Clinical Chemistry 51:12
2326–2332 (2005)

General Clinical Chemistry

Detection of Outliers in Reference Distributions: Performance of Horn's Algorithm

Helge Erik Solberg[1*] and Ari Lahti[2]

Background: **Medical laboratory reference data may be contaminated with outliers that should be eliminated before estimation of the reference interval. A statistical test for outliers has been proposed by Paul S. Horn and coworkers (*Clin Chem* 2001;47:2137–45). The algorithm operates in 2 steps: (*a*) mathematically transform the original data to approximate a gaussian distribution; and (*b*) establish detection limits (Tukey fences) based on the central part of the transformed distribution.**

Methods: **We studied the specificity of Horn's test algorithm (probability of false detection of outliers), using Monte Carlo computer simulations performed on 13 types of probability distributions covering a wide range of positive and negative skewness. Distributions with 3% of the original observations replaced by random outliers were used to also examine the sensitivity of the test (probability of detection of true outliers). Three data transformations were used: the Box and Cox function (used in the original Horn's test), the Manly exponential function, and the John and Draper modulus function.**

Results: **For many of the probability distributions, the specificity of Horn's algorithm was rather poor compared with the theoretical expectation. The cause for such poor performance was at least partially related to remaining nongaussian kurtosis (peakedness). The sensitivity showed great variation, dependent on both the type of underlying distribution and the location of the outliers (upper and/or lower tail).**

Conclusion: **Although Horn's algorithm undoubtedly is an improvement compared with older methods for outlier detection, reliable statistical identification of outliers in reference data remains a challenge.**

Medical laboratory reference data may be contaminated with erroneous values that should be eliminated before estimation of the reference interval and other types of statistical treatment *(1, 2)*. If erroneous values are hidden within the distribution of reference data, these values can be detected and removed only by following a strict protocol for production of reference values. However, erroneous values that deviate significantly from the proper reference values (outliers) might be identified by statistical techniques. The ideal algorithm for outlier detection should find any number of deviating values in each tail of the reference distribution and should operate equally well when applied to various forms of probability distribution. The plethora of proposed statistical outlier tests *(3, 4)* shows that it is difficult to find a single algorithm that can handle all relevant situations.

The simple Dixon range test *(5)*, i.e., identify the extreme value as an outlier if the difference between the 2 highest (or lowest) values in the distribution exceeds one third of the range of all values, was proposed by the IFCC in the recommendation for statistical treatment of reference values *(1)* and included in previous versions of the RefVal program *(6, 7)*. This method is reasonably insensitive to distribution type, but it has the major drawback of being unable to successfully handle clusters of 2 or more outliers.

A more promising method was proposed by Horn et al. *(8)*. This method (hereafter referred to as Horn's algorithm) is based on 2 general assumptions: (*a*) that the central part of the distribution contains most of the information of the genuine reference values; and (*b*) that outliers may be detected as values lying outside limits based on the properties of this central part. The algorithm operates in 2 steps: In the first step, the original data are transformed to approximate a gaussian distribution, to the extent this is possible in the presence of outliers. Horn et al. *(8)* used for this purpose the Box–Cox function *(9)*. In the second step, 2 detection limits (fences) are estab-

[1] Department of Medical Biochemistry, Rikshospitalet-Radiumhospitalet HF, Oslo, Norway.
[2] Department of General Psychiatry, University Hospital of Northern Norway, Tromsø, Norway.
* Address correspondence to this author at: Bergersletta 28, N-1349 Rykkinn, Norway. E-mail heesolbe@online.no.
Received July 27, 2005; accepted September 21, 2005.
Previously published online at DOI: 10.1373/clinchem.2005.058339

lished based on the middle 50% of the transformed distribution, as suggested earlier by Tukey *(10)*. Possible outliers are identified as the values located outside of these fences.

