Background: We introduce a new criterion, the percentile inclusion probability, for comparing methods for calculating reference intervals. The criterion is compared with a previously published measure of reliability suggested by Linnet (Linnet K. Clin Chem 1987;33: 381–6), the ratio of the width of the confidence interval for the percentile to that of the reference interval.

Methods: Data are simulated from a range of theoretical statistical distributions representing the shapes of data sets encountered in clinical investigations. The two-stage transformation of the data to a gaussian distribution recommended by the IFCC is compared with a nonparametric approach.

Results: The percentile inclusion probability criterion identifies that the parametric approach is in some cases seriously affected by bias. Using different parametric models, we compare nonparametric and parametric methods for two sets of clinical data and show that the parametric approach is susceptible to model choice.

Conclusions: Sample sizes significantly greater than those currently recommended are required to establish reference intervals, regardless of whether parametric or nonparametric methods are used. Parametric methods are preferable when the data are truly gaussian, but are only marginally better than nonparametric methods when data transformation is needed to achieve a gaussian shape.

Reference intervals are commonly used in clinical studies as an initial signal to determine whether further patient testing is indicated. In this report we deal with the estimation of reference intervals for measurements where there is no change in the distribution of the measurement with age, i.e., time-independent reference intervals. In our companion report (1), we discuss methods for the calculation of time-specific reference intervals.

A recommended procedure for establishing reference intervals has been determined by the IFCC and is described in Solberg (2, 3). Briefly the method entails the estimation of percentiles of the reference population so that the required percentage of the population is included in the range defined by these percentiles. Fig. 1A illustrates the distribution of birthweights for full-term male births (gestational age, 40 weeks) in Wales (UK) between 1988 and 1997, obtained from the Child Health Register for Wales. Both unusually underweight or overweight infants may be referred for further monitoring; this therefore is an example of a case where a two-sided reference interval is of interest. Fig. 1B shows the distribution of neonatal thyroid-stimulating hormone (TSH) measured in newborns in Cardiff, Wales in 2003 (up to October 13, 2003), obtained from the Biochemistry Department at the University Hospital of Wales, Cardiff. These data were obtained as part of the screening program in Wales for inborn errors of metabolism. In this case an infant is referred for further tests if this measure is increased; therefore, a one-sided reference interval is needed. The birthweight and TSH data sets are very large (63 589 and 17 705, respectively). In a sense these are complete populations, the small number of infants not measured will not produce any noticeable bias. With smaller samples, of course, issues of sample bias and generalizability undoubtedly arise.

Materials and Methods

There are two types of statistical approaches that are used for the estimation of reference intervals. Nonparametric methods make no assumptions about the functional form of the reference distribution. Methods based on the order statistics are the most commonly used, as described, for example, in Harris and Boyd (4) and Bland (5). A review of modifications to this simple method can be found in Wright and Royston (6). The formulas used here are given in the Technical Appendix (available as Data Supplement...
Parametric approaches, on the other hand, assume that the reference distribution can be modeled by a standard statistical distribution. Usually the gaussian distribution is assumed; methods based on this assumption are described by Bland (5) and Altman (7), who also describe the calculation of related confidence intervals. Distributions other than the gaussian can be used for parametric estimation of reference intervals; an analysis of the TSH data given in the Results assumes that the data are sampled from an exponential distribution. For the parametric approach the percentiles of the standard distribution are known exactly, but uncertainty is introduced because of the need to estimate the parameters defining the reference population.

Where the data do not conform to a gaussian distribution, some authors, rather than seeking an alternative statistical model for the data, have suggested that an initial transformation of the data should be found such that the transformed data resemble a sample from a gaussian parent. Altman (7) discusses the reasons for transforming data to a gaussian distribution. The IFCC have recommended a two-stage transformation (2, 3). The detailed formulas are given in the Technical Appendix (http://www.clinchem.org/content/vol50/issue5/).

