory with such patient-specific reference limits. But because doctors may wish to check reference intervals also when not considering laboratory reports, such calculations should ideally be within their reach in any actual situation of clinical decision-making. Therefore, regression-based continuous reference intervals will probably have limited impact on clinical practice until portable computers, preferably offering mobile on-line connection to the LIS, have been substituted for the laboratory booklets on reference intervals in doctors' pockets!

In the case of categorical independent variables, such as country in this study, the stratification problem cannot be avoided by constructing continuous reference intervals. On the other hand, arbitrary pre-test division of the data is not a concern, as opposed to continuous variables. The simplest method used to examine stratification of reference data into several subgroups is to apply the *t*-test (or its non-parametric analogues) pair-wise to two distributions at a time. This approach has, however, two drawbacks:

1. The *t*-test performs a comparison between the means of two distributions, but because the subgroup distributions can have different standard deviations, differences between their means do not necessarily reflect the behavior of their reference limits. Conclusions based on comparisons between the means are therefore not automatically relevant for comparing reference intervals. Non-parametric equivalents of the *t*-test, such as the Mann-Whitney test, recommended for comparisons of non-gaussian distributions, are hardly better than the *t*-test in this respect.
2. When performing multiple independent tests, some of these will likely lead to a significant conclusion by chance alone. If each test is made at the 0.05 level, one comparison will be found non-significant with a probability of 0.95, assuming that there are no real differences between the groups. The probability for two independent comparisons to be non-significant is $0.95^2 = 0.90$, for three $0.95^3 = 0.86$, and so on. The probability for at least one comparison among, say, six independent comparisons to be false significant would be $1 - 0.95^6 = 0.26$, which is considerably higher than the level of one test (0.05).

Despite these major drawbacks, multiple pair-wise tests are used frequently to assess partitioning into several subgroups in clinical chemical literature. Statistical methods to perform comparisons between several groups simultaneously, such as ANOVA (for gaussian distributed sample means) and the non-parametric Kruskal-Wallis test, are an improvement as compared to multiple pair-wise tests, because they do not have drawback 2. The same is not true for drawback 1, however.

The partitioning method developed by Harris & Boyd [4, 5] seems to be an attempt to avoid drawback no. 1. Recommended by the NCCLS [6], this method is probably widely used at present by clinical chemists to solve partitioning problems. On p. 90 of their outstanding book *Statistical Bases of Reference Values in Laboratory Medicine* [5], Harris and Boyd describe how their method could be used to examine stratification of reference data into several subgroups:

> "Another alternative [to performing pair-wise tests], particularly when more than three categories are involved, is to start by carrying out an analysis of variance of all results together, using a generalized least squares program if necessary to take account of varying numbers of subjects in each subgroup. A significant *F*-test from this analysis should then be followed by simultaneous comparison of paired means, preferably, in our opinion, the Tukey test, which controls all such comparisons at the 0.05 probability level while maintaining a high probability of detecting real differences between pairs. Any pair of means whose difference is statistically significant should be re-examined against the more stringent z^* critical levels suggested earlier."

In the partitioning method used by Harris and Boyd, pair-wise comparisons are made using the test parameter of the standard normal deviate test:

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (1)$$

where $\bar{x}_i$, s_i, and n_i are the mean, standard deviation, and sample size, respectively, of subgroup *i*. This equation gives the difference between the means of two subgroups divided by the standard deviation of that difference. Initially, Harris & Boyd [4] proposed the following critical value for *z*:

$$z3 = 3 \cdot \sqrt{\frac{n}{120}} \qquad (2)$$

where *n* is the average of the numbers of reference values in the two subgroups [8]. In a subsequent publication [5], they were of the opinion that this criterion could be too permissive of partitioning, and suggested a new critical value, which was 5/3 times as large as the original one:

$$z5 = 5 \cdot \sqrt{\frac{n}{120}} \qquad (3)$$

This latter criterion (Eq. 3) is not mentioned in the NCCLS guidelines for partitioning [6], however, and it seems that the original criterion (Eq. 2) is the one that practitioners mostly use.

Apart from an adjustment to the numbers of reference values in the subgroups, made in these threshold values to keep the stringency of the test stable at varying sample sizes, the method of Harris and Boyd is seemingly an ordinary normal deviate test. However, the purpose of this method is not just to perform statistical comparisons between group means but primarily to control the proportions of the subgroups outside the reference limits of the combined distribution. This basic idea was expressed by Harris *et al.* [8] as follows:

> "In our view, a clinically important difference exists [between subgroups] 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 $z^* = 3$ $(\mathrm{N}/120)^{1/2}$ was based on this consideration."

In addition to the modified normal deviate test just described, the Harris-Boyd method involves another, independent test for partitioning, which is based on comparing the standard deviations of the subgroup distributions. According to this test, subgroup-specific reference limits should be used whenever the ratio (*R*) between these standard deviations, the larger one divided by the smaller one, exceeds 1.5. Performing computer simulations, Harris and Boyd observed that one of the two gaussian subgroup distributions used by them to determine the partitioning criteria of their method, would in that case have $<1.0\%$ outside one of the common reference limits while the other distribution would have $>4.0\%$ (4). They apparently considered these proportions to be different enough from 2.5% to imply partitioning.

The value of 1.5 for *R* was in reality unnecessarily high. Calculations based on threshold values of 0.9% and 4.1% for the proportions suggest that the value of $R = 1.36$ could suffice ([2], Appendix 1). In any case, the standard deviation test should be applied independently of the modified normal deviate test. Even though the distance between the means and the value for *z* as calculated from

Eq. 1 were equal to zero, the subgroups should be partitioned if the ratio between their standard deviations exceeds 1.5 (in the present study, the original critical values suggested by Harris and Boyd will be used when assessing the performance of their method). What the locations of the means with respect to each other are is clearly immaterial to whether the subgroups should be partitioned or not, when applying the Harris-Boyd method to two subgroups.

The same is not quite true if the recommendation of Harris and Boyd concerning stratification of several subgroups, cited above, is followed. ANOVA and Tukey tests as preliminary steps of the stratification assure that the group means cannot be equal for those of the subgroups which remain to be tested using the Harris-Boyd criteria after these preliminary steps have been carried out. However, the reason why Harris and Boyd based the preliminary testing on comparisons between means, putting aside the standard deviation criterion, has probably been the fact that methods using means to compare several gaussian distributions simultaneously are readily available while methods using standard deviations for such comparisons hardly exist. As far as I can understand, the recommendation to apply ANOVA to multi-partitioning problems was not intended to turn the focus to distances between means from considering these distances as correlates to the proportions of the subgroups outside the common reference limits. Rather, those proportions probably are supposed to be the ultimate concern also in the pair-wise comparisons following the preliminary steps, because that has apparently been the primary objective of the Harris-Boyd methodology for partitioning.

For both theoretical and technical reasons, discussed in detail previously [1–3], it seems likely that the partitioning criteria of the Harris-Boyd method are not very accurate in terms of the proportions, which raises the concern whether they are accurate enough to be qualitatively useful for partitioning. One of the aims of this study was to examine the relationship between these criteria and the proportions by calculating the proportions and comparing them with the outcomes of the Harris-Boyd partitioning test.

The NCCLS guidelines on partitioning [6] include a recommendation to transform the subgroup distributions if they are "highly skewed", before the Harris-Boyd method is applied. Simple transformations, like the logarithmic one, are assumed to suffice, and if such a transformation "produces a distribution of values much closer to Gaussian form, then it is preferable to apply the *z*-test to the transformed values" [6]. The NCCLS guidelines also suggest that "the *z*-test is essentially a nonparametric test [if both subgroups have at least 60 subjects] and may be applied to the original data whether or not the values conform to a Gaussian distribution." Transforming the distributions to gaussian ones before applying the Harris-Boyd method is clearly not considered important by these guidelines, as if that method were robust against non-normality. Also, some practitioners seem to count on such robustness, because it is not unusual to see reference interval studies where the investigators do not consider the normality of the reference distributions before applying the Harris-Boyd criteria.

To examine the robustness of the Harris-Boyd method against non-normality, partitioning tests will in this study be performed using the data as untransformed (supposed to lead to such conclusions that uninformed users of that method would obtain), as logarithmic transformed (supposed to lead to such conclusions that those who follow the NCCLS guidelines could obtain), and as normalized (supposed to lead to such conclusions that the Harris-Boyd method could give under ideal circumstances, reflecting the situation where the criteria of that method were established). Because strict normalization of several subgroup distributions simultaneously, applying the same transformation to each of them, is seldom feasible, the conclusions corresponding to those ideal circumstances may be impossible to achieve, however. The reason why the same transformation should be applied to all of the distributions is that otherwise their structures with respect to each other would be changed, and the conclusions on partitioning obtained for the transformed distributions would possibly not be valid for the untransformed ones. Apart from the assumed robustness of the Harris-Boyd method against non-normality, another reason why the NCCLS guidelines expect rough normalization to suffice must be that strict normalization, although desirable, is unfeasible as a general requirement.

The partitioning method, described in more detail in a recent report [3], which will be considered as an alternative to the Harris-Boyd method in this study, is very simple: as opposed to using distances between means or ratios between standard deviations as more or less accurate correlates of the proportions, these proportions will be measured directly. Previously, a three-stage classification of proportions was recommended (see the guidelines in ref. [2] for details), but in the NORIP project, the requirements for partitioning were set rather high to avoid clinically unimportant stratifications. Hence, only two classes have been used, applied also in this study and expressed as the following rule: whenever at least one of the four proportions of two subgroup distributions outside the reference limits for their combined distribution either exceeds 4.1% or lies below 0.9%, partitioning is recommended, but otherwise the subgroup distributions can be combined. The critical proportions of 4.1% and 0.9% are slightly more stringent than those suggested by Harris and Boyd (4.0% and 1.0%), for reasons that are explained elsewhere [1 – 3].

