Reference Range Determination: The Problem of Small Sample Sizes

W.D. Leslie and I.D. Greenberg

Section of Nuclear Medicine, Saint Boniface General Hospital (WDL) and Section of Nuclear Medicine, Health Sciences Centre (IDG), Winnipeg, Canada

The process of developing and validating a quantitative test includes determination of a reference range. Traditionally this has been taken as the mean ±2 standard deviations for a random sampling from a reference population. However, this method fails to recognize the substantial variability in the sample mean and standard deviation for the small sample sizes frequently encountered in nuclear medicine. A new approach, which involves calculating confidence intervals for the upper and lower bounds of the traditionally defined range, recognizes three ranges of values: normal, indeterminate, and abnormal. The principles of this approach are illustrated using differential renal function in twelve renal transplant donors. The ^{99m}Tc-DTPA differential uptake between 1 and 2 min gave a traditionally-defined single-kidney range of 50% ± 8%, whereas with our method the normal range would be 50% ± 6% with indeterminate ranges of 37%–44% and 56%–63%. These values are consistent with the wide variation in reference ranges reported in the literature, and suggest that much of this variability may be a statistical artifact resulting from inadequate sample sizes. A nomogram has been derived that permits the power of the reference range determination to be easily calculated from the sample size. Analysis of the effect of sample size on the accuracy of the upper and lower bounds of the reference range is advocated whenever small reference populations are used.

J Nucl Med 1991; 32:2306–2310

Determination of the reference range is one of the most fundamental steps in validating a quantitative test for clinical use. Traditionally this has been accomplished by prospectively performing the new test on a reference group that is considered representative of some larger population. When the data appear to fit a normal (Gaussian) distribution then the reference range, intended to include 95% of the values present in this larger population, is commonly taken as

$$\bar{X} \pm 2\ s, \qquad \text{Eq. 1}$$

where $\bar{X}$ and s are the sample mean and standard deviation, respectively. (The more statistically sophisticated reader will recognize that $\mu \pm 2\sigma$ actually gives a 95.4% coverage and that the "correct" formula is $\mu \pm 1.96\sigma$. We have elected to use the former since it is familiar to more readers; the difference is small and, for practical purposes, unimportant.) On the surface, Equation 1 has the appearance of statistical validity. So much so, in fact, that medical journals rarely (if ever) question this way of defining and presenting a reference range. Despite widespread use, few clinicians appreciate the assumptions, limitations, and potential dangers of this approach. A major difficulty arises from the confusing statistical practice of applying the terms "mean" and "variance" to both population parameters and sample statistics (Table 1). The distribution function for measurements arising from a given population are characterized by parameters which are, in general, unknown and unknowable (with absolute precision). A randomly sampled subset of the population provides the data needed to compute statistics such as the sample mean and sample variance that are used as estimators of the distribution mean and distribution variance, respectively. It is the precision with which the sample statistics estimate their respective distribution parameters that determines the validity of a traditionally-defined reference range. This precision, however, is highly dependent upon the sample size so that sample sizes of less than 50 can lead to serious inaccuracies (*1*). On the other hand, the suggestion that each moderate- to large-sized hospital should evaluate 100–120 reference individuals for each sex and age group (*2,3*) is not realistic for the constrained resources of the average nuclear medicine department (not to mention the ethical question of irradiating such a large population).

TABLE 1
Relationship Between Distribution Parameters, Sample Statistics and Reference Ranges Defined as Mean ± 2 Standard Deviations

	Mean	Standard deviation	Proportion of values within the reference range
Distribution	μ	σ	95.4%
Sample	$\bar{X}$	s	Variable

Received Jan. 23, 1991; revision accepted July 18, 1991.
For reprints contact: W. D. Leslie, Section of Nuclear Medicine (C3-006), St. Boniface General Hospital, 409 Tache Ave., Winnipeg, Canada, R2H 2A6.

We feel that explicitly determining confidence intervals for the reference range boundaries represents a compromise between the small sample sizes inherent to nuclear medicine and the desire for statistical rigor. This approach is illustrated using data derived from differential renal function measurements in a small series of normals.