Briefly, the method uses two consecutive transformations of the data: the first is intended to transform the data to a symmetric distribution, and the second is to remove kurtosis. It is this transformation that is investigated in this report. The sample variation in the estimates of the parameters defining the transformation leads to a rather larger sample size being needed for estimation purposes than would be needed if no transformation were required. In general, the effect that this sample variation has on the reference intervals is to make the percentile estimates less precise and hence render the reference intervals more imprecise than is the case when the untransformed data have a known distribution.

Sample size determination is an important integral part of study design and must be determined in advance because it may not be possible to extend the study if analyses indicate that this may be necessary. Such an extension of the sample size might be particularly difficult in the case of a clinical trial. Linnet (8) gave recommendations on sample sizes based on a ratio criterion, the width of the 90% confidence interval for estimating the 97.5th percentile divided by the width of the reference interval itself. In the gaussian case his suggestion is that between 50 and 450 individuals are needed. He added, however, that the ratio defined above can be expected to be ~25% larger when parameters are estimated to determine a transformation to gaussian form. The sample size needed for a confidence interval of prescribed width is proportional to the square of the width, and Harris and Boyd (5) used this fact, together with Linnet’s recommendation (8), to conclude that sample sizes would need to be 1.25^2, or just over 50% greater when transformation parameters are estimated. Their general guidance is that between 150 and 200 individuals should first be measured to establish the parameters of the transformation. Use is then made of charts to determine how many more individuals should be measured (if any) to estimate reliably the reference interval. The necessary sample size depends on the transformation used, and in extreme cases a sample size of 1500 is advocated. In the nonparametric case sample sizes between 125 and 700 are recommended by Linnet (8).

Whether a parametric or a nonparametric approach is used, the effect of bias in the estimation of the percentiles should be considered. Suppose it is the case that an estimator of a percentile is biased; perhaps its mean value is equal to the 95th percentile instead of the intended 97.5th. In this case, instead of referring 2.5% of the population to the upper end of the range, on average 5% will be referred, a 100% increase in the referral rate.

Assessment of the efficacy of the estimated reference interval, either by an examination of the confidence interval derived for the estimate or by Linnet’s ratio (8), accompanying the online version of this article at http://www.clinchem.org/content/vol50/issue5/).

(A), histogram of birthweights for full-term (gestational age, 40 weeks) male births in Wales (UK) between 1988 and 1997. (B), histogram of TSH measurements (mL/L) for newborns measured at the University Hospital of Wales (Cardiff, UK) in 2003 (up to October 13, 2003).

Fig. 1. Distribution of birthweights (A) and TSH concentrations (B) in newborns.
would not take account of the effect of this bias. It is also possible for a biased estimator to have a value for Linnet’s ratio that is in the range 0.1–0.3, which Linnet used to define acceptable sample sizes, but for the percentage referred to be very different from that required. The aim of a reference interval is to include a specified percentage of the population. In assessing methods for setting reference intervals, therefore, it is of value to measure the extent to which this percentage is achieved. Thus, for example, the proportion of occasions on which the estimated 97.5th percentile lies, e.g., between the true 95th and 99th percentiles, is a direct measure that can be used to assess the efficacy of the calculated reference interval in excluding the specified percentage of the population in the upper tail. We call this proportion the percentile inclusion probability (PIP). If the PIP for the estimated 97.5th percentile is 0.9, then the percentage of the population excluded from the reference interval at the upper end of the range will be between 1% and 5% on 90% of occasions. As is the case with confidence intervals, the proportion targeted to occur in the prescribed true percentile range of 0.95 is arbitrary, but ideally is a high value. The limits of the true percentile range 0.95–0.99 are also arbitrary but ideally would be close to the percentile that is to be estimated. This measure is sensitive to both the bias and imprecision associated with the percentile estimator, whereas Linnet’s ratio assesses only the imprecision (8). Clearly it is not possible to calculate the PIP without reference to the true percentiles of the population; therefore, the criterion cannot be used for the assessment of a single sample. It is, however, a criterion that can be readily used in simulation studies designed to compare the nonparametric and parametric approaches.