Because the new partitioning method has been implemented for two subgroups, stratifications of the four countries will be examined by applying the partitioning tests pair-wise. Although the new method is thereby seemingly more liable to drawback 2 (presented above) than the Harris-Boyd method for several subgroups, this does not invalidate comparisons between these two methods, because when desired, ANOVA and Tukey tests can be used in the same way to limit the number of pair-wise partitioning tests in both of them.

A well-known problem inherent to stratifications into more than two subgroups is intransitivity of conclusions. The simplest form of this problem, possible in the case of three distributions, is illustrated in Figure 1.

If it happens that both pairs of adjacent distributions—in this example the pairs Norway-Sweden and Sweden-Finland—could be combined, but the pair comprising the two outermost distributions—Norway-Finland in our case—should be partitioned, the three countries could not use a common reference interval. Although Norway and Sweden could use one reference interval, and Sweden and Finland another, the situation would be problematic for Sweden, since it would have two different reference intervals to choose from. In cases like this, consistent solutions do not exist.

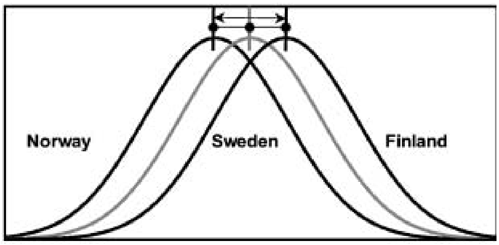

FIG. 1. Intransitivity of conclusions on partitioning. In this example, illustrating the problem of intransitivity in the case of three distributions, the distributions of Norway and Sweden are supposed to be combined with each other, and so are those of Sweden and Finland, whereas those of Norway and Finland should be separated, as suggested by partitioning tests. The three countries should not use a common reference interval, and the situation is, in addition, paradoxical for Sweden. Consistent solutions do not exist in situations like this.

MATERIALS AND METHODS

The reasons why focus was set on enzymes in the present study are: 1) As was speculated in the "Introduction", if ethnicity-related differences exist between countries, these would probably be more clear-cut for enzymes than for non-enzymes. 2) Because the distributions of enzymes are in many cases skewed, they could be appropriate to examine the sensitivity of the Harris-Boyd method to normality of the subgroup distributions. 3) Data analyses performed so far suggest that there might be differences between countries for some enzymes in particular. Notably, Danes seem to have higher levels of GGT than the other nationalities, and Finns perhaps an excess of high-value results of CK. I sought to examine whether partitioning tests support these ideas. The data used in this study are available on the web page of the NORIP project [9]. Collection of samples, laboratory analyses, and post-analytical treatment of the data will be documented in other reports dedicated to that project and published in *SJCLI*.

Iceland was excluded from the present study because of small sample sizes that varied between 0 and 85 (these numbers reflect the size of the country much more than the population-related contribution of the Icelanders to the project, however!). If any of the remaining four countries had collected fewer than 120 reference values for an enzyme, that enzyme was excluded. Age, gender, and laboratory analytical methods were considered as major confounding factors, and to control the effects of these on the enzyme concentrations, the analysis of covariance was used, as implemented in the GLM procedure of the SAS 8.2 statistical program package. Only those enzymes which showed non-significant levels (>0.05) for each of the interaction terms between the three confounding factors and the categorical variable representing country were accepted. The aim of this preliminary step was to ensure that any differences between the four countries found in subsequent partitioning tests could to as great an extent as possible be attributed to true differences between these countries. After exclusion of the three interaction terms from the statistical model, a general ANOVA was performed using the GLM procedure of the SAS 8.2 program, followed by Tukey tests.

Normality of the country-specific distributions for each enzyme was assessed by using conventional normality tests, primarily that of Anderson-Darling, and by observing the values for the skewness and the kurtosis of these distributions, as calculated by the RefVal 4.0 program [10]. In addition to the original distributions, natural logarithmic transformations of these were considered, and because neither the original distributions nor the logarithmic transformed ones were all gaussian for any of the enzymes, attempts to normalize all of the distributions simultaneously were pursued using the RefVal program, by varying the parameters of the exponential and modulus functions [11] used in that program to normalize distributions, until each distribution passed the Anderson-Darling test. If this task seemed impossible to accomplish or was perhaps of minor importance for the conclusions on partitioning, the highest degree of simultaneous normalization obtained using reasonable efforts was accepted. Normality tests and calculation of the moments were performed for most of the distributions also using the UNIVARIATE procedure of the SAS 8.2 program, which gave without any exceptions essentially the same results as RefVal.

To estimate the proportions of the country-specific distributions outside the reference limits of the combined distribution in each pair-wise comparison, the RefVal 4.0 program was slightly modified, using the Object Pascal programming language of the Borland Delphi 6 Professional package. The modified RefVal, called “partitioning program” in this report, reads in two vectors of real numbers, estimates the reference limits of the combined distribution non-parametrically, using existing subroutines of RefVal, and calculates the proportions corresponding to these limits from the data vectors. Ranks (r) were expressed as $r = p^*(n+1)$, where p is the desired proportion and n is the sample size. When appropriate, fractional ranks were calculated using linear interpolation.

Because the numbers of reference values obtained in the NORIP project from the four countries have in the cases of all the enzymes examined ratios that differ from those between the populations in the respective countries, the distributions should be weighted for partitioning calculations by using appropriate factors so as to make these two ratios equal for each pair of countries. In previous reports [2, 3], the importance of observing the prevalences of the populations underlying the subgroups for drawing correct conclusions on partitioning was discussed, and one aim of this study was to obtain further elucidation on the practical significance of this phenomenon. Hence, the proportions were calculated in two ways, both weighting the distributions (“prevalence-adjusted proportions”) and not weighting them (“proportions calculated in the usual way”).

The following population sizes were used to calculate the weight factors: 5.3 million for Denmark, 5.1 million for Finland, 4.5 million for Norway, and 8.9 million for Sweden. To exemplify, the NORIP project has obtained 661 reference values for CK from Norway and 242 from Sweden. Because $661{:}242 = 2.73$, these sample sizes are far from being representative for the respective population sizes, which have a ratio of $4.5{:}8.9 = 0.51$. The inconsistency between these two ratios does not invalidate calculation of country-specific reference limits, but whenever those of the combined distribution need to be calculated, as is the case when performing partitioning tests, the sample sizes should be scaled so that they have the same ratio as those of the underlying populations.

Otherwise the conclusions will be true for the samples—which vary more or less accidentally between laboratory tests and from one reference interval study to another—but not necessarily for the populations, while the primary objective of a reference interval study should be to obtain conclusions that are applicable to the populations. To perform a proper scaling of the data vectors in this example, the distribution of Sweden could be multiplied by (8.9/4.5) * (661/242) = 5.402 or the distribution of Norway by the inverse of this value, 0.185. However, to preclude non-integer reference values, the multiplying factors should be integers, because when multiplying a distribution, each of its reference values will be multiplied by such a factor. Hence, to obtain an adjusted ratio of 0.185 between the distributions of Norway and Sweden, a factor of 185 could be used for the distribution of Norway and a factor of 1000 for that of Sweden. This is the basic idea of the "multiplication method" to take account of prevalences, described in more detail elsewhere [3] and implemented in the partitioning program (see Note).

Another source of error that earlier partitioning methods seem unable to cope with but that is accounted for in the partitioning program used in the present study is tied reference values. If there are several equal reference values, called "tied observations" by statisticians, at a common reference limit, and both subgroups have copies of that particular reference value (one of the subgroups must have at least one copy and the other subgroup at least two copies), it is impossible to know, which one of the copies within each subgroup should be selected to calculate the proportions. The standard way in non-parametric statistics to treat arrays of tied observations is to select the one lying at the midpoint of such an array. However, this solution is unsatisfactory for partitioning purposes, because in a case where both subgroups have a large array of tied reference values and the midpoints of these would give large proportions, the standard solution would lead to a conclusion that those subgroups should be partitioned, while in reality they could be identical. A couple of suggestions to solve the problem posed by tied reference values were presented in a previous study [3]. The partitioning program uses the one of these solutions, which divides the arrays of tied reference values within the subgroups using the same ratio as the common reference limit divides that array in the combined distribution. The effect of varying the precision of reference values, leading to different numbers of tied observations in the data vectors, was examined by performing the partitioning tests on the same data expressed as rounded to various numbers of decimals.

RESULTS

The main results of the present study are summarized in Tables I–IV. Three enzymes passed the exclusion criteria, described above: ALP, CK, and GGT. Each of Tables I–III presents the results given by the partitioning tests for one of these enzymes, and Table IV shows such results for untransformed data of ALP when the calculations were performed using those data as rounded to two decimals, one decimal, or to integers.

Each of Tables I–III has three horizontal blocks, one showing the results obtained for untransformed data, one for logarithmic transformed data, and one for normalized data. Every country-specific distribution in the block for normalized data is not necessarily gaussian (cf. "Materials and Methods"), and if this is the case, the word "*normalized*" was set between quotation marks. The distributions are described in the first vertical block, titled "*Distributions*" (title text used in Tables I–IV will be cited in italics). The moments of the transformed distributions are given as calculated from non-standardized data. A significant level (<0.05) obtained from an Anderson-Darling normality test ("*A.–D.*"), suggesting that the tested distribution is non-gaussian, is marked by using a gray background color for the cell of that distribution. Ideally, there should be no gray cells in the block "*Distributions*" for normalized data, and the values for the skewness and the kurtosis should be as close to zero as possible for each normalized distribution.

