Estimation of Reference Ranges: How Many Subjects are Needed?

John A. Lott,¹ Lynn C. Mitchell,² Melvin L. Moeschberger,² and Donald E. Sutherland²

We measured Na, K, Cl, glucose, hemoglobin, erythrocytes, and hematocrit in the serum or blood from ~800 male and 200 female second-year medical students in an effort to define the size of an acceptable reference population. Using the data from the men and Monte Carlo simulations of 10–400 samples, each carried out 5000 times, we found that for most of the above tests, ~200 people are required for stable lower (2.5%) and upper (97.5%) reference limits to be obtained. This agrees with the 198 subjects required by strictly statistical criteria to define the same limits with a 99% confidence level.

Additional Keyphrases: nonparametric statistics · reference values · sodium · potassium · chloride · glucose · hemoglobin · erythrocytes · hematocrit · hypoglycemia

Reference values must be established for new or modified clinical laboratory tests. Superficially, it is a simple problem; actually, "it is one of the most stubborn and difficult problems limiting the usefulness of clinical laboratory data" (1). Sunderland (2) listed the many variables affecting reference values; if we try to compare any individual's data with those of a representative peer group, we would need a huge population encompassing all peer groups, ages, sexes, life-styles, drug use, ethnic background, and so on. In practice, compromises and assumptions are made, and diverse peer groups are often lumped together, an inappropriate practice. For example, many laboratories report single reference ranges for serum alkaline phosphatase and creatine kinase, which is clearly incorrect (3).

The detailed data on reference populations described by Siest et al. (4) are probably the best currently available for the common laboratory tests. For example, for serum cholesterol, Siest et al. present data on 31 875 people, including 7744 children ages 4–14 years.

A question as old as statistical theory itself is, How large a sample size is needed to give a reasonable estimate of the population variables? Minimum sample sizes of 30 (5) and 120 (6, 7) have been recommended, largely on the basis of experience, intuition, or a priori assumptions. Gaussian statistics are inappropriate for deriving reference ranges (8, 9), and hence definitions of sample size based on a gaussian assumption are incorrect or at least questionable.

Our goal was to use actual laboratory data obtained from young adults in good health to estimate what constitutes an adequate sample of the population. With a Monte Carlo simulation technique and 5000 samplings of different sizes from our population, we wanted to determine the stability of the estimates of the 2.5 and 97.5 percentiles.

Material and Methods

Subjects. Blood specimens were collected in the afternoon by venipuncture from the antecubital vein from apparently healthy subjects who had been seated ~30 min and who had eaten lunch ~2 h earlier. The group comprised second-year medical students, ages 22–31 years, attending The Ohio State University College of Medicine during February or March of 1976–1979, 1981, and 1982. The specimens were assayed within 2 h of collection for the various chemistry and hematology tests listed in Table 1.

Assay methods. We used Technicon's methods and their SMA 6/60 continuous-flow analyzer (Technicon Instrument Corp., Ardsley, NY 10591) for Na, K, and Cl; for glucose, we used a modified glucose oxidase procedure (10). All hematology tests were performed with a Coulter Counter Model S+2 (Coulter Electronics, Inc., Hialeah, FL 33012). The quality-control data are given in Table 1.

Data analysis. Data from the same test for the different years were combined; however, data from men and women were not combined. There were no changes in methods during the study. For the men, 5000 random samples of 10, 20, 50, 100, 153, 200, and 400 values were selected by using a random-number function and our Cray supercomputer. To estimate the 95% confidence level for these percentiles, 198 subjects are needed; 198 subjects are required for the 99% confidence level. The values of the 2.5 and 97.5 percentiles for each sample as well as the mean, upper, and lower values of these percentiles were determined at this time.

We reviewed College of American Pathologists (CAP) Comprehensive Chemistry Survey data for the above tests and years to determine our agreement with the means of our peer groups.