In the Results section we show how the PIP criterion of assessment compares with Linnet’s ratio (8) by means of simulation studies based on random samples taken from the $\chi^2$ family of distributions. For a small number of degrees of freedom, this distribution is highly positively skewed. As the number of degrees of freedom increases, the distribution becomes less skewed, and for $\geq 20$ degrees of freedom the main part of the $\chi^2$ distribution resembles the gaussian, although the right-hand tail is somewhat thicker than the tail of the gaussian distribution. This family of distributions has previously been used [e.g., Boyd and Lacher (9)] to model the range of unimodal positively skewed data that are generally encountered in clinical studies; we therefore consider it appropriate for simulation purposes here. We also generated data from a gaussian parent distribution to make a further comparison of the parametric and the nonparametric methods for which no transformation of the data is needed. Additional details

### Results

The results of simulations to compare the parametric and nonparametric estimates of the 97.5th percentiles for a range of distributions are given in Table 1. In other simulation studies, not reported here for brevity, we have observed similar results for the 2.5th percentile. The first four sampling distributions in Table 1 show the results of simulations from the $\chi^2$ family of distributions with 4, 7, 10, and 20 degrees of freedom, respectively, and the sampling distribution gives results from simulations using a two-stage transformation where the data are drawn from a gaussian parent distribution. In all of these simulations the two-stage transformation was used to convert

| Table 1. Results of simulation studies comparing the nonparametric and parametric approaches for estimation of the 97.5th percentile.\(^a\) |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Sampling distribution | True 97.5th percentile | 50 | 100 | 250 | 500 |
| $\chi^2(4)$ | 11.143 | 10.57 (1.84) | 10.62 (1.21) | 10.62 (0.69) | 10.60 (0.48) |
| P | 12.37 (2.41) | 11.70 (1.55) | 11.34 (0.96) | 11.24 (0.67) |
| NP | 15.35 (1.88) | 15.52 (1.30) | 15.59 (0.81) | 15.61 (0.56) |
| $\chi^2(7)$ | 16.013 | 17.43 (2.86) | 16.62 (1.78) | 16.20 (1.10) | 16.12 (0.77) |
| P | 19.78 (2.07) | 20.01 (1.47) | 20.12 (0.91) | 20.16 (0.64) |
| NP | 21.92 (3.12) | 21.17 (2.00) | 20.70 (1.23) | 20.61 (0.86) |
| $\chi^2(10)$ | 20.483 | 33.24 (2.59) | 33.64 (1.87) | 33.84 (1.17) | 33.92 (0.82) |
| P | 36.00 (3.79) | 35.01 (2.50) | 34.44 (1.52) | 33.40 (1.08) |
| NP | 1.960 | 1.89 (0.30) | 1.92 (0.22) | 1.94 (0.14) | 1.95 (0.10) |
| Gaussian* | 1.960 | 2.15 (0.40) | 2.04 (0.27) | 1.99 (0.17) | 1.98 (0.12) |

\(^a\)In the first five sampling distributions entries are means of $>10\ 000$ simulations together with SD (in parentheses). The first member (P) of each pair is the estimate based on the parametric approach using the two-stage transformation. The second member (NP) is the estimate based on the nonparametric approach. The last row (Gaussian*) gives theoretical results for the gaussian distribution.
the data to a gaussian shape. The last row of Table 1, referred to as Gaussian*, gives the results for the estimator ($x + 1.96s$), without any transformation being used, using well-known approximate results quoted, for example by Bland (5).

The first sampling distributions in Table 1 are subdivided into two estimates. The parametric estimates (labeled P) were obtained with the two-stage transformation, and the nonparametric estimates (labeled NP) were obtained with the formula described in Altman (7). Each entry is the mean of >10,000 simulations for the estimated 97.5th percentile together with the SD (in parentheses). The last row of Table 1 refers to the estimator ($x + 1.96s$).