The country-specific distributions will not be plotted, because figures showing four frequency distributions with similar means and standard deviations could be messy and useless to evaluate the situations (solving multi-partitioning problems is not a matter of intuition based on

Table I. Partitioning of NORIP reference data on ALP by countries.

						Partitioning calculations																		
						Harris-Boyd method									New method									
						Parameters[c]				Conclusions[d]					Proportions[e]									
	Distributions[a]														Usual					Prevalence-adjusted				
	D	F	N	S		z	z3	z5	sr	z3 or sr	z5 or sr	GLMT	(z3 or sr) and T	(z5 or sr) and T	Lower D1	Lower D2	Upper D1	Upper D2	Concl.	Lower D1	Lower D2	Upper D1	Upper D2	Concl.
Untransformed data																								
#	167	178	347	176	D-F[b]	1.92	3.60	6.00	1.09	NP	NP	NS			2.8	2.6	2.4	2.9	NP	2.8	2.7	2.4	2.9	NP
Mean	66.8	63.3	65.2	64.9	D-N	0.99	4.39	7.32	1.11	NP	NP	NS			3.6	2.1	1.7	3.1	NP	3.4	1.9	1.8	3.4	NP
St. dev.	16.5	17.9	18.3	15.5	D-S	1.12	3.59	5.99	1.06	NP	NP	NS			3.9	1.5	2.7	2.6	NP	4.8	1.6	2.7	2.6	P
A.-D.	1.00	0.02	0.00	0.00	F-N	1.18	4.44	7.40	1.02	NP	NP	NS			4.0	1.8	2.0	2.9	NP	3.4	1.7	2.2	3.4	NP
Skewn.	−0.02	0.66	1.09	0.66	F-S	0.91	3.64	6.07	1.16	NP	NP	NS			3.8	1.4	3.1	2.2	NP	4.9	1.5	3.1	2.3	P
Kurt.	0.54	0.83	2.24	0.37	N-S	0.23	4.43	7.39	1.18	NP	NP	NS			3.2	1.6	3.5	1.0	NP	3.7	2.1	4.3	1.8	P
Log-transformed data																								
#	167	178	347	176	D-F	1.96	3.60	6.00	1.02	NP	NP	NS			2.8	2.6	2.4	2.9	NP	2.8	2.7	2.4	2.9	NP
Mean	4.17	4.11	4.14	4.14	D-N	0.98	4.39	7.32	1.04	NP	NP	NS			3.6	2.1	1.7	3.1	NP	3.4	1.9	1.8	3.4	NP
St. dev.	0.28	0.29	0.27	0.24	D-S	0.79	3.59	5.99	1.18	NP	NP	NS			3.9	1.5	2.7	2.6	NP	4.8	1.6	2.7	2.6	P
A.-D.	0.00	1.00	1.00	1.00	F-N	1.32	4.44	7.40	1.07	NP	NP	NS			4.0	1.8	2.0	2.9	NP	3.4	1.7	2.1	3.4	NP
Skewn.	−1.36	−0.25	0.14	−0.04	F-S	1.34	3.64	6.07	1.21	NP	NP	NS			3.8	1.4	3.1	2.2	NP	4.9	1.5	3.1	2.3	P
Kurt.	4.17	0.25	0.28	0.12	N-S	0.14	4.43	7.39	1.13	NP	NP	NS			3.2	1.6	3.5	1.0	NP	3.7	2.1	4.3	1.8	P
“Normalized” data																								
#	167	178	347	176	D-F	2.06	3.60	6.00	1.06	NP	NP	NS			2.8	2.6	2.4	2.9	NP	2.8	2.7	2.4	2.9	NP
Mean	4.18	4.13	4.16	4.16	D-N	1.18	4.39	7.32	1.03	NP	NP	NS			3.6	2.1	1.7	3.1	NP	3.4	1.9	1.8	3.4	NP
St. dev.	0.24	0.26	0.25	0.22	D-S	1.08	3.59	5.99	1.10	NP	NP	NS			3.9	1.5	2.7	2.6	NP	4.8	1.6	2.7	2.6	P
A.-D.	0.08	1.00	0.00	0.14	F-N	1.20	4.44	7.40	1.03	NP	NP	NS			4.0	1.8	2.0	2.9	NP	3.4	1.7	2.1	3.4	NP
Skewn.	−0.56	0.17	0.49	0.28	F-S	1.12	3.64	6.07	1.17	NP	NP	NS			3.8	1.4	3.1	2.2	NP	4.9	1.5	3.1	2.3	P
Kurt.	0.97	−0.08	0.31	−0.14	N-S	0.01	4.43	7.39	1.14	NP	NP	NS			3.2	1.6	3.5	1.0	NP	3.7	2.1	4.3	1.8	P

NORIP = Nordic Reference Interval Project; ALP = alkaline phosphatase.

[a] The number of reference values, mean, standard deviation, significance level of the Anderson-Darling test, skewness, and kurtosis of each distribution.

[b] Countries used in the pair-wise partitioning tests: D = Denmark, F = Finland, N = Norway, S = Sweden.

[c] z and sr are the test parameters of the Harris-Boyd method, and z3 and z5 are different suggestions of Harris and Boyd for critical values of z.

[d] If z > z3 or z5, or the st. dev. criterion is fulfilled (see text), the conclusion is P (partitioning), otherwise NP. T denotes use of Tukey as a preliminary test.

[e] Proportions of each subgroup distribution outside the lower and the upper common reference limit (D1 = distribution 1, D2 = distribution 2).

TABLE II. Partitioning of NORIP reference data on CK by countries.

						Partitioning calculations																		
						Harris-Boyd method									New method									
						Parameters[c]				Conclusions[d]					Proportions[e]									
	Distributions[a]											GLMT			Usual					Prevalence-adjusted				
	D	F	N	S		z	z3	z5	sr	z3 or sr	z5 or sr		(z3 or sr) and T	(z5 or sr) and T	Lower		Upper			Lower		Upper		
															D1	D2	D1	D2	Concl.	D1	D2	D1	D2	Concl.
Untransformed data																								
#	384	487	661	242	D-F[b]	1.33	5.72	9.53	1.24	NP	NP	NS			3.3	2.0	1.9	3.0	NP	3.1	1.9	2.0	3.1	NP
Mean	102.4	108.6	101.0	97.8	D-N	0.35	6.26	10.44	1.00	NP	NP	NS			2.1	2.8	3.1	2.3	NP	2.3	2.8	2.9	2.1	NP
St. dev.	61.7	76.4	61.8	49.4	D-S	1.02	4.85	8.08	1.25	NP	NP	NS			2.5	2.7	3.0	1.9	NP	2.6	2.7	3.2	2.1	NP
A.-D.	0.00	0.00	0.00	0.00	F-N	1.80	6.56	10.94	1.23	NP	NP	NS			1.9	3.0	3.2	2.0	NP	2.0	3.2	3.1	1.9	NP
Skewn.	2.86	3.16	3.23	1.57	F-S	2.30	5.23	8.72	1.55	P	P	NS			2.3	2.9	3.5	0.8	P	2.3	2.9	4.2	1.6	P
Kurt.	13.11	14.02	17.78	2.90	N-S	0.80	5.82	9.70	1.25	NP	NP	NS			2.5	2.6	2.4	2.9	NP	2.6	2.6	2.4	2.6	NP
Log-transformed data																								
#	384	487	661	242	D-F	0.86	5.72	9.53	1.08	NP	NP	NS			3.3	2.0	1.9	3.0	NP	3.1	1.9	2.0	3.1	NP
Mean	4.50	4.53	4.48	4.47	D-N	0.52	6.26	10.44	1.04	NP	NP	NS			2.1	2.8	3.1	2.3	NP	2.3	2.8	2.9	2.1	NP
St. dev.	0.49	0.53	0.51	0.46	D-S	0.63	4.85	8.08	1.07	NP	NP	NS			2.5	2.7	3.0	1.9	NP	2.6	2.7	3.2	2.1	NP
A.-D.	0.03	0.00	0.00	0.14	F-N	1.48	6.56	10.94	1.04	NP	NP	NS			1.9	3.0	3.2	2.0	NP	2.0	3.2	3.1	1.9	NP
Skewn.	0.41	0.58	0.47	0.23	F-S	1.42	5.23	8.72	1.16	NP	NP	NS			2.3	2.9	3.5	0.8	P	2.3	2.9	4.2	1.6	P
Kurt.	0.64	1.04	0.70	0.00	N-S	0.22	5.82	9.70	1.12	NP	NP	NS			2.5	2.6	2.4	2.9	NP	2.6	2.6	2.4	2.6	NP
Normalized data																								
#	384	487	661	242	D-F	0.68	5.72	9.53	1.04	NP	NP	NS			3.3	2.0	1.9	3.0	NP	3.1	1.9	2.0	3.1	NP
Mean	87.1	89.7	84.5	85.1	D-N	0.75	6.26	10.44	1.04	NP	NP	NS			2.1	2.8	3.1	2.3	NP	2.3	2.8	2.9	2.1	NP
St. dev.	54.1	56.5	56.5	52.1	D-S	0.48	4.85	8.08	1.04	NP	NP	NS			2.5	2.7	3.0	1.9	NP	2.6	2.7	3.2	2.1	NP
A.-D.	1.00	0.08	0.05	1.00	F-N	1.54	6.56	10.94	1.00	NP	NP	NS			1.9	3.0	3.2	2.0	NP	2.0	3.2	3.1	1.9	NP
Skewn.	0.03	0.09	0.00	0.05	F-S	1.10	5.23	8.72	1.08	NP	NP	NS			2.3	2.9	3.5	0.8	P	2.3	2.9	4.2	1.6	P
Kurt.	−0.18	−0.18	−0.04	−0.20	N-S	0.14	5.82	9.70	1.08	NP	NP	NS			2.5	2.6	2.4	2.9	NP	2.6	2.6	2.4	2.6	NP

NORIP = Nordic Reference Interval Project; CK = creatine kinase.