As we wanted to include Horn's algorithm in the RefVal program *(6, 7)*, which implements the IFCC-recommended methods for statistical treatment of reference values *(1)*, we used Monte Carlo computer simulations in a study of the specificity and sensitivity of the algorithm:

Specificity. In the absence of outliers, the ideal statistical test for outlier detection should provide a low and predictable probability of false detection regardless of the underlying distribution type. We studied this by means of simulation experiments using computer-generated distributions with known statistical properties and varying positive and negative skewness. We compared the specificity of the original Horn's algorithm with that observed for 2 modified versions of this algorithm, using other data transformations.

Sensitivity. Although the probability of detection of true outliers should ideally be high, one must—for inevitable statistical reasons—accept a compromise between sensitivity and specificity. When increasing the former, the latter decreases, and vice versa. We studied the sensitivity properties of the outlier test algorithms in similar experiments.

Materials and Methods

GENERATION OF RANDOM DATA

We implemented the Monte Carlo simulations of this study by adding appropriate pascal source code to the standard RefVal program *(6)*, using Borland Delphi 5 (Inprise Co.) as the development system. Gaussian-distributed pseudo-random numbers were generated by use of Delphi's built-in *RandG* function, initialized by a single call to Delphi's *Randomize* procedure, which takes a seed value from the system clock. Validation experiments (results not presented) showed that the quality of the gaussian distributions produced was sufficiently high for the present study. To avoid problems with negative data values when transforming the generated distributions and performing the outlier detection experiments, we displaced these distributions so as to make them also lie safely above zero at their lower end, except when χ^2 distributions were produced (see below).

Asymmetric distributions with various degrees of skewness were produced from the computer-generated gaussian distributions. To cover the spectrum of asymmetric distributions that is typical for clinical biochemical reference data, the following nongaussian distributions were examined in this study: the square root gaussian distribution, the logarithmic gaussian distribution, and χ^2 distributions of df = 3, 4, 8, and 16. The square root gaussian and the logarithmic gaussian distributions were produced by use of the transformations $x = g \cdot g$ and $x = \exp(g)$, respectively, where x is a value of a transformed distribution and g a gaussian-distributed random value with mean = 100 and SD = 25.51. The χ^2 distributions with df degrees of freedom were produced by summing up the squares of standard gaussian values (g_i): $x = \Sigma g_i^2$ ($i = 1, \ldots, df$; df = 3, 4, 8, and 16). All of these asymmetric distributions were positively skewed. To obtain corresponding distributions with negative skewness, values of the generated distributions were transformed into those of mirror distributions: $x'_i = w - x_i$, where $w = x_{min} + x_{max}$.

The details of these distributions are shown in Table 1.

TRANSFORMATIONS

Mathematical functions can transform data of nongaussian distributions to approximate the theoretical gaussian distribution. Three functions of this kind were used in our study. One of these, the Box–Cox transformation function *(9)*, which was used in the original Horn's algorithm for outlier detection *(8)*, is as follows: $y = (x^\lambda - 1)/\lambda$ if $\lambda \neq 0$; $y = \ln(x)$ if $\lambda = 0$. The parameter λ of this transformation was determined by maximizing likelihood [formula 8 in Ref. *(9)*]. The 2 other transformations considered in the present study are those of the 2-stage normalization procedure recommended by the IFCC for parametric estimation of reference limits *(1)*. Manly's exponential function *(11)* corrects for nongaussian skewness: $y = \{\exp(\gamma \cdot x) - 1\}/\gamma$ if $\gamma \neq 0$; $y = x$ if $\gamma = 0$. The John and Draper modulus function *(12)* rectifies remaining nongaussian kurtosis: $z = \text{sign}[\{(|y| + 1)^\delta - 1\}/\delta]$ if $\delta \neq 0$; $z = \text{sign}[\ln(|y| + 1)]$ if $\delta = 0$. [Here the sign (+ or −) is that associated with the value y, previously transformed by Manly's exponential function.] The 2 latter functions have always been part of the RefVal program *(6)*. The function parameters (γ and δ, respectively) were determined by use of an iterative "brute force" method, guided by monitoring the coefficient of skewness (the exponential function) or the coefficient of kurtosis (the modulus function).

