Clinical Chemistry 47:12
2137–2145 (2001)

Laboratory Management

Effect of Outliers and Nonhealthy Individuals on Reference Interval Estimation

Paul S. Horn,[1*] Lan Feng,[2] Yanmei Li,[3] and Amadeo J. Pesce[4]

Background: **Improvement in reference interval estimation using a new outlier detection technique, even with a physician-determined healthy sample, is examined. The effect of including physician-determined nonhealthy individuals in the sample is evaluated.**

Methods: **Traditional data transformation coupled with robust and exploratory outlier detection methodology were used in conjunction with various reference interval determination techniques. A simulation study was used to examine the effects of outliers on known reference intervals. Physician-defined healthy groups with and without nonhealthy individuals were compared on real data.**

Results: **With 5% outliers in simulated samples, the described outlier detection techniques had narrower reference intervals. Application of the technique to real data provided reference intervals that were, on average, 10% narrower than those obtained when outlier detection was not used. Only 1.6% of the samples were identified as outliers and removed from reference interval determination in both the healthy and combined samples.**

Conclusions: **Even in healthy samples, outliers may exist. Combining traditional and robust statistical techniques provide a good method of identifying outliers in a reference interval setting. Laboratories in general do not have a well-defined healthy group from which to compute reference intervals. The effect of nonhealthy individuals in the computation increases reference interval width by ~10%. However, there is a large deviation among analytes.**

One problem facing the clinical chemist is how to derive reference intervals (RIs)[5] from healthy populations. The effects of inclusion of individuals different because of age, race, exceptional exercise, or diet and inclusion of nonhealthy populations on the healthy estimate have not been examined. This may be considered an outlier problem. The advent of National Health and Nutrition Examination Survey (NHANES) data *(1)* and our Fernald population *(2)* with extensive subject history has made it possible to determine the effect of such individuals (e.g., outliers) on RI estimation. We have proposed the use of robust estimators as a way of reducing the effect of outliers *(3, 4)*. However, we propose that removal of outliers before analysis may yield better estimates of the RI for both robust and nonparametric estimators. The availability of the NHANES and Fernald data, which include physician-determined health status, makes it possible to determine whether our proposed methodology is useful. It also allows us to examine the effect of including nonhealthy individuals on the estimates. The long-term goal is to help clinical chemists decide whether their data approximate those of a healthy population *(5, 6)*. We use the term nonhealthy rather than unhealthy to differentiate between a physician health status score and a known pathology.

The NHANES III, 1988–1994 CD-ROM (for purposes of abbreviation referred to as NHANES III) contains data for 33 994 persons 2 months of age and older who participated in the survey *(1)*. The CD-ROM was obtained from the National Center for Health Statistic Data Dissemination Branch Centers for Disease Control and Prevention (6525 Belcrest Road, Room 1064, Hyattsville, MD 20782-2003). The data are the result of a complex survey design involving stratification and clustering, and thus, weights

[1] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221-0025.
[2] Mount Carmel Health System, Columbus, OH 43222.
[3] HGO Technology, Wheeling, WV 26003.
[4] Department of Pathology and Laboratory Medicine, University of Cincinnati, Cincinnati, OH 45267-0714.
*Author for correspondence. Fax 513-556-3417; e-mail paul.horn@uc.edu.
Received July 19, 2001; accepted September 17, 2001.

[5] Nonstandard abbreviations: RI, reference interval; NHANES, National Health and Nutrition Examination Survey; MEC, mobile examination center; RIW, reference interval width; IQR, interquartile range; AST, aspartate aminotransferase; BUN, blood urea nitrogen; GGT, γ-glutamyltransferase; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; RBC, red blood cell; FMMP, Fernald Medical Monitoring Program; and RMSE, root mean square error.

were assigned to each individual. The weight for an individual indicates the number of people represented by that individual. For most analyses, the weight variable *WTPFHX6* should be used in conjunction with the mobile examination center (MEC) and the home-examined sample and with items collected at both the MEC and the home. However, in this report, we will treat the individuals as coming from a random sample, i.e., individual weights will be ignored. Analyses involving the individual weights will be explored in future work.