[a] The number of reference values, mean, standard deviation, significance level of the Anderson-Darling test, skewness, and kurtosis of each distribution.

[b] Countries used in the pair-wise partitioning tests: D = Denmark, F = Finland, N = Norway, S = Sweden.

[c] z and sr are the test parameters of the Harris-Boyd method, and z3 and z5 are different suggestions of Harris and Boyd for critical values of z.

[d] If z > z3 or z5, or the st. dev. criterion is fulfilled (see text), the conclusion is P (partitioning), otherwise NP. T denotes use of Tukey as a preliminary test.

[e] Proportions of each subgroup distribution outside the lower and the upper common reference limit (D1 = distribution 1, D2 = distribution 2).

Table III. Partitioning of NORIP reference data on GGT by countries.

						Partitioning calculations																		
						Harris-Boyd method									New method									
						Parameters[c]				Conclusions[d]					Proportions[e]									
	Distributions[a]											GLMT			Usual					Prevalence-adjusted				
	D	F	N	S		z	z3	z5	sr	z3 or sr	z5 or sr		(z3 or sr) and T	(z5 or sr) and T	Lower		Upper			Lower		Upper		
															D1	D2	D1	D2	Concl.	D1	D2	D1	D2	Concl.
Untransformed data																								
#	234	203	515	345	D-F[b]	3.51	4.05	6.75	1.56	P	P	S	P	P	3.0	2.3	3.8	1.3	NP	3.2	2.3	4.3	1.5	P
Mean	34.2	27.1	25.3	26.6	D-N	4.91	5.30	8.84	1.55	P	P	S	P	P	2.6	2.5	5.3	1.3	P	2.6	2.5	3.8	1.1	NP
St. dev.	25.5	16.3	16.5	19.3	D-S	3.87	4.66	7.77	1.32	NP	NP	S	NP	NP	2.3	2.7	3.6	1.8	NP	2.3	2.7	3.8	1.8	NP
A.-D.	0.00	0.00	0.00	0.00	F-N	1.33	5.19	8.65	1.01	NP	NP	NS			2.1	2.7	3.4	2.3	NP	2.3	3.2	3.3	2.1	NP
Skewn.	2.79	2.68	3.46	3.60	F-S	0.32	4.53	7.56	1.18	NP	NP	NS			1.9	3.1	2.0	2.9	NP	1.9	3.1	2.0	2.9	NP
Kurt.	11.13	9.69	17.55	17.73	N-S	1.04	5.68	9.46	1.17	NP	NP	NS			1.9	3.6	1.9	3.5	NP	1.6	3.1	1.7	3.0	NP
Log-transformed data																								
#	234	203	515	345	D-F	3.40	4.05	6.75	1.23	NP	NP	S	NP	NP	3.0	2.3	3.8	1.3	NP	3.2	2.3	4.3	1.5	P
Mean	3.35	3.18	3.10	3.13	D-N	5.68	5.30	8.84	1.24	P	NP	S	P	NP	2.6	2.5	5.3	1.3	P	2.6	2.5	3.8	1.1	NP
St. dev.	0.58	0.47	0.46	0.50	D-S	4.64	4.66	7.77	1.15	NP	NP	S	NP	NP	2.3	2.7	3.6	1.8	NP	2.3	2.7	3.8	1.8	NP
A.-D.	0.00	0.00	0.00	0.00	F-N	1.90	5.19	8.65	1.01	NP	NP	NS			2.1	2.7	3.4	2.3	NP	2.3	3.2	3.3	2.1	NP
Skewn.	0.63	0.79	1.07	1.00	F-S	1.07	4.53	7.56	1.07	NP	NP	NS			1.9	3.1	2.0	2.9	NP	1.9	3.1	2.0	2.9	NP
Kurt.	0.29	0.79	1.44	1.61	N-S	0.83	5.68	9.46	1.08	NP	NP	NS			1.9	3.6	1.9	3.5	NP	1.6	3.1	1.7	3.0	NP
Normalized data																								
#	234	203	515	345	D-F	2.94	4.05	6.75	1.15	NP	NP	S	NP	NP	3.0	2.3	3.8	1.3	NP	3.2	2.3	4.3	1.5	P
Mean	3.26	3.11	3.03	3.05	D-N	5.38	5.30	8.84	1.15	P	NP	S	P	NP	2.6	2.5	5.3	1.3	P	2.6	2.5	3.8	1.1	NP
St.dev.	0.56	0.49	0.49	0.54	D-S	4.52	4.66	7.77	1.05	NP	NP	S	NP	NP	2.3	2.7	3.6	1.8	NP	2.3	2.7	3.8	1.8	NP
A.-D.	1.00	1.00	0.09	0.07	F-N	2.00	5.19	8.65	1.00	NP	NP	NS			2.1	2.7	3.4	2.3	NP	2.3	3.2	3.3	2.1	NP
Skewn.	−0.33	−0.10	0.18	−0.20	F-S	1.41	4.53	7.56	1.10	NP	NP	NS			1.9	3.1	2.0	2.9	NP	1.9	3.1	2.0	2.9	NP
Kurt.	0.06	0.09	−0.05	0.60	N-S	0.49	5.68	9.46	1.10	NP	NP	NS			1.9	3.6	1.9	3.5	NP	1.6	3.1	1.7	3.0	NP

NORIP = Nordic Reference Interval Project; GGT = γ-glutamyl transpeptidase.

[a] The number of reference values, mean, st. dev., significance level of the Anderson-Darling test, skewness, and kurtosis of each distribution.

[b] Countries used in the pair-wise partitioning tests: D = Denmark, F = Finland, N = Norway, S = Sweden.

[c] z and sr are the test parameters of the Harris-Boyd method, and z3 and z5 are different suggestions of Harris and Boyd for critical values of z.

[d] If z > z3 or z5, or the st. dev. criterion is fulfilled (see text), the conclusion is P (partitioning), otherwise NP. T denotes use of Tukey as a preliminary test.

[e] Proportions of each subgroup distribution outside the lower and the upper common reference limit (D1 = distribution 1, D2 = distribution 2).

Table IV. Partitioning of NORIP reference data on ALP by countries: Study on the effect of decimals.

						Partitioning calculations																		
						Harris-Boyd method									New method									
						Parameters[c]				Conclusions[d]					Proportions[e]									
Distributions[a]												GLMT			Usual					Prevalence-adjusted				
	D	F	N	S		z	z3	z5	sr	z3 or sr	z5 or sr		(z3 or sr) and T	(z5 or sr) and T	Lower		Upper			Lower		Upper		
															D1	D2	D1	D2	Concl.	D1	D2	D1	D2	Concl.
Untransformed data, 2 decimals: changes to the results obtained using 1 decimal																								
#					D-F[b]					NP	NP	NS					−0.1		NP					NP
Mean					D-N	−0.01				NP	NP	NS				0.1			NP					NP
St. dev.					D-S					NP	NP	NS							NP	▨				P
A.-D.		▨	▨	▨	F-N					NP	NP	NS							NP			−0.1		NP
Skewn.					F-S					NP	NP	NS							NP	▨				P
Kurt.				−0.01	N-S					NP	NP	NS							NP		−0.1	−0.1	−0.1	P
Untransformed data, 1 decimal ("basic" results)																								
#	167	178	347	176	D-F	1.92	3.60	6.00	1.09	NP	NP	NS			2.8	2.6	2.4	2.9	NP	2.8	2.7	2.4	2.9	NP
Mean	66.8	63.3	65.2	64.9	D-N	0.99	4.39	7.32	1.11	NP	NP	NS			3.6	2.1	1.7	3.1	NP	3.4	1.9	1.8	3.4	NP
St. dev.	16.5	17.9	18.3	15.5	D-S	1.12	3.59	5.99	1.06	NP	NP	NS			3.9	1.5	2.7	2.6	NP	4.8	1.6	2.7	2.6	P
A.-D.	1.00	0.02	0.00	0.00	F-N	1.18	4.44	7.40	1.02	NP	NP	NS			4.0	1.8	2.0	2.9	NP	3.4	1.7	2.2	3.4	NP
Skewn.	−0.02	0.66	1.09	0.66	F-S	0.91	3.64	6.07	1.16	NP	NP	NS			3.8	1.4	3.1	2.2	NP	4.9	1.5	3.1	2.3	P
Kurt.	0.54	0.83	2.24	0.37	N-S	0.23	4.43	7.39	1.18	NP	NP	NS			3.2	1.6	3.5	1.0	NP	3.7	2.1	4.3	1.8	P
Untransformed data, 0 decimals: changes to the results obtained using 1 decimal																								
#					D-F					NP	NP	NS					−0.1		NP		−0.1	−0.1		NP
Mean	0.1		0.1		D-N	−0.01				NP	NP	NS			−0.1	0.2			NP	−0.4	0.1	0.1		NP
St. dev.					D-S					NP	NP	NS							NP	0.6	0.1		0.1	P
A.-D.		▨	▨	▨	F-N	0.01				NP	NP	NS			−0.2	0.1			NP	0.1	0.2			NP
Skewn.		0.01			F-S					NP	NP	NS					0.1		NP	0.4		0.1	0.3	P
Kurt.			−0.03	−0.01	N-S	0.01				NP	NP	NS							NP	0.7	−0.2	−0.1	−0.1	P

NORIP = Nordic Reference Interval Project; ALP = alkaline phosphatase.