HORN'S ALGORITHM AND ITS MODIFICATIONS

The algorithm described by Horn et al. *(8)* has the following consecutive steps:

(1) Transform the original data so as to achieve a distribution that is as close as possible to gaussian shape. The original algorithm used the Box–Cox transformation for this purpose.
(2) Estimate the lower and upper quartiles (Q_1 and Q_3, respectively) and the interquartile range (IQR = $Q_3 - Q_1$) for the transformed data.
(3) Define 2 Tukey fences *(10)*: $Q_1 - 1.5 \cdot \text{IQR}$ and $Q_3 + 1.5 \cdot \text{IQR}$.
(4) Identify as possible outliers all reference values located outside the 2 fences.

We also studied 2 modifications of Horn's algorithm. In the first step of the algorithm, we replaced the Box–Cox

transformation either with the exponential transformation or with a 2-stage transformation consisting of the exponential transformation followed by the modulus transformation.

SIMULATION EXPERIMENTS

Pseudo-random data were generated for the distributions described above (Table 1). All experiments presented here were based on distributions with n = 1000 values each. When studying the sensitivity of the outlier algorithm, we replaced 3% of the random values in the distributions by outliers, keeping the sample size fixed at 1000 values. Outliers were generated as uniform random values in the interval from 2.7 to 3.9 SD below and/or above the mean of the original gaussian distribution and then transformed (see section on generation of random data above).

We analyzed each data set by applying 2 (sensitivity study) or 3 (specificity study) of the versions of Horn's algorithm for outlier detection that were presented above, the original one (using the Box–Cox transforming function) and the 2 modifications (using the exponential transformation and the 2-stage transformation, respectively). The output monitored was the estimated transformation parameters, the coefficients of skewness and kurtosis of the transformed distributions, and the number of transformed values located below and above the lower and upper Tukey fences, respectively. The simulations were always iterated 6000 times for each distribution type.

STATISTICAL ANALYSIS

We computed the coefficients of skewness and kurtosis by applying established routines of the RefVal program *(13)*. The output from simulation experiments was analyzed with Microsoft Excel.

Results

SPECIFICITY STUDY

We studied the specificity of outlier detection in simulation experiments performed on 13 computer-generated distributions of various shapes (Table 1) without added outliers. The mean percentages of data values located outside the 2 Tukey fences, i.e., values falsely identified as outliers, are shown in Fig. 1. The top panel in Fig. 1 shows the results obtained for the original Horn's algorithm, which is based on Box–Cox transformation. The results for the 2 modifications of this algorithm described above are shown in the middle and bottom panels.

The observed percentages in Fig. 1 should be compared with the theoretical expected probability of false detection. For a gaussian distribution, the quartiles (Q_1 and Q_3) are located at a distance of $0.674 \cdot SD$ on both sides of the mean, giving $IQR = 2 \cdot 0.674 = 1.349 \cdot SD$. According to the formulas given above, the Tukey fences are thus $0.674 + 1.5 \cdot 1.349 = 2.698 \cdot SD$ below and above the mean. The cumulated gaussian probability at $-2.698 \cdot SD$ is 0.0035. The expected frequency of false detection is thus 0.70%. This is shown as a horizontal line in each panel of Fig. 1.

For the same simulation experiment, the remaining kurtosis after transformation by the Box–Cox and exponential functions is shown in Fig. 2. (The results for the 2-stage transformation are not shown because the coefficient of kurtosis necessarily always is zero.) The Box–Cox transformation failed to make negatively skewed distributions symmetric (see the coefficients of skewness in the top panel of Fig. 2).