Clinical chemistry measurements were made on several analytes, including glucose, sodium, and potassium. Health status was also determined by a physician. The Fernald population is a group of residents who lived near a nuclear feed plant ~15 miles west of Cincinnati, OH. Clinical chemistry analyte measurements were made on the Fernald population similar to those recorded for the NHANES III. The health status of the 9000 residents from Fernald was evaluated in a manner similar to that of the NHANES III study population. The scoring differentiated degrees of health into five (NHANES III) or six (Fernald) categories. In this report, we use the highest (i.e., best, as opposed to largest number) category of health status to define "healthy", whereas all other categories are combined and referred to as "nonhealthy".

We examine the RI width (RIW) estimators because they are insensitive to instrumental, additive bias (although not for proportional bias). Treating the known healthy individuals as a "gold standard", we examine the effect of the addition of nonhealthy individuals on these estimators. Finally, we evaluate the effect of an outlier detection procedure on these estimators.

The NCCLS *(6)* recognizes that outliers in the data are a real possibility. However, the recommendation is that unless it is known that such points are aberrant for known reasons (e.g., a mistake in the analysis), attempts should be made to retain the values instead of deleting them. The NCCLS recommends that the Dixon outlier range statistic be used, especially for RIs determined by the nonparametric procedure. The Dixon test is as follows: let R = the range of the values (maximum–minimum) and let D = the absolute difference between the most extreme (large or small) value and the next most value (large or small). If the ratio D/R exceeds 1/3, then the extreme value in question is deleted. The NCCLS points out that if there are two or three outliers on the same side of the sample, this rule may fail because of masking, i.e., the less extreme outliers mask the aberrance of the most extreme (and vice versa). The recommendation is to test the least extreme outlier as if it were the only outlier. If the D/R test rejects the least extreme outlier, then the more extreme outliers are rejected as well. Unfortunately, the NCCLS has never indicated how these outlier candidates are determined.

Outlier detection methods are mathematically valid based on assumptions about the underlying distribution, which often is assumed to be gaussian. If such a technique were used on skewed data, such as those found in clinical chemistry analytes, it is not clear that the outlier detection would achieve its goal. For example, it could be the case that many good values would be deemed outliers and thus omitted from calculations. This could possibly make the resulting RI unreasonably narrow. On the other hand, if there are a large number of outliers that "mask" their aberrant location, the outlier detection scheme will not flag enough values to be omitted. This could make the resulting RI unreasonably wide. Thus, there is a trade-off between these two situations.

For this report, we recognize that the underlying population of analytical values from healthy individuals is probably skewed (often toward higher values), but in addition, outliers may exist in the observed sample. We also recognize that no matter what the situation, it is difficult to distinguish an extreme "healthy" observation from that of an "nonhealthy" observation. Thus, for this report we propose a two-stage outlier detection scheme that will attempt to balance the two situations cited above. The method will first transform the entire sample to achieve an approximately gaussian random sample. The method of transformation used will be that of Box and Cox *(7)*. Once the data have been transformed, a robust approach to labeling outliers from a gaussian distribution, as described by Tukey *(8)*, will be used. The efficacy of this new methodology, along with the traditional nonparametric approach, will first be examined in a simulation study.

Materials and Methods

One of the most popular methods for transforming data is the family of transformations described by Box and Cox *(7)*. This methodology involves finding, through mathematical techniques, a value for λ (and c, if necessary) for the following transformation of the original data, x:

$$y = \begin{cases} (x^{\lambda} - 1)/\lambda; & \lambda > 0 \\ \ln(x + c); & \lambda = 0 \end{cases} \quad (1)$$

The first stage of our proposed outlier detection scheme is to compute the maximum likelihood estimate of λ or c as described in the Box–Cox model *(7)*. This transformation is based on the entire sample; therefore, it is recognized that outliers will have influence on the transformation. However, if no outliers are present, then the resulting sample should resemble that of a gaussian distribution.

The second stage involves the labeling of extreme values using only the middle 50% of the sample, thus reducing, or even eliminating, the masking effect of possibly many outliers. This method is based on the work of Tukey *(8)* and involves the computation of the lower and upper quartiles (i.e., the 25 and 75 percentiles) of the transformed data. Call these statistics Q_1 and Q_3. The interquartile range (IQR), or Q_3–Q_1, is then computed. Lastly, the lower and upper fences are computed as follows: lower fence = $Q_1 - 1.5$(IQR); upper fence = $Q_3 + 1.5$(IQR). Any transformed data points outside of the fences, i.e., either less than the lower fence or greater than

the upper fence, are considered as outliers, and the original data points are omitted from subsequent RI estimation.