MATERIALS AND METHODS

Assume that $\bar{X}$ and s are the mean and standard deviation derived from a random sample drawn from a population that is known to be normally distributed. The calculation of 95% confidence intervals for $\bar{X}$ is a technique covered in any introductory course in statistics. According to the central limit theorem the bounds are given by

$$\bar{X} \pm \frac{2s}{\sqrt{n}}, \qquad \text{Eq. 2}$$

where n is the sample size. Confidence intervals for s are also easily computed, although this is rarely discussed in the medical literature (*4*). Such an oversight is regrettable since the error associated with s is always larger than the error associated with $\bar{X}$. The 95% confidence interval for s is given by

$$\left(\sqrt{\frac{n-1}{a}}\cdot s, \quad \sqrt{\frac{n-1}{b}}\cdot s\right), \qquad \text{Eq. 3}$$

where a and b are the 0.975 and 0.025 critical values for a Chi-square distribution with n − 1 degrees of freedom (available from any set of statistical tables). When the true distribution mean (μ) is known there is a statistical gain in calculating the confidence interval for s which becomes

$$\left(\sqrt{\frac{\sum(X_i-\mu)^2}{a}}, \quad \sqrt{\frac{\sum(X_i-\mu)^2}{b}}\right), \qquad \text{Eq. 4}$$

where a and b are the 0.975 and 0.025 critical values for a Chi-square distribution with n degrees of freedom (note that a degree of freedom has not been lost in estimating μ). The critical points 0.025 and 0.975 can be altered to reflect the tolerance of the error in s that is considered acceptable.

The upper and lower reference range limits, $\bar{X}$ − 2s and $\bar{X}$ + 2s, are themselves statistics for which 95% confidence intervals can be determined. These intervals can then be used to define a new type of reference range. First, consider the case where μ is known and let the lower and upper bounds for s described above be denoted s_{min} and s_{max}, respectively. It is clear that any value greater than $\mu + 2\cdot s_{max}$ or less than $\mu - 2\cdot s_{max}$ is highly likely to be abnormal, whereas values between $\mu - 2\cdot s_{min}$ and $\mu + 2\cdot s_{min}$ are highly likely to be normal. This leaves two ranges of values for which the error in s does not provide sufficient power to characterize them as one or the other, and are best regarded as indeterminate. Thus, three classes of values are distinguished rather than the usual "normal-abnormal" dichotomy. With an infinitely large sample the indeterminate range collapses leaving the traditionally defined normal and abnormal classes (Fig. 1).

A similar approach can be used when μ is not known, except that the simultaneous variation in $\bar{X}$ needs to be considered in addition to the error in s. Our method represents an extension of the tolerance interval method as it is applied to reference range quantitation. We make use of one-sided tolerance factors which

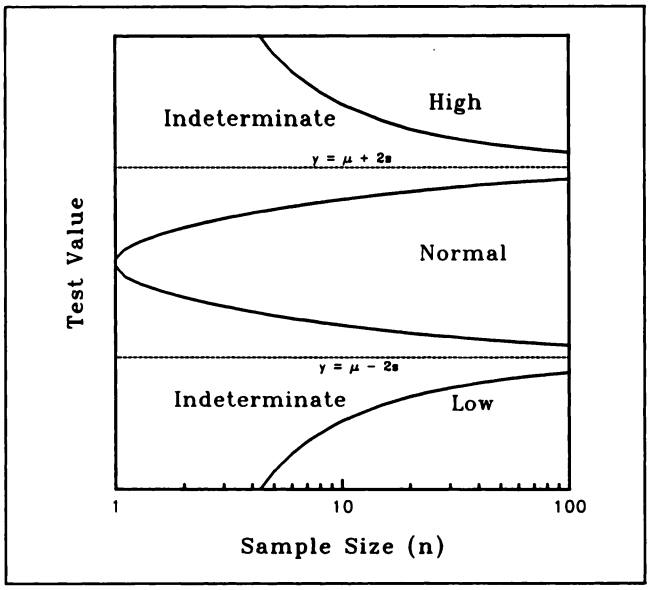

FIGURE 1. Effects of sample size on the nondichotomous reference range.

are values reflecting our confidence that a given percentile point of the population lies below a given cutoff. After selecting a level of probability for containing the true boundary point (typically 80%, 90%, 95%, or 99%), we can find factors k_1 and k_2 in Table 2 such that

$$\bar{X} \pm k_1\cdot s \quad \text{and} \quad \bar{X} \pm k_2\cdot s \qquad \text{Eq. 5}$$

give normal-indeterminate and indeterminate-abnormal boundary points, respectively, for a reference range that will include 95% of the reference population(s).

