A new approach for the determination of reference intervals from hospital-based data

D. Concordet [a], A. Geffré [b], J.P. Braun [a,b], C. Trumel [b]

[a] UMR181 Physiopathologie & Toxicologie Expérimentales INRA, ENVT, Ecole Nationale Vétérinaire, 23 Chemin des Capelles, 31076 Toulouse Cedex 3, France
[b] Department of Clinical Sciences, Ecole Nationale Vétérinaire, 23 Chemin des Capelles, 31076 Toulouse Cedex 3, France

A B S T R A C T

Background: Reference limits are some of the most widely used tools in the medical decision process. Their determination is long, difficult, and expensive, mainly because of the need to select sufficient numbers of reference individuals according to well-defined criteria. Data from hospitalized patients are, in contrast, numerous and easily available. Even if all the information required for a direct reference interval computation is usually not available, these data contain information that can be exploited to derive at least rough reference intervals.

Methods: In this article, we propose a method for the indirect estimation of reference intervals. It relies on a statistical method which has become a gold-standard in other sciences to separate components of mixtures. It relies on some distributional assumptions that can be checked graphically. For the determination of reference intervals, this new method is intended to separate the healthy and diseased distributions of the measured analyte. We assessed its performance by using simulated data drawn from known distributions and two previously published datasets (from human and veterinary clinical chemistry).

Results and discussion: The comparison of results obtained by the new method with the theoretical data of the simulation and determination of the reference interval for the datasets was good, thus supporting the application of this method for a rough estimation of reference intervals when the recommended procedure cannot be used.

1. Introduction

Reference values are some of the most powerful tools in medical decision-making both in human and veterinary medicine, even though decision limits are being introduced in an increasing number of cases. Recently, the recommendations for "Determining, establishing, and verifying reference intervals in the clinical laboratory" have been updated by the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) and the Clinical and Laboratory Standards Institute (CLSI) [1].

This working group acknowledges that the *de novo* determination of reference limits, and their regular updating for the main partitioning groups is an enormous, time-consuming and very expensive undertaking, even in human clinical pathology, for which very large data bases are available, *e.g.* through Preventive Medicine Centers [2]. This task is even more difficult in animal clinical pathology, due to the various species, breeds, breeding conditions, productions, etc., and often the small number of animals available, *e.g.* in wild species. Moreover, the absence of well documented reference intervals (RI) for many analytes sometimes makes it impossible to transfer or verify previously established reference intervals in veterinary clinical pathology.

When direct sampling is not possible, the recommendation states that indirect sampling techniques may be used "based on the assumption, confirmed by observation, that most results, even on hospital and clinic patients appear "normal"" [1]. Data thus obtained, preferably from relatively healthy individuals, can be used for calculations of reference intervals. RI determined in this manner "should be considered rough estimates at best", as the datasets are contaminated by an unknown number of values obtained from individuals that are not healthy.

Very large clinical pathology databases exist in human hospitals, and also in veterinary hospitals and animal research centers. Several attempts have been made [3,4] to use them in the estimation of human reference intervals and some are still being pursued, often with very large numbers of data [5]. To our knowledge, few attempts have been made in animal clinical pathology [6].

Several statistical methods have been proposed to estimate the 2.5 and 97.5 centiles from such "polluted" datasets, the simplest ones being based on a cut-off value below or above which the observed data are discarded [7,8]. This approach is easy to implement but can

severely bias results. After recursively removing the values of individuals considered as diseased outside, Kairisto and Poola [8] modelled the distribution of the remaining values as two half Gaussian distributions. More sophisticated methods have been proposed (see [6] for a brief comprehensible review). Bhattacharya [9] proposed an indirect method based on the assumption that the distributions of healthy and diseased are Gaussian. As advocated by Baadenhuijsen and Smit [10] this method does not properly describe the distributions of most analytes which present skewed distributions. To solve this problem, Baadenhuijsen and Smit extended Bhattacharya's method to mixtures of log-normal and gamma distributions. These two methods are very practical since the estimations can be performed graphically without using a modern computer. Oosterhuis et al. [11] proposed a weighting scheme that decreased the influence of outliers in the Baadenhuijsen and Smit method. However, such approaches are specific to the chosen distributions and even for mixtures of Gaussian, log-Gaussian or gamma distributions, they do not give the maximum likelihood estimate which is the best estimate that it is possible to build when the number of data available (n) is large.

The aim of this paper is to propose a method that allows separation of the distributions of healthy and diseased individuals from observation of their mixture. This method generalizes and improves Bhattacharya-like methods.

As suggested in [1] when the actual patients status is known, we assume that there are two Box–Cox transformations that respectively make Gaussian the distributions of healthy and diseased individuals. We used this distributional assumption to estimate the percentage of healthy individuals and the distributions of healthy and diseased individuals, from which we derived a reference interval.