[a] The number of reference values, mean, standard deviation, significance level of the Anderson-Darling test, skewness, and kurtosis of each distribution.

[b] Countries used in the pair-wise partitioning tests: D = Denmark, F = Finland, N = Norway, S = Sweden.

[c] z and sr are the test parameters of the Harris-Boyd method, and z3 and z5 are different suggestions of Harris and Boyd for critical values of z.

[d] If z > z3 or z5, or the st. dev. criterion is fulfilled (see text), the conclusion is P (partitioning), otherwise NP. T denotes use of Tukey as a preliminary test.

[e] Proportions of each subgroup distribution outside the lower and the upper common reference limit (D1 = distribution 1, D2 = distribution 2).

visual assessment, except in extreme cases). Moreover, plots illustrating the typical profile for the distributions of each enzyme are available on the web page of the NORIP project [9], and these profiles are not hard to imagine by considering the values for skewness ("*Skewn.*") and kurtosis ("*Kurt.*") of each distribution, either. However, to give an idea of the effects of the two transformation steps used in this study to normalize distributions, the data on GGT obtained from Norway is shown as an example in Figure 2, as plotted using both the untransformed data and those data after each transformation step.

Tables I – IV have also a large vertical block titled "*Partitioning calculations*". This block is divided into two sections, titled "*Harris-Boyd method*" and "*New method*". The first column ("*z*") of the section "*Harris-Boyd method*" shows the values for the test parameter of the modified normal deviate test, as calculated from Eq. 1 using the data presented in the block "*Distributions*" (because more decimals were used in the calculations than is shown in this block, the values for test parameters given in Tables I – IV may deviate slightly from those obtained by using the shown precision). The next two columns ("*z3*" and "*z5*") give the two critical values suggested by Harris and Boyd for that test parameter, as calculated from Eqs. 2 and 3. Observe that these critical values remain the same between transformation stages (horizontal blocks), because the numbers of reference values ("#") are the only variables needed to calculate them. In contrast, the value for the test parameter *z* varies from one transformation stage to another, and the same is true for the test parameter of the standard deviation test ("*sr*" which stands for "ratio of standard deviations"). This is because the arithmetical operations needed to calculate these parameters are not invariant under the transformations. Values of *z*, which exceed at least the value of *z3*, and values of *sr*, which exceed 1.5, are marked in gray color.

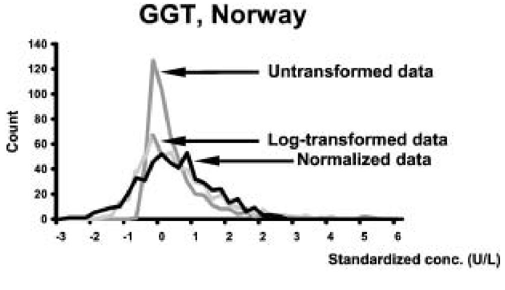

FIG. 2. Example illustrating the effect of the two transformation steps on frequency distributions. In this example, the Nordic Reference Interval Project (NORIP) data on γ-glutamyl transpeptidase (GGT) obtained from Norway (n = 661) are shown as unsmoothed frequency distributions, as calculated for untransformed, logarithmic (ln) transformed, and normalized data. The normalization was performed by transforming logarithmic transformed data further by repeatedly applying the exponential and modulus functions, as implemented in the RefVal 4.0 program. Each distribution was standardized to have mean = 0 and standard deviation = 1.

The next five columns of the section "*Harris-Boyd method*" show the conclusions on partitioning as obtained using this method. These columns, with the exception of the midmost one, as well as those of the section "*New method*" showing conclusions, have gray as their background color irrespective of what the suggested conclusions are. The first two columns in this area show the conclusions for each pair of two countries disregarding the outcomes of the general ANOVA and the subsequent Tukey tests. The critical values for the Harris-Boyd test parameters used to draw the conclusions shown in the first of these two columns are *z3* and *sr*, and those for the second column are *z5* and *sr*. If at least one of the two test parameters exceeds its column-specific critical value, the test is interpreted as supporting partitioning of the subgroups, which is indicated by "*P*", otherwise as supporting combining, or "non-partitioning", indicated by "*NP*".

The column titled "*GLMT*" shows the significance levels of the Tukey tests, classified qualitatively as either significant ("*S*") or non-significant ("*NS*") using a level of <0.05 as the requirement for significance. This column works as a filter between the first two and the last two columns of the area titled "*Conclusions*" within the section "*Harris-Boyd method*": if the significance level "*S*" has been obtained for the Tukey test of some comparison between two countries, the contents of the first two columns were copied to the last two columns on the row corresponding to that comparison. Hence, the last two columns show the conclusions as obtained when applying the method of Harris and Boyd as recommended by them for several subgroups. These columns are left empty for non-significant levels of the Tukey test, because following the instructions of Harris and Boyd, partitioning tests are then not needed, but the final conclusion is obviously "*NP*" also in such cases.

The section titled "*New method*" shows the values for the proportions of country-specific distributions outside the common reference limits obtained for each comparison between two countries, as calculated by using the partitioning program. The first five columns, titled "*Usual*", give these proportions and the conclusions suggested by them when the prevalences, i.e. the populations of the countries, were not considered, and the last five columns, titled "*Prevalence-adjusted*", give the same data after weighting the country-specific distributions so as to make the ratios between the numbers of reference values in them agree with the ratios between the populations. The proportions are listed in the following order for each of these two calculation methods: the proportion for 1) distribution 1 ("*D1*") at the lower end of the distributions, 2) distribution 2 ("*D2*") at the lower end, 3) distribution 1 at the upper end, 4) distribution 2 at the upper end, where distributions 1 and 2 refer to the distribution of the first and the second country, respectively, involved in a comparison and indicated by the initial letters of those countries in the row titles for the block "*Partitioning calculations*" (e.g. "*D–F*" as such a row title means that distribution 1 in that comparison has been Denmark and distribution 2 Finland). If a proportion is either equal to or exceeds 4.1%, or is equal to or lies below 0.9%, it is highlighted with gray color. If at least one of the four proportions obtained in a comparison is shown highlighted, the conclusion of that partitioning test is interpreted as being partitioning ("*P*"), otherwise as non-partitioning ("*NP*").

As opposed to the Harris-Boyd method, the results obtained by direct calculation of the proportions remain unchanged between transformations. This is because rank-based, nonparametrically calculated proportions depend only on the order of the reference values in the data vectors, but the transformations used in this study do not change that order. Hence, the proportions and the conclusions in the section "*New method*" are the same in each horizontal block, whereas variation between these blocks is seen in both the test parameters and the conclusions when considering the results presented in the section "*Harris-Boyd method*". However, I did not simply copy the proportions obtained for untransformed data to the horizontal blocks for transformed data but actually recalculated each proportion, for two reasons: 1) Although the order of the reference values is invariant under a transformation, the fractions obtained by using linear interpolation between reference values may vary slightly between transformations. As could be expected, this variation had a minimal effect on the proportions. One out of the 144 proportions presented in Tables I–III showed a change between transformations that affected the 1st decimal of a percentage (in the comparison F-N for ALP, the prevalence-adjusted proportion obtained for Finland at the upper end of the distributions was decreased from 2.2% to 2.1% when the data were transformed), but otherwise the changes concerned higher decimals. 2) If the proportions obtained for transformed data are not practically the same as those obtained for untransformed data, there is probably something wrong with the transformation. Recalculation of the proportions is in fact an effective way to check that each distribution has been transformed correctly, using the same transformation parameters.

Table IV, which shows the results obtained for the untransformed data of ALP as calculated using various precisions, has a slightly different structure from those of Tables I–III. The data of the three enzymes examined in this study are given with a precision of two decimals on the web page of the NORIP project (9). Because this precision is probably too high, I performed the calculations reported in Tables I–III using the precision of one decimal, although I was not sure if this precision was the "correct" one, either. However, considering such results as the "basic" ones, I copied the uppermost horizontal block of Table I, containing those basic results for untransformed ALP, and pasted that block in the second horizontal block of Table IV, titled "*Untransformed data, 1 decimal ("basic" results)*". The results obtained for the same data as calculated using the precision of two decimals and that of integers, are shown in the horizontal blocks lying above and below, respectively, the block for the basic results in Table IV.

To make the effect of using the two new precisions on the results conspicuous, the data in the uppermost and the lowermost horizontal block of Table IV are not shown in absolute terms but as differences from the corresponding basic results. To exemplify, in the uppermost horizontal block of that table, titled "*Untransformed data, 2 decimals: changes to the results obtained using 1 decimal*", the vertical block "*Distributions*" includes only one figure, −0.01, placed in the cell reserved for the kurtosis of the distribution of Sweden. Because that kurtosis as calculated using the precision of one decimal is 0.37, as shown in the horizontal block for the basic results, the same kurtosis as calculated using two decimals is $0.37-0.01=0.36$ in absolute terms. All the other cells in the vertical block "*Distributions*" of the uppermost horizontal block are empty, which means that the increased precision has not evoked any other changes to the basic results in that vertical block. The background colors of the cells in the uppermost and the lowermost horizontal blocks reflect the status of each result in absolute terms.