The width of the indeterminate classes is inversely related to the sample size. A simple index of the width of the indeterminate intervals is given by the ratio of intermediate-to-normal class size:

$$R = (s_{max} - s_{min})/S_{min} \quad \text{if } \mu \text{ is known} \qquad \text{Eq. 6}$$

$$R = (k_2 - k_1)/k_1 \quad \text{if } \mu \text{ is not known.} \qquad \text{Eq. 7}$$

As the sample size increases R decreases and approaches zero in the limit. Figure 2 illustrates this sample size dependence and can be used as a nomogram for rapid calculation of the sample size required to provide a given level of precision in the reference range. Predictably there is a large reduction in R for each increment in the sample size when the initial sample size is less than 30. However, even with 100 subjects the indeterminate interval is still more than one-third the size of the normal interval if a 95% confidence is desired.

Study Population

The study population consisted of 13 consecutive potential related renal transplant donors who had been referred to the nuclear medicine department for assessment of renal function. Of these, twelve had archived studies that were complete and analyzable. All subjects were free of a history of previous renal disease and hypertension and had normal renal function as assessed by creatinine clearance, urinary sediment analysis, and urinary protein excretion. Scintigraphic examinations were performed with technetium-99m-DTPA 740 MBq followed by ^{131}I-orthoiodohippurate (OIH) 7 MBq. An Elscint LFOV camera-

TABLE 2
One-sided Tolerance Factors (k_1, k_2) for $\mu \pm 2\sigma$ at Four Different Levels of Uncertainty (p = 80, 90, 95, 99%)*

n	p = 80%	p = 90%	p = 95%	p = 99%
2	(0.96, 15.59)	(0.71, 31.26)	(0.52, 62.56)	(0.15, 312.86)
3	(1.09, 6.24)	(0.88, 8.99)	(0.71, 12.82)	(0.43, 28.85)
4	(1.17, 4.64)	(0.98, 6.02)	(0.84, 7.71)	(0.58, 13.43)
5	(1.24, 3.98)	(1.06, 4.91)	(0.92, 5.98)	(0.68, 9.20)
6	(1.29, 3.62)	(1.12, 4.33)	(0.99, 5.11)	(0.76, 7.32)
7	(1.33, 3.39)	(1.17, 3.97)	(1.05, 4.60)	(0.83, 6.27)
8	(1.36, 3.23)	(1.21, 3.72)	(1.09, 4.24)	(0.88, 5.60)
9	(1.39, 3.11)	(1.24, 3.54)	(1.13, 3.99)	(0.93, 5.13)
10	(1.41, 3.01)	(1.27, 3.40)	(1.16, 3.80)	(0.97, 4.79)
11	(1.43, 2.94)	(1.30, 3.29)	(1.19, 3.65)	(1.00, 4.53)
12	(1.45, 2.87)	(1.32, 3.20)	(1.22, 3.53)	(1.04, 4.32)
13	(1.47, 2.82)	(1.34, 3.13)	(1.24, 3.43)	(1.06, 4.15)
14	(1.48, 2.77)	(1.36, 3.06)	(1.26, 3.34)	(1.09, 4.00)
15	(1.50, 2.74)	(1.38, 3.01)	(1.28, 3.27)	(1.11, 3.88)
16	(1.51, 2.70)	(1.39, 2.96)	(1.30, 3.20)	(1.13, 3.78)
17	(1.52, 2.67)	(1.41, 2.91)	(1.32, 3.15)	(1.15, 3.69)
18	(1.53, 2.64)	(1.42, 2.88)	(1.33, 3.10)	(1.17, 3.61)
19	(1.54, 2.62)	(1.43, 2.84)	(1.34, 3.05)	(1.19, 3.54)
20	(1.55, 2.60)	(1.44, 2.81)	(1.36, 3.01)	(1.20, 3.48)
21	(1.56, 2.58)	(1.45, 2.78)	(1.37, 2.98)	(1.22, 3.42)
22	(1.57, 2.56)	(1.46, 2.76)	(1.38, 2.95)	(1.23, 3.37)
23	(1.57, 2.54)	(1.47, 2.73)	(1.39, 2.91)	(1.24, 3.32)
24	(1.58, 2.52)	(1.48, 2.71)	(1.40, 2.89)	(1.25, 3.28)
25	(1.59, 2.51)	(1.49, 2.69)	(1.41, 2.86)	(1.27, 3.24)
30	(1.61, 2.45)	(1.52, 2.61)	(1.45, 2.76)	(1.32, 3.08)
35	(1.64, 2.41)	(1.55, 2.55)	(1.48, 2.68)	(1.36, 2.97)
40	(1.66, 2.37)	(1.58, 2.50)	(1.51, 2.62)	(1.39, 2.88)
45	(1.67, 2.34)	(1.60, 2.46)	(1.53, 2.57)	(1.42, 2.81)
50	(1.68, 2.32)	(1.61, 2.43)	(1.55, 2.54)	(1.44, 2.75)
60	(1.71, 2.28)	(1.64, 2.38)	(1.58, 2.48)	(1.48, 2.67)
70	(1.72, 2.26)	(1.66, 2.35)	(1.61, 2.43)	(1.51, 2.60)
80	(1.74, 2.24)	(1.68, 2.32)	(1.63, 2.39)	(1.53, 2.55)
90	(1.75, 2.22)	(1.69, 2.30)	(1.64, 2.37)	(1.56, 2.51)
100	(1.76, 2.20)	(1.71, 2.28)	(1.66, 2.34)	(1.57, 2.48)