SENSITIVITY STUDY

We studied the sensitivity of outlier detection in simulation experiments using probability distributions 4–10 described in Table 1, each having 3% of the values replaced by random outliers. Three types of experiments, with different locations of the outliers, were performed: (*a*) all outliers placed in the interval ($-3.9 \cdot SD$ to $-2.7 \cdot SD$) of the lower tail of the distribution; (*b*) all outliers placed in the interval ($2.7 \cdot SD$ to $3.9 \cdot SD$) of the upper tail of the distribution; and (*c*) one half of the outliers, i.e., 1.5% of the observations, placed in each of these 2 intervals. The inner limits of these intervals, $\pm 2.7 \cdot SD$, were set to coincide with the Tukey fences (see above). The cumulative gaussian probabilities at the limits $3.9 \cdot SD$ and $2.7 \cdot SD$ are 0.00005 and 0.0035, respectively,

Table 1. Properties of computer-generated probability distributions.[a]

	Original			Mirrored		
Distribution type	Sequence[b]	Skewness	Kurtosis	Sequence	Skewness	Kurtosis
Gaussian	7	0.01	0.00			
Square root gaussian	6	0.41	0.23	8	−0.41	0.22
χ^2 (df = 16)	5	0.70	0.73	9	−0.70	0.72
Logarithmic gaussian	4	0.87	1.38	10	−0.87	1.32
χ^2 (df = 8)	3	0.99	1.47	11	−1.00	1.47
χ^2 (df = 4)	2	1.41	2.94	12	−1.40	2.91
χ^2 (df = 3)	1	1.61	3.80	13	−1.62	3.92

[a] For each of the 13 distribution types used in this study, 1000 distributions with 1000 values were generated, as described in the text. The coefficients of skewness and kurtosis were estimated by established procedures of the RefVal program *(13)*. The values shown are the observed mean values for these coefficients. The distributions are presented in the order of increasing absolute value for the skewness.

[b] Sequence number used for identification of distributions in graphs (Figs. 1–3).

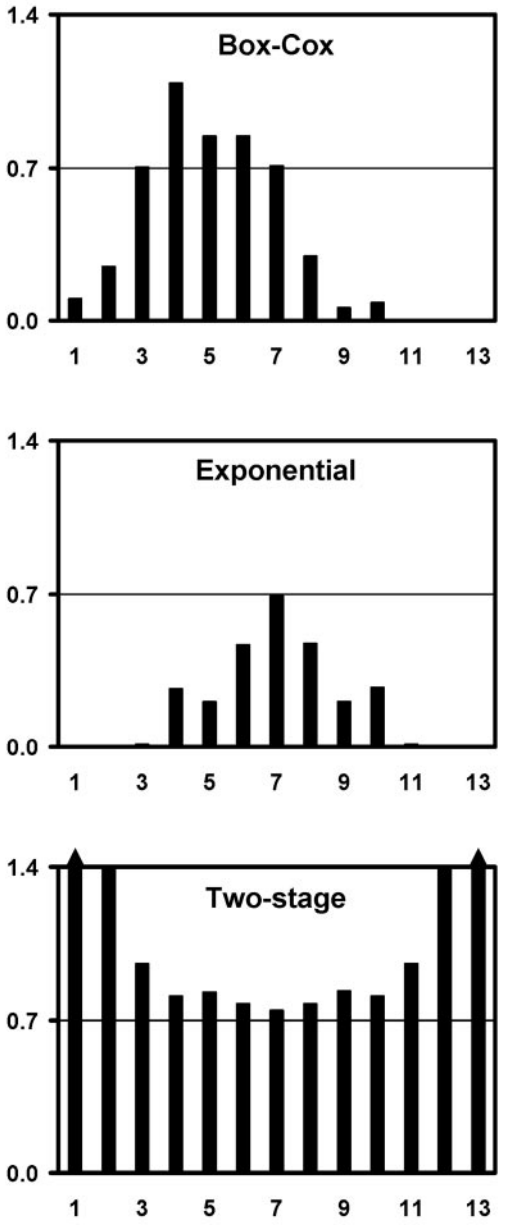

Fig. 1. Specificity of Horn's outlier algorithm.