<table>
<thead>
<tr>
<th>Table 1. Methods and Quality-Control Data*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Na, mmol/L</td>
</tr>
<tr>
<td>K, mmol/L</td>
</tr>
<tr>
<td>Cl, mmol/L</td>
</tr>
<tr>
<td>Glucose, mmol/L</td>
</tr>
<tr>
<td>Hemoglobin, g/L</td>
</tr>
<tr>
<td>Erythrocyte, 10¹²/L</td>
</tr>
<tr>
<td>Hematocrit</td>
</tr>
</tbody>
</table>

* n = 30 for all; between-day data. * Mean (CV, %).
Results and Discussion

The number in each group and the 2.5 and 97.5 percentiles for the entire population of men and women are shown in Table 2. The data for the 5000 random examinations of various sample sizes are shown in Figure 1. We decided what constituted an adequate sample primarily by the appearance of the figures, rather than mathematically. We found that increasing the size of the sample had a stabilizing effect on both the 2.5 and 97.5 percentiles.

Percentiles were calculated by a simple linear interpolation technique. The sampled data of size n was put into rank order, and the tth percentile was calculated as the weighted average of the two observations surrounding \( x(p(n+1)) \), with \( p = t/100 \). More precisely, the tth percentile is \( (1 - gx_j + g x_{j+1}) \), where \( j \) is the integer part and \( g \) is the fractional part of \( (n + 1)p \).

From the interlaboratory CAP Comprehensive Chemistry Survey data, we concluded that our analytical methods at the times of sampling the populations were stable and that merging of the data from different years was appropriate.

For Na, a reference range based on <100 people is
inadequate (Figure 1); the nonparametric range does not stabilize until ~150 subjects are included. For glucose, a sample of 200 subjects appears to be reasonable (Figure 1).

Our observed range of glucose values requires comment. The students were asymptomatic, and ~2 h had passed since their last meal. Of the 957 men, we found 12 with a serum glucose concentration <2.8 mmol/L, 61 with a glucose concentration <3.3 mmol/L, and 21 with a glucose concentration >7.5 mmol/L. We cannot rule out that some students had a quick snack between their last meal and the test or that some were glucose intolerant. For the 201 women, none had a concentration <2.8 mmol/L, 9 had a concentration <3.3 mmol/L, and 6 had a concentration >7.5 mmol/L.

For hemoglobin data, ~150 subjects are required for the nonparametric reference range to be established (Figure 1). One asymptomatic person had a hemoglobin concentration <90 g/L. For the erythrocyte count, ~200 people are needed for the nonparametric reference range to be established (Figure 1).

For the tolerance limit approach used by Conover (11), at a 99% confidence level to capture the middle 95% of the data (i.e., the reference range), a sample of ~200 subjects is needed. A sample of 100 subjects gives only a 75% confidence level of estimating the same ranges. This approach makes no assumptions about the distribution of the data, gaussian or otherwise; it is strictly nonparametric.

If all reference range data followed a gaussian distribution, good estimates of the 2.5 and 97.5 percentiles of the population would be much simpler, and samples of ~120 people would suffice (6). Because we do not know the distribution of the population, a decision about adequate sample number is necessarily arbitrary and probably determined best by simulation or by a nonparametric method as we have done. We think our approach is reasonable and conclude that generally ~200 persons should be included for defining a reference group. Our sample of men had a narrow range of ages, and our reference interval for this group can be applied to other ages only cautiously. Our goal was not to define reference ranges for our population, but to determine what constitutes a reasonable sample of the population for estimating reference intervals.

In summary, the estimation of nonparametric reference ranges requires ~200 subjects, even if the group is fairly homogeneous, as was ours. Nevertheless, the intended use of the reference values should be kept in mind; for some applications, fewer subjects may suffice (12).

A judgment of adequate numbers can be made from a graph of nonparametric limits of increasing numbers of subjects drawn repeatedly from the same population. Nonparametric ranges are simple to calculate and do not make any assumptions about the distribution of the underlying population. Statistically, 200 subjects are needed at the 99% confidence level, 100 subjects at the 75% confidence level. The agreement of the statistically required number and that found by our simulation technique is more than a coincidence and strengthens our argument for the minimum sample size for estimating reference ranges.

References