* This table is abridged and adapated from Tables 1.1-1.13 of Ref. (5) courtesy of Marcel Dekker Inc.

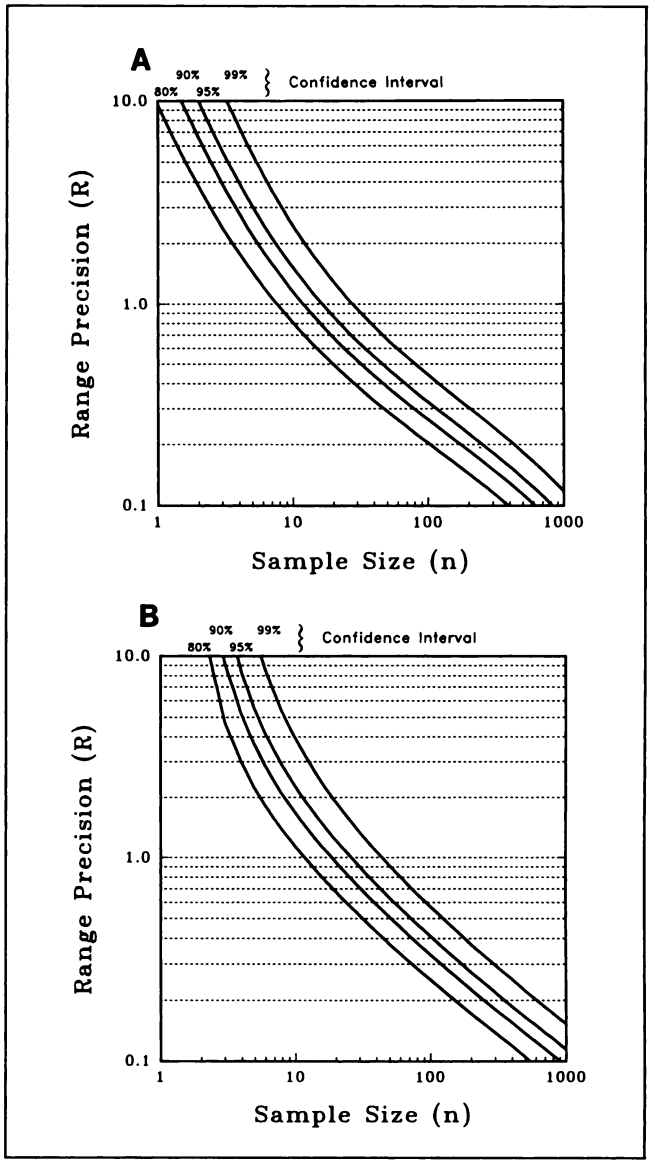

FIGURE 2. Relationship between sample size and reference range precision (R) for confidence intervals of 80%, 90%, 95%, and 99%. (A) If the true mean (μ) is known. (B) If the true mean (μ) is not known.

computer system was used to acquire data over a 30-min period after each injection. The kidney outlines and a single background region of interest were identified by a single operator. Differential renal functions based upon uptakes between 1 and 2 min were determined for the DTPA and OIH studies. The analysis was repeated by a second person without knowledge of the first analysis. There was extremely close correlation in the results (s.e.e. 1.3% for DTPA and 1.6% for OIH). Values for the left kidney constitute the raw data for the remaining discussion.