2. The proposed method

Let us first describe intuitively how the proposed method works. Assume that high values are rather observed on diseased individuals. A high value has thus a small probability to have been observed in a healthy individual and a high probability in a diseased individual. The model (Eq. (1)) described hereafter allows to formally quantify these probabilities. The distribution of healthy is computed as if all observed values came from healthy individuals. However, during this computation, each observed value is weighted by the probability that it has been obtained on a healthy individual. The same process is used for the diseased distribution. As an example, a very high value has a small probability (≈ 0) to come from a healthy individual. Its weight during the computation of healthy distribution is ≈ 0: this means that this value is discarded. One can see that proceeding so leads to a kind of circular reasoning: we need to know the distribution of healthy/diseased to compute the probability that a value has been obtained on a healthy/diseased individual while knowing these probabilities allows to compute these distributions. The method we propose iterates this idea until there is a strict accordance between the probabilities used to perform the computations and the distributions thus obtained. We now describe how to do it practically.

Let us assume that one observes Y_i on the ith individual of the sample

$$Y_i = U_i X_i^1 + (1 - U_i) X_i^2, i = 1, \ldots, n \tag{1}$$

where U_i is the unobserved status of this individual equal to 1 when he is healthy and 0 otherwise. Consequently, when the individual is healthy, one observes X_i^1 while X_i^2 is observed when he is diseased. The distribution of Y_i is therefore a mixture of the distributions for healthy and diseased individuals.

We assume that the random variables X_i^j that appear in the model (Eq. (1)) are mutually independent, independent of U_i and respectively distributed according to a $N(m_j,\sigma_j^2)$ up to a Box–Cox transformation $k_{\lambda j}$ where

$$k_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0 \\ \log(x) \text{ if } \lambda = 0. \end{cases}$$

If the actual status U_i of the ith individual was known, the probability density function (pdf) of the observation Y_i would be

$$\gamma_{m,\sigma^2,\lambda}(y_i) = \frac{y_i^{\lambda - 1}}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(k_\lambda(y_i) - m)^2\right]$$

with $(m,\sigma^2,\lambda) = (m_1,\sigma_1^2,\lambda_1)$ if $U_i = 1$ and $(m,\sigma^2,\lambda) = (m_2,\sigma_2^2,\lambda_2)$ otherwise.

In this case, one would estimate the reference interval

$$\left[\left(1 + \hat{\lambda}_1\left(\hat{m}_1 - 2\hat{\sigma}_1\right)\right)^{1/\hat{\lambda}_1}; \left(1 + \hat{\lambda}_1\left(\hat{m}_1 + 2\hat{\sigma}_1\right)\right)^{1/\hat{\lambda}_1}\right] \tag{2}$$

where $\hat{m}_1$, $\hat{\sigma}_1^2$, and $\hat{\lambda}_1$ are the parameter estimations computed from the data of healthy individuals only.

Since the U_i's are not observed, we need to estimate the percentage of healthy individuals p and the parameters of the "healthy" distribution $(m_1,\sigma_1^2,\lambda_1)$. The healthy distribution can be well separated from the diseased distribution when the parameters from the "diseased" distribution $(m_2,\sigma_2^2,\lambda_2)$ are also estimated. We thus need to estimate the parameter $\theta = (p,m_1,\sigma_1^2,\lambda_1,m_2,\sigma_2^2,\lambda_2)$ to get an estimation of the reference interval.

The best way to proceed is to use the maximum likelihood estimate of the parameter θ that can be obtained as the value of θ that maximizes

$$L(\theta) = \prod_{i=1}^{n} \left(p\gamma_{m_1,\sigma_1^2,\lambda_1}(y_i) + (1 - p)\gamma_{m_2,\sigma_2^2,\lambda_2}(y_i)\right).$$

Unfortunately, the direct optimisation of this function is often intractable. This is the reason why we suggest using the so-called EM algorithm [12] to solve it. The EM algorithm consists of iterations of an Expectation and a Maximization step. At the kth iteration, the E step computes the conditional expectation of the log-likelihood of the complete data (Y, U) with respect to the distribution of the missing, or non-observed data U given the observed data Y at the current estimated parameter value $\theta^{(k)}$:

$$Q\left(\theta, \theta^{(k)}\right) = E\left(\log P(Y, U) \mid Y, \theta^{(k)}\right).$$

The M step finds $\theta^{(k+1)}$ so that

$$\theta^{(k+1)} = \arg\sup_\theta Q\left(\theta, \theta^{(k)}\right).$$

These two-step iterations are repeated until convergence [13].