From a theoretical standpoint, for the standard gaussian distribution, the lower and upper quartiles are ± 0.6745, respectively, the IQR is 1.349, and the lower and upper fences are −0.6745 − 1.5(1.349) and 0.6745 + 1.5(1.349), or ± 2.698, respectively. Therefore, theoretically, only values >2.698 SD from the mean will be labeled as outliers. In other words, for a normally distributed population, the 0.7% most extreme values will be considered outliers. It should be noted at this point that if the (transformed) data are representative of a gaussian population, then ~0.7% of valid values will have to be omitted. For this reason, the width of the RI must be adjusted upward to maintain the nominal values. For example, to compute a nominal 90% RI, a 90.634% RI must be computed if the above outlier detection scheme is used (for a 95% RI, 95.67% must be used).

It could be argued that the width should be adjusted upward even more because of the estimation of λ (or c) required for the Box–Cox transformation. The reasoning is that the Box–Cox transformation may be unsuccessful (the transformed data may not appear gaussian). It is also the case that the outlier detection limits are functions of the data themselves *(9)*. These arguments are based on statistical conservatism: where a false negative, in general, is considered more serious than a false positive. However, in the case of RIs, false negatives (having a "sick" individual declared "healthy") are, in general, considered more serious than false positives. It is for this reason that this outlier detection method takes a middle ground by attempting to maintain the nominal width of the desired RI while not allowing statistical conservatism to force the RI to be unusually wide. It should be noted that the IFCC and NCCLS do not recommend any adjustment to the nominal width of the RI after outlier detection *(5, 6)*.

Once the outliers, if any, are identified and removed, the remaining observations are back-transformed to the original scale and used for subsequent computation of the RI. Hereafter, these remaining untransformed data points (i.e., with outliers removed) will be referred to as the adjusted sample. (In practice, these outliers would be studied further to ascertain the reason for unusual behavior.) Nonparametric and robust RIs will be computed based on the original and adjusted samples. (Obviously, if there are no outliers detected then the adjusted sample is the same as the original.)

The nonparametric estimation procedure is the traditional one based on the observed 2.5 and 97.5 percentiles. The RIW is the (positive) difference between these two values. The robust estimator of the upper endpoint of the RI is based on a function that smoothly down-weights the observations based on their distance from the sample median. Only observations greater than or equal to the median are used, and the weighting function is the biweight described by Horn et al. *(3, 10, 11)*. The robust estimator of the lower endpoint is a smoothed version of the nonparametric estimator described by Harrel and Davis *(12)* and used by Horn et al. *(3, 4)*.

To examine the performance of the outlier removal, we ran a simulation consisting of 1000 replicates of samples of size n = 120 for each of six distributions to represent various types of populations. These distributions were the normal, χ^2 with 1 degree of freedom, log-normal, half-normal (i.e., the positive part of the normal distribution), noncentral χ^2 with 1 degree of freedom and noncentrality parameter = 10, and a χ^2 with 4 degrees of freedom. For each sample the RI was computed with and without outlier removal. Subsequently, each sample had the 5% most extreme values altered to mimic outliers. These outliers replaced the 5% largest values of the sample (right side), with the exception of the normal, which had the outliers replace the 5% smallest values (left side). Large-valued outliers were constructed as follows. Consider the ordered sample: $X_1 < X_2 <, \ldots, X_{114}, X_{115}$ is set equal to $(10/7) \times X_{114} - (3/7) \times X_1$, which ensures that $X_{115} - X_{114} = 0.3 \times (X_{115} - X_1)$, which is $<(1/3) \times (X_{115} - X_1)$. This is repeated for $X_j = (10/7) \times X_{j-1} - (3/7) \times X_1$ for $j = 116$–120. Note that this means that the Dixon outlier detection (as recommended by the NCCLS) would not identify these values as outliers because $X_j - X_{j-1} <(1/3) \times (X_j - X_1)$. Similar computations were performed on the smallest six values for the normal case. (Note that the symmetry of the normal population should preclude the need for left side outliers over right side. However, because the robust procedure uses two different estimators for the endpoints, we wanted to test performance for outliers at the low end. The normal population seemed the most appropriate of the distributions examined.)