It can be seen that the parametric estimates are consistently biased downward and that the nonparametric estimator used in this report is biased upward. The bias is reduced as the sample size increases, but whereas for the nonparametric estimates the bias is small for a sample size of 500, it is still noticeable at this sample size for the parametric estimates. This bias is appreciably larger for the $\chi^2$ simulations in percentage terms as well as absolute terms. The only exception to this pattern is when sampling from a gaussian distribution, where for both parametric and nonparametric estimates the bias decreases as the sample size increases. The last row of Table 1 shows results for the gaussian distribution where no transformation is used.

We now turn to the criteria suggested in the Methods for assessing the efficacy of reference intervals, Linnet’s ratio (8), and the PIP. For samples taken from a gaussian distribution, a gaussian approximation to the distribution of the estimator ($x + 1.96s$) of the 97.5th percentile is described in Bland (5). We have used this in our calculations but have used the correction for the bias in $s$ given in Sokal and Rohlf (10). This allows us to calculate a reliable estimate of Linnet’s ratio (8) and the coverage probabilities. For the assessment of reference intervals derived from samples in $\chi^2$ distributions with 4, 7, 10, and 20 degrees of freedom, no theoretical results are available; we therefore made an assessment using the sampling distributions obtained from the simulation exercise described above. We examined the distribution of these estimates to give estimates of Linnet’s ratio and the values of PIP. Shown in Fig. 2 is a plot of Linnet’s ratio against sample size, and Fig. 3 shows a plot of PIP against sample size.

Linnet (8) tabulated sample sizes needed to achieve ratios of 0.1, 0.2, and 0.3. Taking the central value of 0.2 as a criterion, we can see that a sample size of 50 satisfies the criterion for a parametric estimator when data are from a gaussian distribution. If, however, the two-stage transformation is used, then Linnet’s ratio is generally somewhat larger; therefore, the sample size needed to achieve a ratio of 0.2 increases. A sample size of 85 is needed when samples are drawn from a gaussian parent distribution and a two-stage transformation is used. Notice, however, that these sample sizes take no account of any bias in the estimates of the percentiles. For data from the $\chi^2$ family of distributions, the sample size needed is between 150 (for 20 degrees of freedom) and 260 (for 4 degrees of freedom).

![Fig. 2. Linnet’s ratio (see text for definition) plotted against sample size.](image)

The bottom curve is the theoretical curve for the gaussian distribution. The other curves are based on simulations. The solid curves represent results obtained with the two-stage transformation for the parametric approach; the dashed curves represent the nonparametric approach. From top to bottom, parent distributions are $\chi^2$ with 4, 7, 10, and 20 degrees of freedom (parametric); $\chi^2$ with 4, 7, 10, and 20 degrees of freedom (nonparametric); gaussian (nonparametric), and gaussian (parametric). Simulations were done at sample sizes of 50, 100, 150, 200, 250, 300, 350, and 500. The curves shown were drawn by interpolation.

![Fig. 3. PIP for the estimated 97.5th percentile to lie between the true 95th and 99th percentiles.](image)

The top curve is the theoretical curve for the gaussian distribution. The other curves are based on simulations. The solid curves represent results obtained with the two-stage transformation for the parametric approach. From top to bottom, the parent distributions for the next three curves are gaussian ($\chi^2$ with 20 and 10 degrees of freedom; indistinguishable from this curve) and $\chi^2$ with 7 and 4 degrees of freedom (parametric). The bottom curve (dashed) represents results from the nonparametric approach from all parent distributions investigated. Simulations were done at sample sizes of 50, 100, 150, 200, 250, 300, 350, and 500. The curves shown were drawn by interpolation.
Where a nonparametric estimator is used for samples from a gaussian distribution, a sample size of 120 is needed. For the $\chi^2$ family, sample sizes in the range 27–500 are indicated. There is thus a penalty, in terms of this assessment criterion, for use of a nonparametric estimator.