In our problem, the actual status U_i of each individual is not observed. Estimating the mixture of distributions of diseased and healthy individuals with the EM algorithm amounts to repeating the following iterations:

[1] For each individual compute the weights w_i and v_i that respectively depend on the likelihood that this individual belongs to the healthy and diseased group.
[2] Compute the mean, variance and λ parameter of each group by affecting to each individual the corresponding weight.
[3] Repeat steps 1 and 2 until the estimated parameters no longer change between two successive iterations.

This problem can be reduced to the following iterations. The algorithm starts with some initial value $\theta^{(0)}$ of $\theta = (p,m_1,\sigma_1^2,\lambda_1,m_2,\sigma_2^2,\lambda_2)$. At iteration k of the algorithm, $\theta^{(k)} = (p^{(k)},m_1^{(k)},\sigma_1^{2(k)},\lambda_1^{(k)},m_2^{(k)},\sigma_2^{2(k)},\lambda_2^{(k)})$ is known and $\theta^{(k+1)}$ is computed from $\theta^{(k)}$ by minimizing the following function with respect to θ:

$$\sum_{i=1}^{n} a_i^{(k)}\left[2(1 - \lambda_1)\log y_i + \log\sigma_1^2 + \left(k_{\lambda_1}(y_i) - m_1\right)^2/\sigma_1^2\right] + \left(1 - a_i^{(k)}\right)\left[2(1 - \lambda_2)\log y_i + \log\sigma_2^2 + \left(k_{\lambda_2}(y_i) - m_2\right)^2/\sigma_2^2\right]$$

where

$$a_i^{(k)} = \frac{p^{(k)}\gamma_{m_1^{(k)},\sigma_1^{2(k)},\lambda_1^{(k)}}(y_i)}{p^{(k)}\gamma_{m_1^{(k)},\sigma_1^{2(k)},\lambda_1^{(k)}}(y_i) + (1 - p^{(k)})\gamma_{m_2^{(k)},\sigma_2^{2(k)},\lambda_2^{(k)}}(y_i)}.$$

Solving this problem leads to the following iterations

$$\begin{cases} p^{(k+1)} = \dfrac{1}{n}\sum_{i=1}^{n} a_i^{(k)} \\ w_i^{(k)} = a_i^{(k)} / \sum_i a_i^{(k)} \text{ and } v_i^{(k)} = \left(1 - a_i^{(k)}\right) \Big/ \left(n - \sum_i a_i^{(k)}\right) \\ \lambda_1^{(k+1)} = \arg\inf_\lambda \log\sum_i w_i^{(k)}\left(k_\lambda(y_i) - m_1^{(k)}\right)^2 - 2\lambda\sum_i \log y_i \\ m_1^{(k+1)} = \sum_i w_i^{(k)} k_{\lambda_1^{(k+1)}}(y_i) \\ \sigma_1^{2(k+1)} = \sum_i w_i^{(k)}\left(k_{\lambda_1^{(k+1)}}(y_i) - m_1^{(k+1)}\right)^2 \\ \lambda_2^{(k+1)} = \arg\inf_\lambda \log\sum_i v_i^{(k)}\left(k_\lambda(y_i) - m_2^{(k)}\right)^2 - 2\lambda\sum_i \log y_i \\ m_2^{(k+1)} = \sum_i v_i^{(k)} k_{\lambda_2^{(k+1)}}(y_i) \\ \sigma_2^{2(k+1)} = \sum_i v_i^{(k)}\left(k_{\lambda_2^{(k+1)}}(y_i) - m_2^{(k+1)}\right)^2. \end{cases}$$

These iterations are continued until the difference between $\theta^{(k)}$ and $\theta^{(k+1)}$ becomes small. Let us denote $\hat{\theta} = (\hat{p}, \hat{m}_1, \hat{\sigma}_1^2, \hat{\lambda}_1, \hat{m}_2, \hat{\sigma}_2^2, \hat{\lambda}_2)$ the value of $\theta^{(k)}$ after convergence of the preceding iterations. The reference interval is then obtained as described by equation (Eq. (2)).

As we use the maximum likelihood, this reference interval is the best estimate (with respect to the precision) that can be built when the distributional assumptions we made are reasonable. It is therefore important to check these distributional assumptions. We propose hereafter a test (*P*-value) and a graphical means of checking that fully exploits the statistical properties of $\hat{\theta}$.

A *P*-value lower than 0.05 indicates that the assumptions about the distributions in the model (Eq. (1)) are not adequately chosen. Two main reasons lead to such *P*-values: 1) the choice of the distributions is not consistent with the data suggesting other distributions to be used (*e.g.* gamma, Weibull,…), 2) there exist "outliers". Their effect on the estimation of the upper limit of the reference interval seems to be limited as soon as a rigorous analysis is performed. When the test rejects the model, some extreme values should be removed from the analysis.