Shown are mean percentages of false outliers (*y axis*) in simulation experiments in which 13 computer-generated distributions (*x axis*; see Table 1) were mathematically transformed with 3 functions, as explained in the text. Each distribution had 1000 data values. The simulation was iterated 6000 times. The *horizontal lines* indicate the expected percentage of data values located outside the Tukey fences (0.70%).

which shows that by setting the outer limits at $\pm 3.9 \cdot SD$, the suggested intervals for outliers will cover practically all probability outside the inner limits. The expected total percentage of observations identified as outliers will now be 0.70% (false outliers) + 3.0% (true outliers) = 3.7%. Of this total percentage, 3.35% (in absolute terms) should originate from the tail in which the generated random outliers were placed and 0.35% from the other tail, as far as experiment types (*a*) and (*b*) are concerned, whereas for

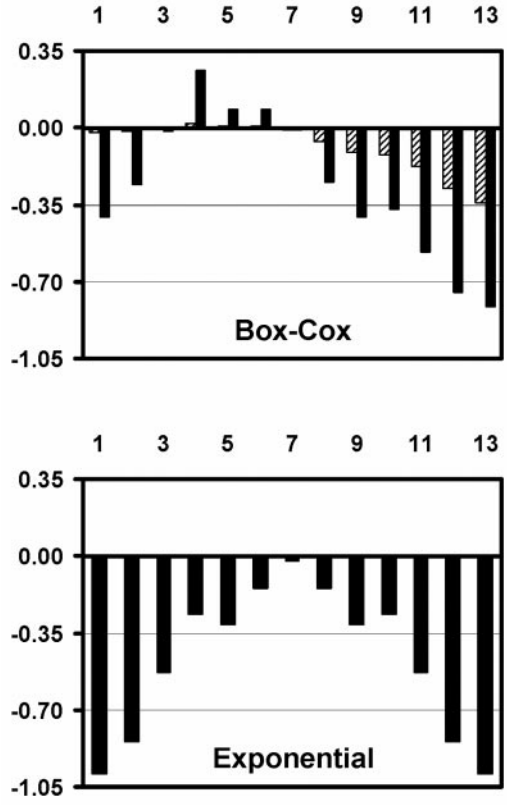

Fig. 2. Coefficients of skewness and kurtosis of transformed distributions.

Shown are mean values of the coefficients (*y axis*) for the 13 distributions (*x axis*), as obtained in the simulation experiment described in the legend of Fig. 1. ■, kurtosis; ▨ (*top panel only*), skewness.

the experiment type (*c*), the expectation is 1.85% from each tail.

The mean percentages of data values located outside the 2 Tukey fences, i.e., values identified as outliers, true or false, are shown in Fig. 3. The filled columns show results for the original Horn's algorithm (using Box–Cox transformation); the open columns show the corresponding results obtained for a modified algorithm (exponential transformation).

Discussion

The results presented here are all based on simulation experiments with a sample size of 1000. Although this size may be larger than that in many real-life studies of reference values, we chose it because it was large enough to reduce unwanted sample variation. Other experiments with smaller and larger samples (results not presented) gave similar findings, showing that the results of this study are valid regardless of the number of values in the distributions.

SPECIFICITY STUDY

A low and predictable probability of false detection is a basic requirement of statistical tests for outliers. In our specificity study, we tested Horn's algorithm and 2 mod-