The NHANES III data set is a survey of individuals in the US. This data set was obtained from NHANES III, Ver. 1.21. The variable *HSSEX* was used to define sex and *HSAGEIR* to define age at interview. The health status was derived from the variable *DMAPEP13A*. Any data with the following missing values were deleted: sex, age, race (as defined by the variable *DMARETHN*), or health status. Clinical chemistry and hematology measurements were made on several analytes and cellular components. These laboratory data were obtained from the NHANES III Laboratory Data File. In this study we examined the following 33 analytes: albumin, alanine aminotransferase, alkaline phosphatase, aspartate aminotransferase (AST), blood urea nitrogen (BUN), calcium, chloride, creatinine, creatinine in urine, γ-glutamyltransferase (GGT), glucose, granulocytes, hematocrit, hemoglobin, potassium, lactate dehydrogenase, lymphocytes, mean corpuscular hemoglobin (MCH), MCH concentrate, mean corpuscular volume (MCV), monocytes, mean platelet volume, osmolality, platelets, phosphorous, red blood cells (RBCs), RBC distribution width, sodium, total bilirubin, total CO_2, total protein, uric acid, and white blood cells.

The Fernald sample was obtained from a population

living near a uranium ore processing plant located in Fernald, OH *(3, 4)*. The population is representative of the greater Cincinnati metropolitan area. As part of a settlement, the residents were offered an extensive medical monitoring program extending for 17 years. This will be referred to as the Fernald Medical Monitoring Program (FMMP) group. The residents were given complete examinations, including physical examination, personal and family medical histories, pulmonary function tests, chest x-rays, and cardiac monitoring as well as blood and urine tests, including the chemistry and hematology tests reported here. All individuals were seen within a 3-year period from December 1990 to November 1993. The FMMP group represents a unique data set. Disease history, family history, psychological history, physical examination, chest x-ray, and laboratory values were derived for each of the 8517 residents. The same 33 analytes mentioned above were examined for the FMMP group.

In both the NHANES III and FMMP groups, physicians using similar rating scales scored the participants' health status. For both groups, we designated individuals as healthy if they achieved the highest health status rating. The remaining individuals were designated nonhealthy. We chose to divide the adult population into 12 groups comprising two genders and six age categories: 20–29, 30–39, 40–49, 50–59, 60–69, and 70–79 years. The numbers of individuals in each age category by gender by health status are given in Table 1.

We examined the RIW because it is insensitive to additive bias and ratios of these widths allow us to compare across analytes. (Because some enzymes may have proportional bias if assayed at different temperatures, for example, measures based on RIWs alone may be sensitive to such bias. However, measures based on ratios of RIWs will not be sensitive to such bias.) Robust and nonparametric estimators of the RIW were computed using a SAS^{TM} program described previously *(4)*. Because we had access to the participants' health status, we were able to determine the effect of the inclusion of a nonhealthy group on the RIW. The RIW is defined as the difference between the $(1 - \alpha/2) \times 100\%$ and the $(\alpha/2) \times 100\%$ of the values of the group under consideration. In this report, we examined 95% RIWs, i.e., $\alpha = 0.05$.

The nonparametric estimation procedure is the traditional one based on the observed upper and lower endpoints, or percentiles. The RIW is the difference between these two values. The robust estimator of the upper endpoint of the RI is based on a function that smoothly down-weights observations based on their distance from the sample median. Only observations greater than or equal to the median are used, and the weighting function is the biweight described by Horn *(11)*. The robust estimator of the lower endpoint is a smoothed version of the nonparametric estimator described by Harrel and Davis *(12)* and used by Horn et al. *(3, 4)*.

The RIW was calculated using the traditional nonparametric 95% RIs and our previously described robust estimator *(4)*, with both methods using the proposed outlier detection. We chose the ratio of RIWs as the measure of the effect of inclusion of nonhealthy individuals (RIW based on the total group divided by RIW based on the healthiest in the group). To average across variables, we used the logarithm of this ratio because the means of these values are not biased. For example, if $x/y = 1.25$, then $y/x = 0.8$; therefore, the average of the two is not equal to 1. However, if $\log(x/y) = 1.25$, then $\log(y/x) = -1.25$, and the average is 0, which has an

Table 1. Frequencies of health status by age category and sex for NHANES III and FMMP.