Next, the results presented in each of Tables I – IV are outlined.

Table I: ALP

Although only one of the untransformed distributions for ALP (that of Denmark) passed the Anderson-Darling test for normality, the distributions are only moderately skewed and peaked (consider the values for skewness and kurtosis, respectively). Logarithmic transformation reversed the situation, making the three other distributions, except that of Denmark, pass the Anderson-Darling test. Applying the exponential and modulus functions with weighted (2:1:1:1 for Denmark:Finland:Norway:Sweden) parameter values to the logarithmic transformed data, a "normalization" was achieved, which made the distribution of Norway non-gaussian but otherwise shows a better balance of normality between the distributions as compared to the logarithmic transformed data. Visually, the profiles of the distributions were only slightly changed by any of the transformations (plots not shown), and because it was obvious that no changes to the conclusions of the partitioning tests were to be expected by improving the normalization, no more efforts were made to find better values for the normalization parameters.

The results given by the Harris-Boyd method suggest that ALP could be a rather unproblematic case for partitioning. The test parameter *z* remains far from the critical value *z3* (and obviously even farther from *z5*) for every comparison between two countries no matter which of the three transformation stages (untransformed, log-transformed, "normalized") is considered. Also, each value for *sr* (ratio of standard deviations) lies rather close to unity, the highest value being 1.21. Hence, the pair-wise comparisons lead to the conclusion "*NP*" without any exceptions, and because each of the Tukey tests gave a non-significant result, it is according to the Harris-Boyd method for partitioning of several subgroups not necessary even to consider the results of those pair-wise comparisons. The overall conclusion is that the Harris-Boyd method does not detect any differences between the four countries.

The situation is not quite that simple if the proportions are estimated directly ("*New method*"), although none of these proportions as calculated without considering the prevalences ("usual" proportions for short in what follows) reaches critical levels (4.1% and 0.9%), and the conclusions based on them agree with those suggested by the Harris-Boyd method. However, a prevalence-adjusted proportion exceeding 4.1% was obtained for three comparisons (D-S, F-S, and N-S). Each of these critically high proportions should have been discovered also by the Harris-Boyd method as applied pair-wise, because using ANOVA and Tukey tests as preliminary steps does not mean that the results obtained for pair-wise comparisons should not be consistent with the objectives of that method.

Before any conclusions concerning the four countries as a whole can be drawn, each result suggesting partitioning must be examined in detail. In the comparison D-S, a proportion of 4.8% was obtained for Denmark at the lower end of the distributions. This means that if the lower common reference limit calculated for Denmark and Sweden were used in Denmark, as much as 4.8% of the Danes would be classified as having unusually low levels of ALP instead of the expected 2.5%. But because low levels of ALP are of limited clinical interest, this finding has hardly any consequences when assessing the usefulness of such a lower common reference limit for Denmark. The same conclusion is valid for the comparison F-S, where Finland has the same role as Denmark has in the comparison D-S. In the comparison N-S, a proportion of 4.3% was obtained for Norway at the upper end of the distributions, suggesting that if the upper common reference limit of Norway and Sweden were used in Norway, 4.3% of the Norwegians would be considered as having unusually high levels of ALP, 42% of these erroneously. Such a situation could have important consequences for the healthcare system in Norway, including economic ones, if it were the final conclusion. However, because the levels of ALP in Norway do not deviate from those in Denmark and Finland, the difference observed between the distributions of Norway and Sweden does alone not suffice for a recommendation to be made to introduce country-specific reference limits in Norway.

Table II: CK

The untransformed distributions of CK are heavily skewed and peaked. The logarithmic transformation removed most of the skewness and kurtosis from each distribution, but three of these did not pass the Anderson-Darling test. A slightly better normalization was achieved by application of the exponential and modulus functions to the untransformed data.

The *z* values of the Harris-Boyd method lie substantially below their respective critical values in each comparison between two countries, while the value of *sr* exceeds 1.5 in the comparison F-S for untransformed data. Because the Tukey test gave a non-significant result for each comparison irrespective of the transformation stage, the final conclusion suggested by the Harris-Boyd method for CK is that there are no differences between the four countries.

The comparison F-S is the only one, which is highlighted also by the new method. Both the "usual" and the prevalence-adjusted proportions suggest the conclusion to partition to be made for this comparison. However, while the reason for that conclusion is a critically low "usual" proportion obtained for Sweden (0.8%) at the upper end of the distributions, the prevalence-adjusted proportion being highlighted is a critically high proportion calculated for Finland (4.2%) at the same end. Because of the larger population of Sweden (8.9 million) as compared with that of Finland (5.1 million), an adjustment made in the calculations to account for the populations has transferred the upper common reference limit toward lower levels of ALP, reflecting the situation for ALP in Sweden, which has led to an increase in the proportion obtained for each distribution at the upper end. This has made the proportion of Sweden grow from a level lying below 0.9% to a level lying above that percentage, changing the conclusion from "*P*" to "*NP*", and the proportion of Finland from a level lying below 4.1% to a level lying above that percentage, changing the conclusion in the opposite direction. Hence, both types of proportion, the "usual" and the prevalence-adjusted ones, accidentally suggest the same conclusion in this particular case, but the results obtained by using prevalence-adjusted proportions should in general be preferred.

Putting aside the filtering activity of the Tukey tests, the Harris-Boyd method and the new method suggest seemingly the same conclusion for the comparison F-S. However, as opposed to the new method, giving the same results for each of the three transformation stages, the conclusion suggested by the Harris-Boyd method is changed from "*P*" to "*NP*" when the distributions are transformed. This happens because the value of *sr* for the comparison F-S is decreased from 1.55 (untransformed data) to 1.16 (logarithmic transformed data) and further to 1.08 (normalized data) under normalization. Hence, the critical level of 1.5 for *sr* is far from being reached when considering the situation of the comparison F-S for normalized data. Because the Harris-Boyd method was created using gaussian distributions, the conclusion, which should be considered as representative for that method, is apparently the one obtained from the normalized data, i.e. "*NP*". Hence, once again the observation is made that the Harris-Boyd method as applied to two distributions in the way it should be applied, is unable to discover a proportion that would require a recommendation to partition to be made.

Table III: GGT

The untransformed distributions of GGT are perhaps even more skewed and peaked on the average than those of CK, and they seem to be more non-gaussian on the average also after the logarithmic transformation than the distributions of CK are after the same transformation. None of the four logarithmic transformed distributions of GGT passed the Anderson-Darling test, and some effort was needed to make all of them pass that test simultaneously (normalized distributions). Starting from the logarithmic transformed data, the exponential and modulus functions were applied twice with appropriately weighted parameter values. Although each of these normalized distributions passed the Anderson-Darling test in the end, they can hardly be considered strictly gaussian (Fig. 2) in the same way as simulated gaussian distributions, used by Harris and Boyd when they created their method [4].

The Harris-Boyd method suggests partitioning for the comparisons D-F and D-N when applied to untransformed data, due to high values of *sr* calculated in these comparisons. Because the Tukey tests gave significant outcomes for the comparisons D-F, D-N, and D-S, which include those two comparisons, these suggestions are also the final conclusions obtained from untransformed data by using the Harris-Boyd method. When the distributions were transformed, the high values of *sr* disappeared, but the *z* value for the comparison D-N was increased above *z3*. The Tukey tests remained significant for the three comparisons just mentioned at each transformation stage, which means that the final conclusion suggested by the Harris-Boyd method is partitioning for the comparison D-N. However, this conclusion is valid only when applying the critical value *z3*. Those who would prefer following the new recommendation of Harris and Boyd to use *z5* as the critical value for *z*, would obtain the conclusion "*NP*" for every pair-wise comparison, including D-N.

Nothing of this diversity of conclusions, varying between transformation stages and different suggestions for the threshold values of the test parameters, is seen when the new method is applied (apart from differences between "usual" and prevalence-adjusted proportions, of course, but the proportions of the former type are presented only to illustrate the importance of making an adjustment for prevalences). The "usual" proportion of Denmark exceeds 4.1% at the upper end of the distributions in the comparison D-N, but the conclusion "*P*" for this comparison is changed to "*NP*" when the population sizes are taken account of. The only comparison leading to the conclusion "*P*" when using the prevalence-adjusted proportions is the comparison D-F.

Table IV: Study on the effects of varying the precision of the data for ALP

As expected, the precision of data has minimal effects on the values of the test parameters used in the Harris-Boyd method. This is because the means and the standard deviations remain approximately the same regardless of with how many decimals—2, 1, or 0—the reference values are expressed. As shown in Table IV, some of the *z* values changed by 0.01 in either positive or negative direction, but the other test parameter values remained the same. The Harris-Boyd method is apparently not particularly sensitive to the precision of reference values, but how this seemingly positive property of that method should actually be interpreted is an intricate issue that I will return to in the "Discussion".

In contrast to the test parameters of the Harris-Boyd method, the proportions may show considerable changes as the precision of the reference values is varied. The changes are small—0.1% at most (in absolute terms)—if the precision is increased from one decimal to two decimals, but if it is decreased from one decimal to that of integers, changes extending up to 0.7% are seen. In one case (prevalence-adjusted proportion at the lower end of the distribution of Norway in the comparison N-S) the change in the proportion would have changed the conclusion (from "*NP*" to "*P*"), had there not been another proportion exceeding 4.1% in the same comparison.