If our model (Eq. (1)) is correct, the actual distribution of the data, summarized by its cumulative distribution function (cdf) *F*, should be equal to the one implied by the model:

$$F_\theta(y) = p\Phi\left(\frac{k_{\lambda_1}(y) - m_1}{\sigma_1}\right) + (1-p)\Phi\left(\frac{k_{\lambda_2}(y) - m_2}{\sigma_2}\right) \quad (3)$$

where Φ is the cdf of the standard Gaussian distribution.

Thus, if this model is correct, *F* and F_θ should be equal. Since both *F* and F_θ cannot be observed, we propose to compare their estimates. An estimate of *F* is given by its empirical cumulative distribution function (cdf)

$$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{[Y_i \le y]},$$

where $1_{[Y_i \le y]} = 1$ if $Y_i \le y$ and 0 otherwise.

On the other hand, an estimate of F_θ is $F_{\hat{\theta}}$. A quick way to proceed is to build a QQ-plot by representing $F_{\hat{\theta}}^{-1}(\hat{F}(y_i))$ as a function of y_i. When the points obtained are about on a straight line, the model can be considered as reasonable. However and even if the chosen model is the good one, such graphics always exhibit points far away from the line because the sampling variation has not been taken into account to build it. Also, a confidence region/a *P*-value is helpful to compare these cdf. We propose to build a 95% uniform region $R_{0.95}$ for the QQ-Plot that can be used as follows: when at least one point $(y_i, F_{\hat{\theta}}^{-1}(\hat{F}(y_i)))$ falls outside $R_{0.95}$ there is less than 5% risk that $F \neq F_\theta$. Equivalently, we propose to compute a *P*-value by simulating the distribution of $\sup_y |F_{\hat{\theta}}(y) - \hat{F}(y)|$ under H_0: $F = F_\theta$.

To perform these simulations, we used the properties of the maximum likelihood estimator: the Central Limit Theorem guarantees that when *n* is large, $\sqrt{n}\left(\hat{\theta} - \theta\right)$ is approximately distributed according to the following Gaussian distribution $N(0, I^{-1}(\theta))$ where $I(\theta)$ is the Fisher information matrix. This matrix depends on the true value of θ which is unknown. Louis [14] proposed a method that gives a good approximation of $I(\theta)$ that we denote by $\hat{I}$.

We therefore propose to build $R_{0.95}$ using Monte-Carlo simulations. First, draw a large number (K_1) of θ_j (say $K_1 = 10\,000$) from $N(\hat{\theta}, \hat{I}^{-1}/n)$ and for each θ_j, we simulated $K_2 = 50$ samples of size *n* using (Eq. (1)). We thus obtained $K_1 \times K_2$ samples of size *n*, each of them giving an empirical cdf $\hat{G}$ and the distance $\sup_y |F_{\hat{\theta}}(y) - \hat{G}(y)|$. The *P*-value of the test is obtained as the percentage of $\sup_y |F_{\hat{\theta}}(y) - \hat{G}(y)|$ greater than $\sup_y |F_{\hat{\theta}}(y) - \hat{F}(y)|$. After discarding the 5% simulated samples with the highest distance $\sup_y |F_{\hat{\theta}}(y) - \hat{G}(y)|$, we built the QQ-plot for each of the 95% remaining samples. The envelope of these QQ-plots gives $R_{0.95}$.

After the distributional assumptions of the model (Eq. (1)) have been checked, the standard error of the estimates of the reference limits are computed as the standard deviation of the K_1 reference limits that can be computed using equation (Eq. (2)) but with the Monte-Carlo sample θ_j instead of $\hat{\theta}$.

3. Performance assessment

The objective of this simulation was to create samples from a predefined model distribution and to see whether or not the proposed method was able to recover this known distribution from the samples thus created.

The following parameters were used for the model distribution $p = 0.9$, $m_1 = 3.1$, $\sigma_1^2 = 0.014$, $\lambda_1 = -0.18$, $m_2 = 2.90$, $\sigma_2^2 = 0.022$, $\lambda_2 = -0.26$. These values are not purely fictive and, except for the value of *p*, come from example 2 given hereafter. In this distribution, the 2 subpopulations of healthy and diseased subjects are not clearly separated and greatly overlap.

The true reference limits of the healthy subpopulation are given in Table 1. As they have been determined from a distribution model and are purely theoretical, there is no confidence interval for these values.

For each sample, the method gives a parameter estimate, the mean of which is given in Table 1 for the simulations made as below. The actual imprecision of the calculation is measured from the sample-to-sample variation of this estimate. The standard deviation of this variation is the standard error of the estimate (fourth line of Table 1). In practice, only one single sample of size *n* is available for the calculations, which makes it impossible to evaluate the sample-to-sample variation. However, when *n* is large enough, it is possible to calculate an approximation of this sample-to-sample variation using the information matrix. The average of such approximations over samples is named the asymptotic standard error (a.s.e.) and is given in the second line of Table 1.