	Age categories, years						
	20–29	**30–39**	**40–49**	**50–59**	**60–69**	**70–79**	**Total**
NHANES III							
Males, n							
Healthy	1046	761	490	235	211	110	2853
Nonhealthy	551	663	701	586	906	687	4094
Total	1597	1424	1191	821	1117	797	6947
Females, n							
Healthy	1175	923	533	244	184	89	3148
Nonhealthy	635	880	769	714	930	808	4736
Total	1810	1803	1302	958	1114	897	7884
FMMP							
Males, n							
Healthy	445	691	646	390	287	78	2537
Nonhealthy	60	159	221	249	239	121	1049
Total	505	850	867	639	526	199	3586
Females, n							
Healthy	592	864	769	404	231	63	2923
Nonhealthy	93	276	324	341	309	120	1463
Total	685	1140	1093	745	540	183	4386

anti-logarithm equal to 1. Therefore, the log-ratio analysis included 6 age categories, 2 sexes, and 33 analytes for a total of 396 comparisons.

Results

The results of the simulation are summarized in Tables 2 and 3 and in Tables S1 and S2 [Tables S1–S4 can be found in a data supplement attached to the electronic version of this article, available at *Clinical Chemistry Online* (http://www.clinchem.org/content/vol47/issue12)]. The means and SD for each of the endpoints of the RI are provided in Table 2; the population or "ideal" values are given in parentheses below each of the six distributions. The effect of 5% outliers on the estimated endpoints is dramatic. For example, for the χ^2 1 distribution, the average values of the upper endpoint were 200–300% of the ideal value of 5.024. With outlier detection in place, the average values of the upper endpoint were within 4% of the ideal value. Similar results were obtained for the normal, half-normal, noncentral χ^2 1, and the χ^2 4 distributions. The log-normal distribution endpoints were also inflated, by ~250%. However, outlier detection was not as effective as with the other distributions: the improvement was 5–13%.

The percentages of outliers removed from each of the two sides of the sample are summarized in Table 3. With no outliers and a normal distribution, we would expect 0.35% outliers removed from each side of the sample. In general, this expectation was met, although the percentage of outliers found for the χ^2 1, half-normal, and χ^2 4 were less than expected. With 5% outliers, the method found, on average, 5% for the normal distribution. For the χ^2 1, half-normal, noncentral χ^2, and χ^2 4 distributions, the percentages of outliers found were, on average, 2.5–4.3%. For the log-normal distribution, only 0.7% were found at the high end, providing an explanation for the weaker performance observed in Table 2.

Lastly, to compare the accuracy of the RIW estimators with and without outlier detection, the root mean square error (RMSE) was examined. For each sample, the RIW was computed, the ideal RIW was subtracted from it, and the difference was squared. (The ideal RIW for each distribution is the difference between its 97.5 and 2.5 percentiles. These population percentiles are given in Table 2.) These squared differences were averaged over the 1000 replications, and the square root was taken, yielding the RMSE. For example, for the normal distribution:

Table 2. Means and SD of simulated RI endpoints.

			No outlier detection				Outlier detection			
			No outliers		5% outliers		No outliers		5% outliers	
Distribution (Population percentiles)	Percentile	Measure	Nonpar[a]	Robust	Nonpar	Robust	Nonpar	Robust	Nonpar	Robust
Normal	0.025	Mean	−1.955	−2.008	−12.204	−14.483	−1.924	−1.946	−1.471	−1.484
		SD	0.230	0.219	1.233	1.479	0.230	0.211	0.168	0.164
(−1.96; 1.96)	0.975	Mean	1.948	1.982	1.948	1.982	1.932	1.983	1.920	1.982
		SD	0.229	0.182	0.229	0.182	0.232	0.195	0.229	0.196
χ^2 1	0.025	Mean	0.002	0.002	0.002	0.002	0.001	0.002	0.001	0.002
		SD	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001
(0.001; 5.024)	0.975	Mean	5.070	4.609	13.154	11.570	4.896	4.295	5.249	4.915
		SD	0.947	0.715	2.271	1.902	0.972	0.704	2.454	1.765
Log-normal	0.025	Mean	0.147	0.143	0.147	0.143	0.149	0.148	0.172	0.171
		SD	0.033	0.029	0.033	0.029	0.038	0.034	0.037	0.034
(0.141; 7.099)	0.975	Mean	7.261	6.900	17.508	15.359	7.243	6.426	16.343	13.216
		SD	1.820	1.829	3.408	2.794	1.992	1.667	6.647	4.893
Half-normal	0.025	Mean	0.035	0.035	0.035	0.035	0.030	0.031	0.030	0.030
		SD	0.018	0.016	0.018	0.016	0.017	0.015	0.017	0.015
(0.031; 2.241)	0.975	Mean	2.228	2.207	6.761	6.141	2.228	2.200	1.844	1.977
		SD	0.210	0.165	0.588	0.508	0.221	0.182	0.154	0.170
Noncentral χ^2 1 (10)[b]	0.025	Mean	1.541	1.480	1.541	1.480	1.515	1.486	2.231	2.205
		SD	0.572	0.503	0.572	0.503	0.616	0.543	0.746	0.712
(1.446; 26.237)	0.975	Mean	26.298	25.773	79.078	72.052	26.187	25.561	25.662	27.291
		SD	2.406	1.812	6.449	5.536	2.534	1.953	8.520	6.417
χ^2 4	0.025	Mean	0.505	0.491	0.505	0.491	0.477	0.468	0.656	0.650
		SD	0.145	0.130	0.145	0.130	0.153	0.136	0.192	0.184
(0.484; 11.143)	0.975	Mean	11.150	10.707	32.175	29.068	11.108	10.558	14.648	14.051
		SD	1.272	0.944	3.291	2.802	1.410	1.027	7.484	5.302