DISCUSSION

The present study had three major objectives: 1) to discuss stratification of reference data into several subgroups as a general problem; 2) to compare the partitioning method of Harris and Boyd with the new method, both as far as pair-wise comparisons and multi-partitioning problems are concerned; 3) to solve some practical partitioning problems raised by preliminary results of the NORIP project. Intending to illustrate new methodology for partitioning, it was necessary to describe both the background and interpretation of the results in some detail. Hence, because of limited time and space, I restricted the scope of this study to the enzymes rather than trying to perform a comprehensive study on all of the biochemical markers involved in the NORIP project.

The three enzymes that passed the inclusion criteria for this study were not pre-examined to find ideal candidates to illustrate the weaknesses of the Harris-Boyd method, suspected in earlier theoretical studies [1–3], and yet those weaknesses are apparent for each of them. To recapitulate, 1) the Harris-Boyd method failed to discover three critically high proportions for ALP in pair-wise comparisons. Although two of these proportions were as high as close to 5.0%, the values for the test parameters of the Harris-Boyd method, designed to detect proportions exceeding 4.0%, lay far from their respective threshold levels in these comparisons (Table I). Hence, the failure of the Harris-Boyd method to highlight these three comparisons cannot be ascribed to considering them as borderline cases, remained undetected because of slightly erroneous approximations. 2) It discovered one critical proportion for CK in a pair-wise comparison when untransformed data, being strongly non-gaussian, was used, but not when it should have discovered that proportion, i.e. when normalized data were used. 3) It discovered one critically high proportion for GGT, but the conclusion was dependent on which critical level of the *z* parameter was selected. Moreover, that proportion turned out to be non-critical when the prevalences were considered, while another critically high prevalence-adjusted proportion remained unrecognized.

It should not be surprising if proportions measured directly deviate from those measured indirectly. The idea of using easy-to-calculate test parameters, such as those of the Harris-Boyd method, as correlates of the proportions, is to make partitioning convenient in practice, and a certain degree of inaccuracy must be accepted as a price for that convenience. However, the correlation between the values of the Harris-Boyd test parameters and the proportions seems in these case studies to be so poor, not only quantitatively, but also in qualitative terms, that methodological problems are a more likely explanation than just inaccuracy of approximations.

The fundamental limitation in the Harris-Boyd method, which could explain a major part of its poor performance in the present study, as evaluated by using prevalence-adjusted proportions as a reference, is that the critical values for its test parameters were established only for a specific ratio between the numbers of reference values in the subgroups, but because the proportions vary with those numbers, the same critical values are not valid as soon as that ratio is changed. On p. 266 of their publication [4], which describes the selection of these critical values, Harris and Boyd wrote, referring to sample sizes, that "N is the number of subjects in each subgroup". This shows that the same number of reference values was used in both subgroups in their study. Hence, it seems that the Harris-Boyd method was unintentionally designed to work in the particular situation where the prevalences of the subgroups are equal. In all other situations, including these case studies, its usefulness may be limited.

In terms of the prevalences, the results of Tables I–III could be interpreted as:

1. The conclusions of the Harris-Boyd method reflect (probably a rough approximation of) a situation where the populations would be equal in the four countries, i.e. where the distributions would have weights of 1:1:1:1 with respect to each other. These conclusions are not reliable, because the Nordic countries do not have equal populations.
2. The conclusions suggested by the "usual" proportions reflect the sample sizes, i.e. a situation where the investigator puts together the data from the two subgroups in each pair-wise comparison without taking account of their prevalences. As an example, the sample sizes for ALP were 167 for Denmark, 178 for Finland, 347 for Norway, and 176 for Sweden, giving weights of 167:178:347:176 = 1:1.07:2.08:1.05 for these countries in the partitioning calculations for ALP. Because the sample sizes vary between enzymes (Tables I–III), another set of weights has been used in the calculations for CK and again another in those for GGT. Each of these sets of weights is useless, however, because none of them reflects the ratios between the four populations. Therefore, these conclusions should not be considered as reliable, either.

3. The conclusions obtained by using prevalence-adjusted proportions reflect the population sizes in the four countries, giving weights of 5.3:5.1:4.5:8.9 = 1:0.96:0.85:1.68 for the distributions. As opposed to the other two types of conclusions, these are valid for the situation, which they are intended to be valid for.

A comparison between these three types of conclusions in each of Tables I – III shows in concrete terms the consequences of using various weights for the four distributions and illustrates the importance of adequate handling of the prevalences, discussed in previous publications [2, 3].

Apart from inability to account for prevalences, there seem to be other concerns that users of the Harris-Boyd method should observe. One of these is its possible sensitivity to non-normality of the subgroup distributions. Consider the three examples of Tables I – III. ALP (Table I) is a straightforward case for the Harris-Boyd method because the levels of the test parameters are so low in each of the six comparisons that they do not come close to the threshold levels in any of them, and transforming the distributions does not change this situation. In contrast, for both CK (Table II) and GGT (Table III), a different set of conclusions was obtained when logarithmic transformation was performed on the original data. The conclusions were changed in neither of these cases between the logarithmic transformed and the normalized data. However, note how much closer to its critical level (5.30) the z value has come to lie in the comparison D-N of GGT when using normalized data (5.38) than when using logarithmic transformed data (5.68). If it had been possible to normalize the distributions perfectly, making them comparable to the simulated gaussian distributions used by Harris and Boyd to construct their method (4), a third set of conclusions would perhaps have been obtained for GGT, different from each of those obtained using the data of the other two transformation stages.

These examples suggest that the Harris-Boyd method may be more sensitive to non-normality than some of its users expect. Because strict normalization of more than one distribution simultaneously is in most cases impossible, uncertainty about what the representative conclusions of that method would have been if the distributions had been strictly gaussian cannot be removed in borderline cases, such as that of GGT in the present study. In contrast, the conclusions obtained by using the new method remain unchanged under transformations (with the reservation made for fractions, discussed above, but this reservation is relevant only for transformed data and even then in most cases is unimportant in qualitative terms), making a similar concern about the forms of the distributions unnecessary.

One more drawback of the Harris-Boyd method is its inability to cope with tied reference values. To understand how such values may affect the proportions, consider the results obtained for untransformed ALP by expressing the reference values in various numbers of decimals (Table IV). These results show that the largest change of a proportion, 0.7%, was obtained in the comparison N-S for Norway at the lower end of the distributions when the precision was reduced from one decimal to that of integers. The beginning of the sorted data vector for Norway as expressed in one decimal looks like this:

27.2 31.2 32.5 33.1 33.2 35.8 36.4
36.6 37.5 38.3 38.7 39.1 39.8...

The prevalence-adjusted lower common reference limit, which lies at 39.6 (as calculated by using one decimal), has a unique position within this vector. Lying between the reference values of 39.1 and 39.8, it gives a proportion of 3.7%, shown in Table IV. Because each of the reference values listed above is unique at the precision of one decimal, the same must be true for the precision of two decimals. Hence, the same proportion was obtained also when using that precision (that there was no change is indicated by an empty cell in Table IV).

However, the situation is quite different if the same data are considered as rounded to integers. In this case, the beginning of the data vector for Norway looks like this (more reference values are listed than for the precision of one decimal):

27 31 33 33 33 36 36 37 38 38
39 39 40 40 40 40 40 40 42...

The lower common reference limit as calculated using integers has the value of 40. Because there are six copies of that value in the data vector for Norway and also two copies (data not shown) of it in the data vector for Sweden, there is no way to know, which of the six copies should be selected to calculate the proportion for the distribution of Norway. Because that proportion as calculated using different copies varies between 3.7% and 5.2%, and this interval includes the critical proportion of 4.1%, it may be decisive for the conclusion on partitioning how the uncertainty due to tied observations is handled in this case. It was proposed recently [3] that the arrays of tied reference values in the subgroups could be divided using the same ratio, as the common reference limit divides that array in the combined distribution. In this example, that ratio was 0.45, and using this value, the proportion of 4.4% was obtained for Norway (Table IV).

Which one of the calculated proportions, 3.7% or 4.4%, is correct, depends on what the true precision of the data is. If this precision is that of one decimal or two decimals, 3.7% should be considered as the correct proportion, whereas if the true precision is that of integers, 4.4% should be preferred although it is a compromise value selected from a wide range of proportions. It may appear unacceptable that a decision on partitioning could depend on such a trifling matter as the precision of the reference values. However, a clinical chemist should respect the data he has. If the true precision for ALP is that of integers and the data collected from Norway in a carefully performed reference interval study show six tied observations of 40, then those six observations are a fact that cannot be handled successfully by manipulating the data. As has been shown previously [3], expressing the reference values with unrealistically high precision or adding random noise to them so as to get a solution that seems easier to accept, would probably be a mistake.

I did not check what the correct number of decimals justified by the analytical quality achieved in the NORIP project should be for each of the three enzymes included in the present study, because it turned out that the results of the partitioning tests remained qualitatively the same for each enzyme although the proportions varied with the precision. However, that the conclusions did not change in any of the cases was just good luck. In general, one should use the true precision, not a single decimal more or less than that, of the reference values in partitioning calculations. With how many decimals the reference limits are reported to the clinicians is another question, but such considerations should not affect the precision selected for the partitioning calculations.

As was reported above, the Harris-Boyd method was insensitive to the precision of the reference values (Table IV). Because there are no tools to detect tied observations in this method, those observations and the additional uncertainty introduced by them to the values of the proportions remain unrecognized by it. That a problem remains unrecognized does not mean that it does not exist, however. The stability of the Harris-Boyd method when varying the number of decimals is therefore not an advantage, but rather one more contribution to its inability to establish a realistic correlation between the values for its test parameters and the actual proportions.