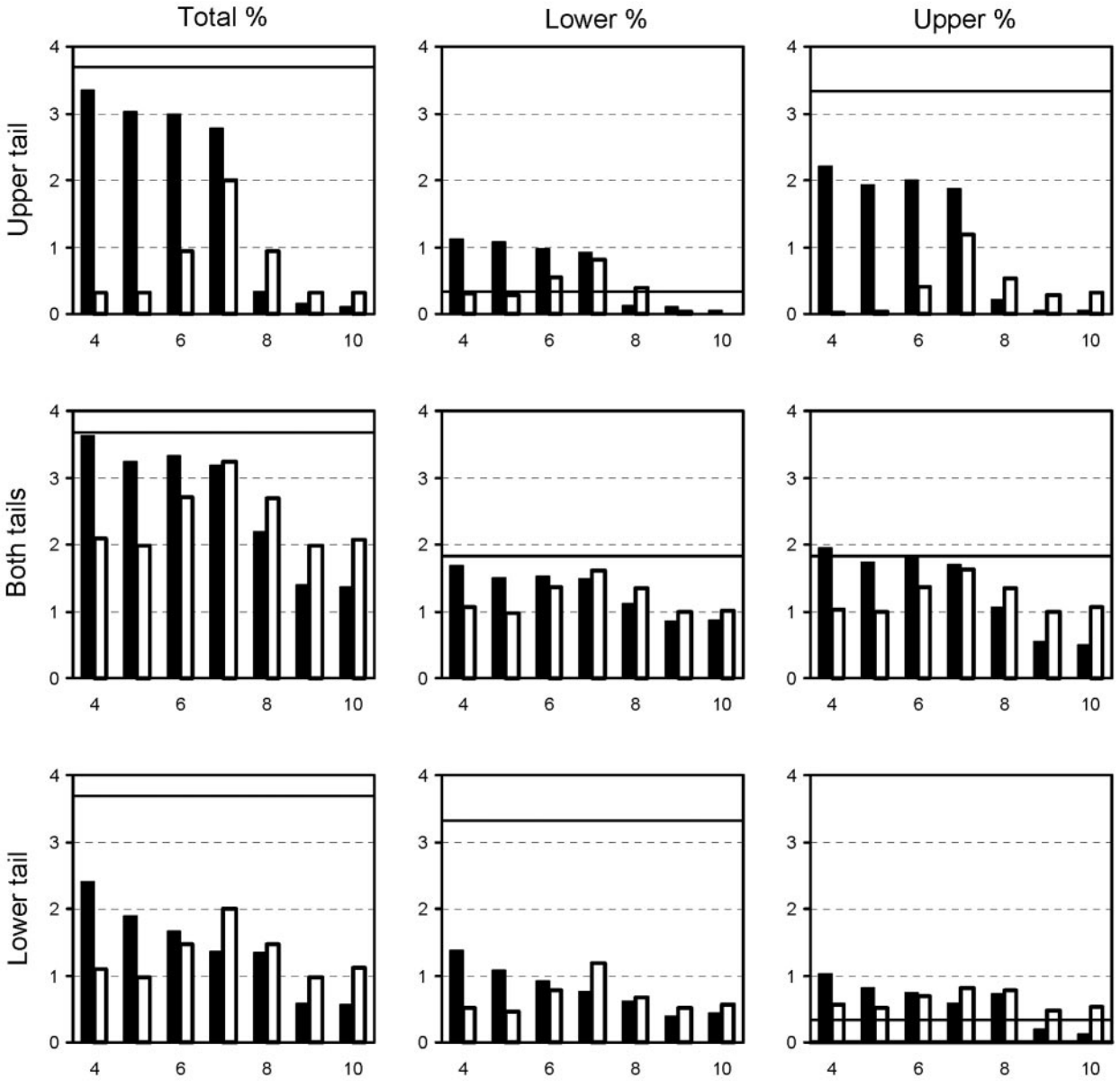

Fig. 3. Sensitivity of Horn's algorithm (original and modified).

Shown are the mean percentage of values identified as outliers (*y axis*) in simulation experiments in which 3% of the random values in computer-generated distributions (types 4–10; see Table 1; *x axis*) were replaced by outliers. The outliers were located in the upper tail (*top row*), in the lower tail *(bottom row*), or in both tails (*middle row*). The *panels* at the *right* and in the *middle* show the percentages of values located above the upper and below the lower Tukey fences, respectively; the sum is shown at the *left*. The *solid horizontal line* in each *panel* shows the respective theoretical expectations that would be obtained for a gaussian distribution. Two series of simulations are shown: ■, results with the original Horn's algorithm (Box–Cox transformation); □, modified algorithm (exponential transformation). Each distribution had 1000 data values, including the outliers. The simulation was iterated 6000 times.

ifications of it, using 13 different types of computer-generated probability distributions without generated outliers. These distributions were all unimodal, and they had a coefficient of skewness varying between −1.6 and 1.6 (Table 1). The distribution types with zero or moderately large positive skewness (distributions 3–7) are typical of distributions found in clinical chemistry. Empirical distributions with negative skewness (distributions 8–13) are admittedly very rare in laboratory medicine, but because they may potentially occur, they were included in the study to make it comprehensive.