Examination of Fig. 3 shows that for data from a gaussian distribution where no transformation is used, a sample size of 50 gives a PIP of 0.85. This is unacceptably low and is caused by both the bias and imprecision in the estimator. To achieve a PIP of 0.95 requires sample sizes of at least 100, this is double the sample size of 50 indicated by Linnet’s ratio of 0.2 (8). If the two-stage transformation is used for data from a gaussian parent distribution, the sample size needed for a PIP of 0.95 increases to ~170, again double the sample size of 85 needed for Linnet’s ratio to equal 0.2 (see above). Results for this two-stage transformation on $\chi^2$ data show that sample sizes needed to achieve a PIP of 0.95 vary between 180 (for 20 degrees of freedom) to 235 (for 4 degrees of freedom). Where a nonparametric estimator is used this criterion of a PIP of 0.95 is satisfied for sample sizes <250, this being independent of the parent distribution from which samples are taken.

The PIP criterion defined earlier is the probability that the estimated 97.5th percentile lies between the true 95th and 99th percentiles. This might be felt to be rather wide in some applications, and we have repeated calculations with a more stringent requirement for PIP, the probability that the estimated 97.5th percentile lies between the true 96.5th and 98.5th percentiles. For samples taken from a gaussian distribution and not transformed, a sample size of 50 gives a value just over 0.5 for this more stringent PIP. To achieve a PIP of 0.95, a sample size of ~400 is needed. We found that for the $\chi^2$ distribution with 4 degrees of freedom the two-stage transformation gave a value of only 0.8 for this more stringent PIP, and a sample size of 1000 is required if a PIP of 0.95 is required. This is undoubtedly attributable to the bias in the estimator of the 97.5th percentile for this skewed distribution. For samples from the $\chi^2$ distribution with 7 degrees of freedom where the two-stage transformation is used, these PIP values are higher, but still only 0.8 with a sample size of 500. For these two distributions in which the nonparametric approach is used, the PIP is ~0.85 for a sample size of 500. In all other cases, regardless of whether a parametric estimator derived from a two-stage transformation or a nonparametric estimator is used, the sample sizes needed for this more stringent PIP criterion to equal 0.95 are between 750 and 1000. A sample size of 500 gives a value of the PIP between 0.85 and 0.9.

Finally we analyze the two data sets described earlier. These two data sets illustrate different facets of the parametric approach to reference interval estimation. Both data sets are very large; indeed they can be viewed as complete populations because every infant in Wales is weighed at birth and also screened for TSH. Hence, the true percentiles in both cases are essentially known to be the appropriate order statistics, and it is of interest to compare these with estimates obtained by a parametric approach.

For the birthweight data the sample size is 63,589, and the 2.5th and 97.5th percentiles are 2715 and 4480 g, respectively, these two values defining the 95% reference interval. A visual examination of Fig. 1A indicates that a gaussian model is plausible for this distribution, and the parametric estimate of the 95% reference interval is 2675–4445 g, differing only slightly from the true interval. However, the Anderson–Darling test for a gaussian distribution is 34.97 ($P < 0.0005$), strongly indicating departure from a gaussian distribution. The data were therefore transformed by the two-stage transformation, and a revised estimate of the 95% reference is 2685–4495 g. These three methods of calculating the reference intervals give very similar results, and the differences, <70 g, are not likely to be of clinical significance. Thus, although the gaussian model is incorrect, the parametric estimate of the 95% reference interval assuming a gaussian distribution is close to the true 95% interval. The majority of these data are closely modeled by a gaussian distribution, but the tails are elongated. Only if more extreme percentiles were of interest would the (incorrect) assumption of a gaussian model give misleading answers for the reference interval.