We simulated 100 samples of size $n = 1000$ and $n = 10{,}000$ and used our method to separate the distributions of healthy and diseased individuals. We estimated p and the limits of the reference interval of the "healthy" distribution thus separated. Estimates of *p* and of the reference limits from the 1000 and 10,000-samples are presented in Figs. 1 and 2, and summarized in Table 1. In both cases, the means were very close to the theoretical values, and the imprecision of the estimates was two-fold higher for the 1000-sample than for the 10,000-sample series. When 10,000-samples were used, the minimum and maximum estimations of *p* were 0.844 and 0.944, whereas the true value was 0.9. Estimates of the lower limit of the reference interval were very precise and accurate, whatever the number of values. The more precise estimation of the upper limit of the reference interval was obtained with the 10,000-samples, ranging from 145.3 to 167.4, whereas the true value was 155.76. This large imprecision can probably be explained by the position of the actual upper limit of the reference interval which is about at the mode of the distribution of diseased individuals. We can see in Fig. 2 that estimations of the upper limit are more often lower than the actual value of this limit. This suggests that the distribution of the corresponding estimator is as skewed as the distribution of the extreme values.

The estimates of the reference limits given by our method are unbiased since the mean of the 100 estimates is close to the true value for the 3 parameters. It can also be observed that the standard error of the estimator is very close to the a.s.e. But a standard statistical result shows that the maximum likelihood is the best estimate that it is possible to build when *n* is large. Consequently, no unbiased estimate with a smaller standard error exists, which implies that there is no better estimation than the one thus provided.

The parameters that define the distribution of diseased individuals are not well estimated in the above example. This can be explained the low number of data available for the diseased individuals. The population contains only 10% of diseased individuals therefore, the samples of size $n = 1000$ and 10,000 contained about 100 and 1000 diseased individuals. These numbers of diseased individuals appear too small for a precise estimation.

4. Two practical "real" examples

In this section, we show the results obtained when this method was applied to two previously reported examples.

4.1. Example of plasma TSH concentration in human males

An indirect determination of the plasma TSH upper reference limit was carried out using more than 19,000 unselected hospital results [15]. A 2056-value sample was selected by the authors comprising males for which repeat analyses had been discarded and the corresponding histogram was published cf Fig. 1B of [15]. The estimated upper limit, using the method proposed by Kairisto and Poola [8] implemented in the program GraphROC, was 3.5 mIU/L. The

Table 1
Parameters of the simulated model distribution and estimations of the percentage of healthy individuals and of the reference interval using the proposed method.

	$n=1000$			$n=10{,}000$		
	p	Lower lim.	Upper lim.	p	Lower lim.	Upper lim.
True value	0.90	53.73	155.76	0.90	53.73	155.76
a.s.e.	0.094	1.01	10.78	0.034	0.33	3.43
Mean	0.899	53.53	155.8	0.886	53.34	154.49
s.e.	0.045	1.18	9.62	0.022	0.6	5.19

(p = percentage of healthy individuals; Lower and Upper lim. = the limits of the reference interval; a.s.e. = average of asymptotic standard errors; s.e. = standard deviation of estimates obtained with the 100 samples).

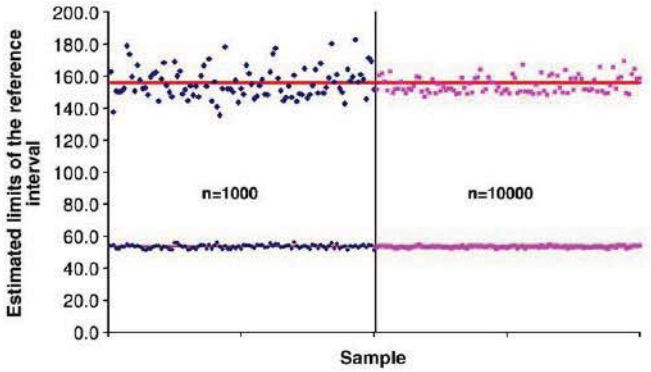

Fig. 2. Estimated reference limits obtained from 100 samples of size $n=1000$ and 100 samples of size $n=10{,}000$. The horizontal lines are the actual reference limits. The lower reference limit is well estimated while the estimated upper reference limit has quite a large imprecision probably because a large percentage of diseased individuals had concentrations close to this limit.