None of the outlier tests studied, neither the original Horn's algorithm that uses Box–Cox transformation nor the 2 modifications of this algorithm involved in the present study and based on other transformations, fulfilled the basic requirement for outlier tests stated above, as is shown in Fig. 1. The theoretical expectation of 0.70% values falsely identified as outliers was obtained only when the distribution was gaussian (distribution 7) and, for the original Horn's algorithm, using the χ^2 distribution with $df = 8$ (distribution 3). With Horn's original algorithm (Fig. 1, top panel), the probability of false identifi-

cation was too high for distributions 4–6, which have the moderate positive skewness frequently found in medical data. It was particularly high for the logarithmic gaussian distribution (distribution 4), a very typical distribution in laboratory medicine. This test was very conservative for extreme positive skewness (distributions 1 and 2) and for all negatively skewed distributions (distributions 8–13). In contrast to this asymmetric behavior, the modified test, using the exponential transformation (middle panel of Fig. 1), handled positively and negatively skewed distributions in the same, conservative way. When the modulus transformation was added for correction of remaining kurtosis (bottom panel of Fig. 1), the percentage of false outliers was only slightly increased (0.77%–0.96%) from the expected value of 0.7%, assuming that the skewness was moderate (positive or negative; distributions 3–6 and 8–11).

The main cause for the varying performance of the outlier tests based on the Box–Cox and exponential functions was the remaining kurtosis after the transformation of data (Fig. 2). A symmetric distribution with negative kurtosis has a flat, central peak and fewer values in the tails than does the gaussian distribution. Therefore, the percentage of values outside Tukey fences will be lower than expected when the distribution after transformation has negative kurtosis. Positive kurtosis has the opposite effect. Comparison of the 2 upper panels of Fig. 1 with the respective panels of Fig. 2 illustrates this kurtosis effect.

Another problem with the original Horn's algorithm was that the Box–Cox transformation failed to produce a symmetric distribution when the original distribution was negatively skewed (Fig. 2, top panel; distributions 8–13). In such cases the algorithm will not handle values equally in the 2 tails of the distribution.

Horn et al. *(8)* correctly pointed out that automatic elimination of 0.70% of the reference values in a gaussian distribution may cause biased reference limits. They accordingly suggested to estimate a nominal 95% reference interval as a 95.67% interval. However, this recommendation is valid only if the transformation step of Horn's algorithm gives a truly gaussian distribution. Our results show that this is not the case for the majority of the distributions studied here.

In summary, the specificity of neither the original Horn's outlier test nor its modified versions will be predictable when analyzing empirical data because the performance of these tests is dependent on the underlying distribution type, which usually is unknown a priori.

SENSITIVITY STUDY

To get a manageable study of sensitivity for the outlier tests, we restricted it to the symmetric and moderately skewed distributions (distributions 4–10 in Table 1). In addition, we omitted the 2-stage transformation, which uses the exponential and modulus functions in sequence. This might seem surprising because the specificity study showed that it had relatively stable performance for symmetric and moderately skewed distributions (Fig. 1, bottom panel); however, the modulus transformation will necessarily corrupt the test in the presence of real outliers. Extra values in one or both tails of a distribution will increase the coefficient of kurtosis, but this is precisely what the modulus transformation attempts to correct. Test simulations (results not shown) confirmed that this was in fact the case.

When the underlying distribution was positively skewed or gaussian (distributions 4–7), the original Horn's outlier test (Fig. 3, filled columns) identified only slightly fewer values outside the Tukey fences than the expected total percentage (leftmost panels of Fig. 3) if the outliers were located in the upper tail or both tails of the distribution (top and middle rows). However, in the case of the upper tail, approximately one third of these values were low false outliers (top row, middle panel), whereas a corresponding percentage of the true outliers located in the upper tail remained unidentified (top row, rightmost panel). The outlier test based on Box–Cox transformation showed a rather poor performance when the outliers were located in the lower tail of the distribution (Fig. 3, bottom row), and this was true for negatively skewed distributions (distributions 8–10) in particular.