Our second illustration is of the TSH data, where there are 17,705 observations. These data are characteristically different from the birthweight data in that 9335 readings <1 mL/L are recorded as “<1” because of the lack of sensitivity of the assay. Because it is the upper percentile that is of interest in this case, this left censoring of the data causes no difficulty in the nonparametric approach. The 95th percentile is 3.76; therefore, the 95% reference interval is all measurements <3.76 mL/L. In this case it is evident from Fig. 1B that a gaussian model is not appropriate for these data; however, it is not clear how a transformation for censored data of this sort can be obtained. Certainly because estimates of the skewness and kurtosis of the data are severely affected by the censoring, the two-stage procedure cannot be used. It is possible, however, to fit an exponential distribution to these data, using the method of maximum likelihood to estimate the parameter of the distribution. Some additional details and descriptions of the tools used (11–14) are given in the Technical Appendix (http://www.clinchem.org/content/vol50/issue5/). See also Wetherill (15). This parameter, the mean, is estimated as 1.429. The fitted exponential distribution with this mean is shown in Fig. 1B, and as can be seen, it may be a plausible model for the data. However, it is not our intention to claim that this is the correct model, merely that it may be an acceptable parametric model for the data. For these data, however, the 95th percentile estimated from the fitted model (4.28 mL/L) is somewhat higher than the true percentile (3.76 mL/L). In fact the percentage of TSH measurements exceeding 4.28 mL/L is 3.25%, much lower than the
required 5%. As we have argued, with such a large data set it is most unlikely that the 95th percentile is wrongly estimated by the nonparametric method, but despite the fact that the fitted exponential model is plausible, there is considerable bias in the estimation of this percentile obtained with this parametric approach. It may be possible to find an alternative parametric model that gives an unbiased estimate of the percentile; the point is that the choice of model can have a profound effect on the percentile estimates.

Discussion
Transformations have been widely used in statistical analyses to obtain data that can be assumed to derive from a gaussian parent distribution, and undoubtedly the procedure has been most useful in many types of statistical analysis. The use of such transformations in establishing reference intervals in clinical studies has been widely advocated, but we have shown in this report that some care needs to be exercised in assessing the reliability of the intervals thus obtained. In interpreting the implications of these ranges the practitioner should consider the risks involved. The precision of the estimator is of less importance in this field than the closeness of the percentage included in the reference interval to the nominal value. Our studies have shown that with sample sizes <200 it is sometimes the case that the estimator of the 97.5th percentile is actually so biased that its expected value lies below the true 95th percentile.

When no parameters need to be estimated, either because the data are drawn from a gaussian parent or the parent distribution is known exactly, a sample of 100 observations gives a value of PIP for the estimated 97.5th percentile between the true 95th and 99th percentiles of 0.95. If the value of PIP is defined to be the probability that the estimate of the 97.5th percentile is between the true 96.5th and 98.5th percentiles, then a sample size of 500 is needed to achieve a PIP of 0.95.

If a transformation is used on the data to make the shape of the distribution resemble that of a gaussian distribution, then larger sample sizes are needed. Our simulations indicate that sample sizes between 500 and 1000 may be necessary, depending on the definition of the PIP. With such large sample sizes there is little disadvantage, in terms of the PIP, in using nonparametric estimators. Our analysis of the birthweight and TSH data show that there is another consideration that leads to a preference for using nonparametric estimators. The birthweight distribution is not gaussian, having noticeably longer tails. In this case extreme values are rare, and the parametric estimates of the 2.5th and 97.5th percentiles are very close to their true values. In the case of the TSH data, however, a different conclusion is reached. The parametric model that we have used gives an estimate of the 95th percentile that is badly biased. For real populations it is extremely unlikely that a simple parametric model will accurately represent the observed frequencies for a very large sample. With smaller samples, however, the lack of sensitivity of statistical goodness-of-fit tests and the effects of sampling error could combine to indicate that an incorrect parametric model can be accepted. The nonparametric approach is less susceptible to such model deficiencies, and where the sample size exceeds 500 this approach is also less affected by extreme values or outliers in the data. This leads us to recommend that it be used as standard. Only if the parametric form of the distribution is truly known should sample sizes significantly <500 be considered, and then only when their efficiency has been assessed, possibly by a simulation exercise.

References