only information available being a histogram, we did not have individual data. We thus simulated individual data by assuming that they were uniformly distributed within each histogram class. We assumed that values included in negative classes corresponded to results below the limit of quantification (LOQ) and deleted them before subsequent analyses. The distribution of values above 0 was assumed to be a mixture of two Gaussian distributions after Box–Cox transformations. The observed and fitted distributions thus obtained as well as the components of the mixture are presented in Fig. 3. The overall shape of the distribution is very similar to the original one. The percentage of individuals with a TSH concentration below the limit of quantification was 4.21%. Using the proposed method, the estimated percentage of healthy individuals with a TSH concentration above the limit of quantification was 77.10%. The distributions of plasma concentration of healthy (with TSH > LOQ) and diseased patients were estimated to be Gaussian after a Box–Cox transformation with parameters 0.2835 and − 1.433 respectively. Since 4.21% of individuals were assumed to be healthy because they had a value below the LOQ, the next step was to calculate the 97.50–4.21 = 93.29% quantile of the distribution of the healthy individuals to obtain the upper limit of the reference interval. This latter was estimated to be 3.770 ± 0.333 mIU/L which is slightly higher but very close to the estimate provided in the original study. The diagnostic plot represented in Fig. 4 shows that the proposed model is acceptable ($P>0.05$) as the observed QQ-Plot is entirely contained within the confidence region.

4.2. Example of creatinine concentration in dog

A study to determine the diagnostic efficiency of plasma creatinine and urea concentrations for the diagnosis of canine kidney diseases finally included 3822 cases, of which 37% were healthy [16]. The nonparametric reference interval was 53–151 µmol/L (90% confidence intervals were 52 to 55 µmol/L and 148 to 159 µmol/L). As discussed in this article, even if the health status of these animals was known, there was no way of ensuring the quality of the diagnoses. We therefore decided to ignore the available diagnoses in the present study. The observed distribution of plasma creatinine in the 3822 dogs is represented in Fig. 5. The diagnostic test described in the Proposed methods section gave a P-value = 0.580 thus showing that the proposed model is acceptable.

The estimated percentage of healthy individuals was $\hat{p}=79.93\%$, quite higher than the "true" percentage. The plasma concentration distribution of healthy patients was estimated to be $N(3.100;0.014)$ after a Box–Cox transformation with parameter − 0.179. This led to an upper limit of the reference interval of 160.9 ± 8.7 µmol/L, which is close to the limit obtained in healthy dogs based on reported clinical status [16].

When the method proposed by Baadenhuijsen and Smit [10] was applied to the same data, the estimated upper limit was 135.1 µmol/L; for a mixture of gamma distributions. Using the same method but with the weighting scheme proposed by Oosterhuis et al. [11] we obtained 142.6 µmol/L. In contrast the method of Kairisto and Poola (with no use of external information) [8] gave 340.8 µmol/L.

These large differences between the estimations could be explained: 1/ by the high sensitivity of Baadenhuijsen's method to the choice of aligned points and 2/ by the threshold implied by the ±4 SD rule in the Kairisto and Poola method that keeps the data from diseased individuals

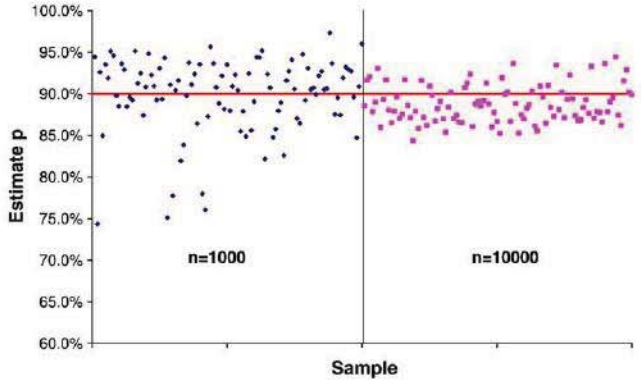

Fig. 1. Estimated percentage of healthy individuals obtained from 100 samples of size $n=1000$ and 100 samples of size $n=10{,}000$. The horizontal line is the value that was used for simulations. This percentage is well estimated; the imprecision decreases with n.

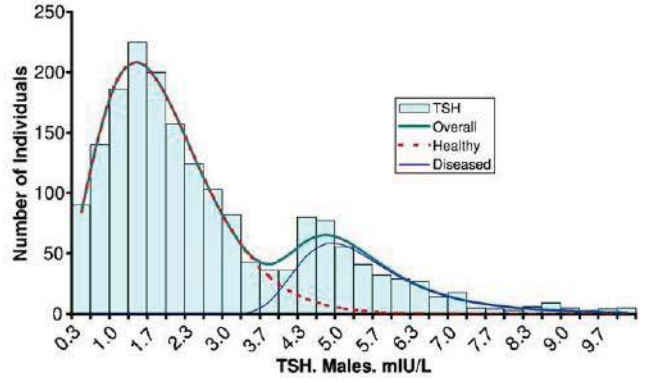

Fig. 3. The histogram represents the observed TSH concentrations in human males obtained from [16]. The concentrations below the LOQ have not been represented (but taken into account for the RI determination (see text). The dotted, thin and thick curves represent the TSH distribution obtained using the proposed method for the diseased, healthy individuals and for the entire population respectively.