We did not observe this kind of asymmetric behavior when we used the exponential transformation in a modified Horn's algorithm (Fig. 3, open columns). It underestimated somewhat the percentage of outliers when they were located in both tails (middle row of Fig. 3). The sensitivity was unacceptably low when the outliers were located in 1 tail only (Fig. 3, top and bottom rows).

CONCLUSIONS

The results of our Monte Carlo simulation experiments concerning outlier detection based on the original Horn's algorithm and 2 modifications of it were rather disappointing. In the specificity study, none of the outlier tests fulfilled the basic requirement of low and predictable probability of false detection. The sensitivity study suggested that the sensitivity tends generally to be too low. The main underlying problem seems to be that the calculation of Tukey fences in Horn's algorithm assumes that the transformed distributions are close to gaussian in shape. Our results indicate that this is most often not the case, as judged from the coefficients of skewness and kurtosis after transformation, not even when outliers were absent (see, for example, the specificity study). The presence of true outliers only increases the problems with the transformations.

We assumed that the following modifications of Horn's algorithm could possibly help to eliminate some of the negative effects of outliers on the transformation: (*a*) truncate the distribution to eliminate possible outliers by temporarily excluding, e.g., 5% of the extreme values at each tail; (*b*) then estimate the transformation parameter on the truncated distribution; and (*c*) finally transform all data, including the outliers, with this parameter and

continue with steps 2–4 of Horn's algorithm (see the *Materials and Methods*). However, test runs using this modification showed that the performance of Horn's algorithm still was not acceptable (results not documented).

Horn's algorithm for outlier detection is based on a promising idea *(8)* to determine outliers using criteria that are calculated from the central part of a hopefully close-to-gaussian distribution. However, our simulation experiments suggest that the normalization of distributions achieved by use of the transformation functions involved in the present study is not good enough to allow Horn's algorithm to work as it is expected to do. Although Horn's algorithm undoubtedly is an improvement compared with older methods for outlier detection, reliable statistical identification of outliers in reference data remains a challenge.

References

1. Solberg HE, ed. International Federation of Clinical Chemistry and International Committee for Standardization in Haematology. Approved recommendation (1987) on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. J Clin Chem Clin Biochem 1987;25: 645–56.
2. Solberg HE. Establishment and use of reference values. In: Burtis CA, Ashwood ER, Bruns DE, eds. Tietz textbook of clinical chemistry and molecular diagnostics, 4th ed. St. Louis: Saunders, 2006:425–48.
3. Barnett V, Lewis T. Outliers in statistical data. Chichester, England: John Wiley, 1994:584pp.
4. Hawkins DM. Identification of outliers. London: Chapman and Hall, 1980:188pp.
5. Dixon WJ. Processing data for outliers. Biometrics 1953;9:74–89.
6. Solberg HE. RefVal: a program implementing the recommendations of the International Federation of Clinical Chemistry on the statistical treatment of reference values. Comput Methods Programs Biomed 1995;48:247–56.
7. Solberg HE. The IFCC recommendation on estimation of reference intervals. The RefVal program. Clin Chem Lab Med 2004;42: 710–4.
8. Horn PS, Feng L, Li Y, Pesce AJ. Effect of outliers and nonhealthy individuals on reference interval estimation. Clin Chem 2001;47: 2137–45.
9. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc 1964;B26:211–52.
10. Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley, 1977:688pp.
11. Manly BFJ. Exponential data transformations. Statistician 1976; 25:37–42.
12. John JA, Draper NR. An alternative family of transformations. Appl Statist 1980;29:190–7.
13. Solberg HE. Statistical treatment of reference values in laboratory medicine: testing the goodness-of-fit of an observed distribution to the Gaussian distribution. Scand J Clin Lab Invest 1986; 46(Suppl 184):125–32.