Because ANOVA and subsequent pair-wise tests like the Tukey test are designed to detect differences between group means, they have the same weakness as the Harris-Boyd method of being unable to account for prevalences and tied observations. Hence, they are in the general case hardly useful for such a filtering activity as suggested by Harris & Boyd [5]. In the present study (Tables I – III) this filtering activity changed the conclusion of a pair-wise comparison once. In the comparison F-S for untransformed data of CK, the conclusion obtained by using the Harris-Boyd method changed from "*P*" to "*NP*", but that change was probably erroneous, as suggested by the high proportion (4.2%) calculated for Finland at the upper end of the distributions.

If the Harris-Boyd method is applied to solve a multi-partitioning problem and the conclusion is partitioning for several pair-wise comparisons, there is a considerable risk that one will be faced with the problem of intransitivity, illustrated in Figure 1. This is because this method gives no information beyond the simple fact that there is a difference between each pair of two distributions in those comparisons. In contrast, the new method provides the investigator with information about at which end(s) of which distribution(s) the reason(s) for the conclusion to partition lie(s). This may help to clarify the situation, e.g. by showing that some of the suggestions would not have much clinical importance, as was illustrated by the case of ALP (cf. "Results").

Because I do not know of any statistical method, analogous to ANOVA, that could be used to assess differences between proportions among several groups simultaneously, and because ANOVA seems to be useless for that purpose, performing the partitioning tests pair-wise may be the only possibility for using the new method at this moment. The results that I consider as the most reliable ones supplied by this study, those presented in the last column of each of Tables I – III, probably are to some extent susceptible to drawback 2 discussed above in "General considerations", although the risk of a false decision to partition may be hard to quantify. However, I feel that those results make educated guesses feasible about the two concerns raised by previous analyses of the enzyme data of the NORIP project. First of these was that Denmark seemed to have higher levels of GGT than the other countries. This concern is likely to be unnecessary, because only one comparison (D-F) of those performed using the data on GGT led to partitioning as the conclusion (Table III). Hence, the distributions of GGT appear reasonably similar between the four countries, as evaluated using the suggested partitioning criteria. The same conclusion could be valid also for the other concern, which was that the levels of CK at the upper end of the distributions appeared higher for Finland than for the other countries, because there was only one comparison (F-S) to support such an idea (Table II). The third enzyme included in this study, ALP, had three comparisons that suggested partitioning (D-S, F-S, and N-S). However, because two of these involved the lower end of the distributions and that end has limited interest for clinical practice, the situation for ALP (Table I) is essentially similar to those for GGT and CK.

Overall, the results of the present study suggest that common reference intervals should be applicable in each of the four Nordic countries as far as the examined three enzymes are concerned. Because neither CK nor GGT—which in the preliminary data analyses had raised the question of whether application of common reference intervals was feasible in their cases—showed major overall differences between the four countries, it seems reasonable to assume that such differences will possibly not be found in future calculations for any of the other tests involved in the NORIP project. Although it is obviously risky to extend the results obtained in the present case studies to the whole project, my working hypothesis for those future calculations is that the assumption made by the initiators and supporters of this project concerning the usefulness of common reference intervals in the Nordic countries has been correct.

At the same time another important conclusion needs to be drawn, however: the methodology that is available today to examine multi-partitioning problems is unsatisfactory. For reasons that have been elucidated in the present study, ANOVA and other methods based on comparisons between group means are of limited usefulness for partitioning purposes, and calculating proportions pair-wise to two distributions at a time is not an ideal method to resolve complex situations, either. Interest in such methodology will possibly increase in the near future, because large-scale reference interval studies similar to the NORIP project are being performed and also planned elsewhere, following the idea that laboratories in several countries could share the effort required to produce reference intervals, both to reduce the costs for each laboratory and to improve the statistical quality of the results. It is important in such international joint projects to ensure that the reference limits produced are applicable and useful in each of the participating countries. A decision made by the clinical chemists in a country to recommend reference limits, which would classify, say, 1% of the tested population more than is considered acceptable, as having a false-positive result, could be an expensive decision for that country, depending on the test and the clinical conclusions drawn from those false results. Country-specific reference limits might be preferable in such a situation, but that country would then lose the advantages of sharing common reference limits within a large geographical area. To make the choice, clinical chemists should have reliable information about the behavior of reference distributions as compared to each other in the partitioning tests.

CONCLUSIONS

A. Comparison between the properties of the Harris-Boyd method and those of the new method, as suggested by previous theoretical publications and illustrated by the results of the present study, could be summarized as:

1. The Harris-Boyd method was developed using simulated gaussian distributions, which have probably been strictly normalized. The method is more sensitive to non-normality of the subgroups than some users expect and perhaps even more sensitive than the guidelines of the NCCLS on partitioning suggest. To obtain a conclusion that is representative for the performance of the Harris-Boyd method, particularly in borderline cases, each subgroup should be strictly normalized, but normalization of several distributions simultaneously by applying the same transformation to them is in many cases impossible to achieve. In contrast, normality of the distributions is not a concern when using the new method because it gives the same results when applied to the original data vectors as when applied to transformed ones.
2. The Harris-Boyd method was developed using equal numbers of reference values in the subgroups, which means that the relationship between the proportions and the recommended partitioning criteria was established and is valid for such a specific situation where the prevalences of the subgroup populations are equal. In any other situation the usefulness of those criteria may be questionable. While adjustment for prevalences would be difficult with the Harris-Boyd method, they can be coped with in several ways using the new method [3].
3. Using quotients calculated from means and standard deviations as test parameters, the Harris-Boyd method is unable to detect and take account of tied reference values. Because such values may affect the proportions considerably, this inability contributes further to impair the relationship between the proportions and the Harris-Boyd partitioning criteria. In contrast, focusing on the proportions directly, the new method is readily able to deal with tied reference values in an appropriate way.
4. Because the drawbacks of the Harris-Boyd method described in items 2 and 3 above are also drawbacks of ANOVA, the usefulness of ANOVA for stratifying reference data is probably limited. Nor is applying the new method pair-wise hardly an ideal way to solve multi-partitioning problems, and more sophisticated methods, based on estimation of the proportions rather than on distances between means, could be developed.

B. The results obtained for the three enzymes involved in the present study suggest that there are no important differences between reference distributions among the four Nordic counties for those enzymes, and to use common reference intervals for each of them in these countries seems feasible. Because the enzymes showed the largest differences between countries among the analytical tests included in the NORIP project, this conclusion can possibly be extended to cover also the other tests not examined in this study.

ACKNOWLEDGEMENTS

I gratefully acknowledge the support for this study given by the Department of Clinical Chemistry at Rikshospitalet University Hospital of Oslo and the Nordic Reference Interval Project.

NOTE

A couple of technicalities should be borne in mind when operating with multiplied data vectors. A multiplied distribution is supposed to have exactly the same structure as the original one, i.e. at each reference value, both distributions should give the same proportions. If the proportions in the original distribution are calculated as $p = r/(n+1)$, they should in the multiplied distribution be calculated as $p_m = r_m/(n_m + F)$, where F is the multiplier, and $r_m = F*r$ and $n_m = F*n$ are the rank and the number of reference values in the multiplied distribution, respectively, because otherwise p_m would not be equal to p. Another thing to remember is that the interspaces between the sets of F equal reference values in a multiplied vector should be considered as having been multiplied by F. If fewer than F rank numbers are missing to reach a common reference limit from a certain reference value, one should not advance to the next set but calculate a fraction between the sets by using interpolation.

REFERENCES

1 Lahti A, Hyltoft Petersen P, Boyd JC, Fraser CG, Jørgensen N. Objective criteria for partitioning gaussian-distributed reference values into subgroups. Clin Chem 2002; 48: 338–52.

2 Lahti A, Hyltoft Petersen P, Boyd JC. Impact of subgroup prevalences on partitioning gaussian-distributed reference values. Clin Chem 2002; 48: 1987–99.

3 Lahti A, Hyltoft Petersen P, Boyd JC, Rustad P, Laake P, Solberg HE. Partitioning of non-gaussian distributed biochemical reference data into subgroups. Clin Chem 2004; 50: 891–900.

4 Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranges. Clin Chem 1990; 36: 265–70.

5 Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York, Basel, Hong Kong: Marcel Dekker, Inc.; 1995

6 NCCLS. How to define and determine reference intervals in the clinical laboratory; approved guideline. 2nd ed. NCCLS document C28-A2. NCCLS, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087–1898 USA, 2000.

7 Virtanen A, Kairisto V, Irjala K, Rajamäki A, Uusipaikka E. Regression-based reference limits and their reliability: example on hemoglobin during the first year of life. Clin Chem 1998; 44: 327–35.

8 Harris EK, Wong ET, Shaw Jr ST. Statistical criteria for separate reference intervals: race and gender groups in creatine kinase. Clin Chem 1991; 37/9: 1580–2.

9 Web page of the Nordic Reference Interval Project (NORIP) on the Internet: http://www.furst.no/norip/. Accessed 9/30/03.

10 Solberg HE. RefVal: a program implementing the recommendations of the International Federation of Clinical Chemistry on the statistical treatment of reference values. Comput Meth Progr Biomed 1995; 48: 247–56.

11 Solberg HE. Approved recommendation on the theory of reference values (1987). Part 5. Statistical treatment of collected reference values. Determination of reference limits. J Clin Chem Clin Biochem 1987; 25: 645–56.

Received: 2 October 2003
Accepted: 4 November 2003