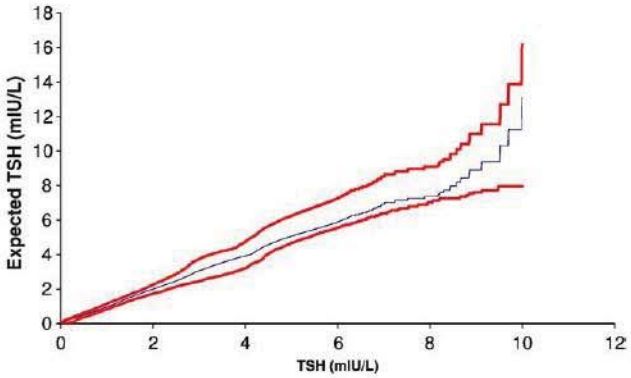

Fig. 4. Diagnostic plot proposed in Section 2 for the TSH data. The empirical cdf F(thin line) is totally included within the 95% confidence region (thick lines) or ($P>0.05$). The assumptions about the model (shape of distributions) are thus in accordance with the data.

for computation when no manual exclusion of data is performed. On this example, the modification of the Bhattacharya method proposed by Oosterhuis, Modderman and Pronk produced a reasonable estimate but its imprecision cannot be computed.

5. Discussion

In this article we propose a method for estimating reference interval of an analyte from a large set of data when the actual status of the individuals in the analyzed sample is unknown. This method is not new; it is currently considered as the gold-standard method for mixtures analysis.

This method compensates some of the deficiencies of existing methods used to determine reference intervals. Bhattacharya and Baadenhuijsen's methods assume specific shapes, mainly Gaussian and log-Gaussian, for the distributions of healthy and diseased. These methods are not flexible enough to adapt to other distribution shapes, unlike the proposed method. With these methods the computation of reference intervals relies mainly on a graphical analysis that requires a subjective choice of aligned points in a scatter plot. This choice strongly influences the results obtained and the confidence intervals for the provided estimates cannot be computed. Actually, these relatively "old" methods were well adapted to a time when computations were not performed with computers.

When the actual distributions have a shape close to the assumed one, the method that we propose provides, for large n, the maximum likelihood estimation of the reference interval, *i.e.* the most precise unbiased estimate that can be computed.

We also propose a method that allows rapid (and visual) checking of whether or not these assumptions about shapes of distributions are reasonable.

Although this method is an improvement on existing ones, it has some weaknesses that restrict its use. First of all, it requires the distributions to be Gaussian up to a Box–Cox transformation. This assumption is essential for the method to give unbiased results.

Secondly, the method is designed to identify two subpopulations and it succeeds in doing this even if the subpopulations are not the ones which were expected. As an example, we used this method on another dataset and it separated young *versus* old individuals instead of healthy *versus* diseased ones.

Thirdly, this method correctly separates healthy individuals, but as previously stated, the limits thus obtained are "rough estimates at best" [1]. Even when this new method which minimizes the imprecision of such estimates was used, the s.e. of the upper limit was within the range of 5 to 10% in the simulations and in the examples studied, which means, that the 90% confidence interval of this limit was in the range of [140.0; 171.6] for $n = 1000$.

Fourthly, this method cannot be used to analyse serial measurements in the same individual. As the method assumes that the random variables are mutually independent, these data should be discarded from the analysis. In practice, omitting repeated measurements often reduces the number of available data.

Finally, this method estimates the distribution of the "minor" *i.e.* less represented subpopulation, with greater imprecision than that of the major. This algorithm converges (*i.e.* stabilizes after some iterations) only when the distributions of healthy and diseased individuals are different enough with regards to the sample size n. Even very close distributions can be distinguished when n is very large. On the contrary, this algorithm may fail to estimate quite different distributions when only a small sample size is available. This problem is connected with the general problem of identifiability of mixtures [17].