RESULTS

The values for $\bar{X}$ and s derived from our study population are presented in Table 3. The results with OIH were very similar to those obtained with DTPA. The statistic s_{50} is the standard deviation based upon the assumption that μ is known to be 50%. Reference ranges based upon s_{50} will be much more convenient to use since they will be symmetric about 50%. There is nothing in the definition of the standard deviation as a measure of dispersion about a central value which is violated as long as we have good theoretical reasons for choosing μ. This choice is supported by anatomical and radiological data which indicates the left kidney to be only marginally larger than the right (*6*).

The reference range for DTPA is compared with previously published reference ranges (*7,8,9*) in Figure 3. Analysis of the pooled data from the above studies gives a standard deviation of 3.6% (95% confidence interval 3.2–4.2). This translates into a traditionally-defined symmetric normal range of 50% ± 7.3%. The nondichotomous tolerance interval method gives a normal range of 50% ± 6.4% flanked by comparatively small indeterminate ranges of 41.6%–43.6% and 56.4%–58.4%. In this case the statistical precision afforded by the pooled sample size (n = 124) probably does not warrant the added complexity of a nondichotomous reference range.

TABLE 3
Differential Left Kidney Functions (%) and Derived Nondichotomous Reference Ranges

	$DTPA_{1-2}$	OIH_{1-2}
Raw Statistics		
Mean ($\bar{X}$)	51.3	51.0
	(48.8-53.7)	(48.7-53.3)
St Dev (s)	4.0	3.9
	(2.8-6.8)	(2.7-6.6)
St Dev (s_{50})	4.2	4.0
	(2.9-6.6)	(2.8-6.4)
Symmetric reference range		
Normal	44.2-55.8	44.4-55.6
Indeterminate	36.8-44.2	37.1-44.4
	55.8-63.2	55.6-62.9
Abnormal	< 36.8	< 37.1
	> 63.2	> 62.9
Asymmetric reference range		
Normal	46.4-56.2	45.2-55.8
Indeterminate	37.2-46.4	37.2-46.2
	56.2-65.5	55.8-64.8
Abnormal	< 37.2	< 37.2
	> 65.5	> 64.8

DISCUSSION

Many methods have been described for extracting a reference range from a set of data. However, it is useful to remember that these are only estimates of the true reference range which is, in general, unknowable. We have attempted to illustrate this principle in the context of differential renal function using DTPA and OIH, the two most frequently used renal radiopharmaceuticals. Considerable variability exists in the published literature with respect to the normal degree of asymmetry in this parameter. Most of this variability can be explained by the uncertainty in the standard deviation arising from the use of relatively small sample sizes. In our study the normal range for single kidney function would typically be stated as 50% ± 8%. Our nondichotomous approach leads to a normal range of 50% ± 6%, with indeterminate ranges of 37%–44% and 56%–63%, and abnormal ranges of less than 37% or greater than 63%.

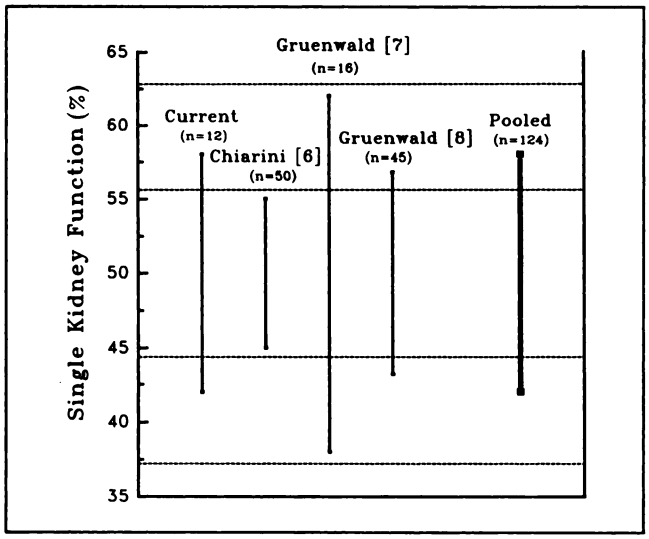

FIGURE 3. Published reference intervals for differential renal function compared with the current study. Dotted lines represent the normal-indeterminate and indeterminate-normal divisions for the DTPA data from the current study.

In general, nonparametric methods are preferred since they make no assumptions about the shape of the underlying distribution. Two such approaches are the 2.5–97.5 interpercentile intervals (*10*) and nonparametric tolerance limits (*11,12*). Although the former can theoretically be used with as few as 40 cases, in practice a considerably larger sample is necessary to give reasonable statistical accuracy. The latter is also unusable with small samples and would only give an 18% probability of containing the true 95% reference interval if only 12 cases were available as in the current study (*13*).

The major advantage of a parametric method is its ability to generate distribution predictions based upon a smaller number of patients. The choice of the distribution then becomes critical since a wrong choice can produce misleading results. Ironically, determining the shape of a distribution typically requires a much larger sample than estimating the 2.5–97.5 percentile points (*14,15*). Parametric tolerance limits have been a popular way to quantitate the effect of sample size on the reference interval (*12*). Finding a 95% coverage requires the use of a wider reference range as the sample size decreases, while the abnormal range simultaneously becomes smaller. A value falling outside of the bounds $\bar{X} \pm 2s$ might then be accepted as normal (ie., the statistical power of the reference sample is insufficient to "prove" that this value falls outside the range $\mu \pm 2\sigma$). Clearly this leads to a reduction in the ability to correctly classify abnormal results. Intuitively both the normal and abnormal ranges should shrink as the error in our estimates for μ and σ increases.

Criticisms that our method is overly conservative in its definition of the upper and lower bounds are justified insofar as values laying outside of a 95% confidence interval based upon the upper end of a 95% confidence interval for the standard deviation are almost certain to be truly abnormal. However, flexibility in the choice of P for the tolerance factors (or, if μ is known, in the critical points of the Chi-square distribution) provides a mechanism for specifying less severe conditions for "normalcy."

Clearly, the preoccupation with finding the "correct" cut-off point between normal and abnormal ignores the arbitrary nature in any such choice. We have formalized what should be intuitively obvious: some values fall within a statistical grey zone and cannot be categorized as normal or abnormal, but this borderline range can be decreased with larger sample sizes. Since small reference populations will continue to be a fact of life in most nuclear medicine facilities, we advocate an explicit analysis of sample size effects as a routine step in reference range determination.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Dr. T. Hassard, Dr. C. Schwarz and Dr. A. Arnason for their valuable advice during the preparation of this work.

REFERENCES

1. International Federation of Clinical Chemistry: the theory of reference values. Part 6. Presentation of observed values related to reference values. *Clinica Chimica Acta* 1983;127:441f–448f.
2. Reed AH, Henry RJ, Mason WB. Influence of statistical method used on the resulting estimate of normal range. *Clin Chem* 1971;17(4):275–284.
3. Cerembrowski GS, Sullivan AM. Quality control and statistics. In: Biello LA, ed. *Clinical chemistry.* Philadelphia: J.B. Lippincott Co.; 1985:57–86.
4. Hogg RV, Craig AT. *Introduction to mathematical statistics, fourth edition.* New York: Macmillan Publishing Co.; 1978:222–227.
5. Odeh RE, Owen DB. Tables for normal tolerance limits, sampling plans, and screening. In: *Statistics: textbooks and monographs, volume 32.* New York: Marcel Dekker Inc.; 1980.
6. Moell H. Size of normal kidneys. *Acta Radiol* 1956;32:8–13.
7. Chiarini C, Esposti ED, Lossino F et al. Renal scintigraphy versus renal vein renin activity for identifying and treating renovascular hypertension. *Nephron* 1982;32:8–13.
8. Gruenewald SM, Collins LT. Renovascular hypertension: quantitative renography as a screening test. *Radiology* 1983;149:287–291.
9. Gruenwald SM, Collins LT, Antico VF et al. Can quantitative renography predict the outcome of treatment of atherosclerotic renal artery stenosis? *J Nucl Med* 1989;30:1946–1954.
10. Solberg HE. Establishment and use of reference values. In: Tietz NW, ed. *Fundamentals of clinical chemistry.* Philadelphia: W. B. Saunders Co.; 1987:197–212.
11. Wilks SS. Statistical prediction with special reference to the problem of tolerance limits. *Ann Math Statist* 1941;13:400.
12. Kendall M, Stuart A. *The advanced theory of statistics (volume 2), fourth edition.* London, England: Charles Griffin and Co.; 1978:139–142, 547–549.
13. Brunden MN, Clark JJ, Sutter ML. A general method for determining normal ranges applied to blood values for dogs. *Am J Clin Pathol* 1970;53:332–339.
14. Elveback LR. How high is high? A proposed alternative to the normal range. *Mayo Clin Proc* 1972;47:93–97.
15. Bezener PD, Netelenbos JC, Mulder C, et al. Determining reference ('normal') limits in medicine: an application. *Stats Med* 1983;2:191–198.