A large number of refinements and extensions can be considered. It is likely that the skewness we assumed (Gaussian after a Box–Cox

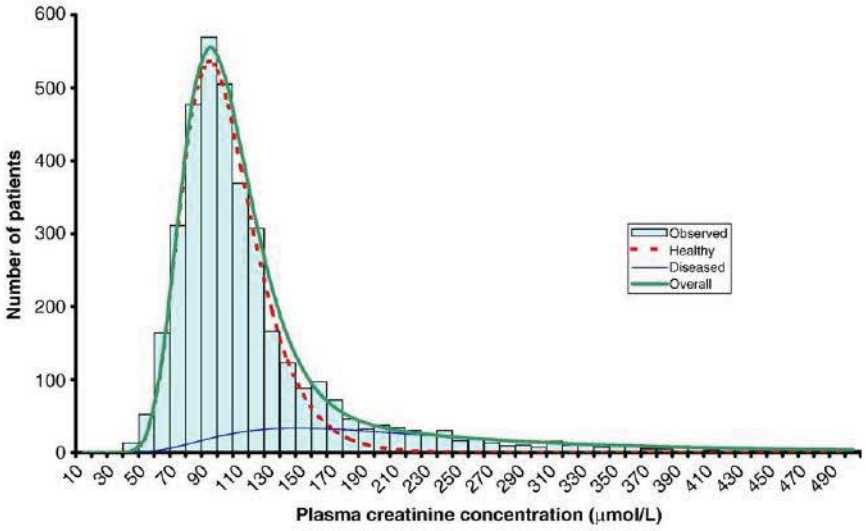

Fig. 5. Histogram of the plasma creatinine concentrations in dogs obtained from [17]. The dotted, thin and thick curves represent the creatinine distribution obtained for diseased animals, for healthy animals and for the entire population respectively, using the proposed method.

transformation) for the distributions cannot cover all the situations encountered in practice. However, following exactly the same EM steps, other distributions can be used, such as the exponential family (*e.g.* gamma, Weibull,...). An interesting study would be to try distributions with different shapes and to use the graphical diagnostic plot proposed in the paper to identify the shapes that are compatible with the data.

Examples composed of 2 subpopulations (healthy *versus* diseased individuals) are simplistic. Actual populations are more complex, *e.g.* subpopulations of individuals affected by a different disease than the one for which the analyte has been measured. Thus, a natural extension of this work will be to build an algorithm that can deal with an unknown number of subpopulations.

Finally, demographic variables such as age, gender, weight could be used as covariates as in [18]. Even if these variables are not known for each individual, they would help identify the populations separated, thus minimizing the risk of wrongly "labelling" the identified subpopulations as healthy or diseased.

References

[1] CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline. Third ed. Wayne, PA: CLSI; 2008.

[2] Siest G. Study of reference values and biological variation: a necessity and a model for Preventive Medicine Centers. Clin Chem Lab Med 2004;42:810–6.

[3] Neumann GJ. The determination of normal ranges from routine laboratory data. Clin Chem 1968;14:979–88.

[4] Glick JH. Statistics of patient test values: application to indirect normal range and to quality control. Clin Chem 1972;18:1504–13.

[5] Grossi E, Colombo R, Cavuto S, et al. The REALAB project: a new method for the formulation of reference intervals based on current data. Clin Chem 2005;51:1232–40.

[6] Ferre-Masferrer M, Fuentes-Arderiu X, Puchal-Ane R. Indirect reference limits estimated from patients' results by three mathematical procedures. Clin Chim Acta 1999;279:97–105.

[7] Petitclerc C. Normality: the unreachable star? Clin Chem Lab Med 2004;42:698–701.

[8] Kairisto V, Poola A. Software for illustrative presentation of basic clinical characteristics of laboratory tests GraphROC for Windows. Scand J Clin Lab Invest Suppl 1995;222:43–60.

[9] Bhattacharya CG. A simple method of resolution of a distribution into Gaussian components. Biometrics 1967;23:115–35.

[10] Baadenhuijsen H, Smit JC. Indirect estimation of clinical chemical reference intervals from total hospital patient data: application of a modified Bhattacharya procedure. J Clin Chem Clin Biochem 1985;23:829–39.

[11] Oosterhuis WP, Modderman TA, Pronk C. Reference values: Bhattacharya or the method proposed by the IFCC? Ann Clin Biochem 1990;27:359–65.

[12] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 1977;39:1–38 (With discussion).

[13] Wu CF. On the convergence properties of the EM algorithm. Ann Stat 1983;11:95–103.

[14] Louis TA. Finding the observed information matrix when using the EM algorithm. J R Stat Soc B 1982;44:226–33.

[15] Giavarina D, Dorizzi RM, Soffiati G. Indirect methods for reference intervals based on current data. Clin Chem 2006;52:335–6.

[16] Concordet D, Vergez F, Trumel C, Diquélou A, Lanore D, et al. A multicentric retrospective study of serum/plasma urea and creatinine concentrations in dogs using univariate and multivariate decision rules to evaluate diagnostic efficiency. Vet Clin Pathol 2008;37:96–103.

[17] Holzmann H, Munk A, Gneiting T. Identifiability of finite mixtures of elliptical distributions. Scand J Statist 2006;33:753–63.

[18] Royston P, Wright EM. Goodness-of-fit statistics for age-specific reference intervals. Stat Med 2000;19:2943–62.