[a] Nonpar, nonparametric.
[b] 10 represents the noncentrality parameter.

Table 3. Average percentage of outliers removed from left and right tails of simulation.

	No outliers		5% outliers	
Distribution[a]	Left	Right	Left	Right
Normal	0.5	0.4	5.0	0.5
χ^2 1	0.0	0.6	0.0	2.7
Log-normal	0.5	0.4	1.5	0.7
Half-normal	0.0	0.4	0.0	4.3
Noncentral χ^2 1 (10)[b]	0.3	0.4	2.4	3.5
χ^2 4	0.1	0.4	2.1	2.5

[a] All distributions have outliers on the right except for the normal.
[b] 10 is the noncentrality parameter.

$$\text{RMSE} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(RIW_i - 3.92)^2}$$

where RIW_i is the RIW for the ith sample. Thus, the ratio described above is equal to: [RMSE (no outlier detection) ÷ RMSE (with outlier detection)] × 100%. These values were derived for each of the six distributions, and the results are presented in Table S1 of the data supplement. With no outliers, as expected, the RIWs without outlier detection were in general 5–10% more accurate. With 5% outliers, the RIWs with outlier detection were 2.5–24 times more accurate for all of the distributions except for log-normal, where the nonparametric was 4% less accurate and the robust was only 12% more accurate.

A similar measure of accuracy was computed for the coverage attained by the RIs. Simply stated, the methods using outlier detection had larger RMSEs, by ~30%, than those without outlier detection. This is not surprising because undetected outliers will often yield a very large estimate (theoretically unbounded) of the upper endpoint of the 95% RI, yielding a coverage of almost 97.5%. Such RIs will have an RMSE equal to ~0.025, whereas those with outliers removed will often have smaller upper endpoints, but their RMSEs are not bounded by 0.025. (If outliers exist on both sides, a procedure that does not remove outliers has RMSE bounded by 0.05, i.e., it attains 100% coverage, and no observation falls outside the resulting RI.) These results appear in Table S2 of the data supplement.

We applied the outlier detection scheme with robust and nonparametric RI estimators to the NHANES and Fernald data sets. We determined the percentage of observations removed as a result of outlier detection for both the NHANES III and Fernald groups, as well as their healthy subgroups. This was done separately for each of the 33 analytes, ignoring the sex and age categories. Outliers were found in every case. Table S3 of the data supplement provides for each analyte and study group (FMMP or NHANES and total or healthy groups) the number of (nonmissing) observations and the percentage of outliers found among these values. In Table S3, in general, larger percentages of observations were removed from the total groups compared with their respective healthy subgroups. The analyte with the largest percentage of detected outliers was glucose. This result was expected because diabetic and prediabetic individuals are present in the general population but may be deemed healthy by a physician.

The RIW analysis included 6 age groups (20–29, 30–39, 40–49, 50–59, 60–69, and 70–79 years), 2 sex groups, and 33 analytes, for a total of 396 comparisons. The effect of outlier detection on the RIWs of the two groups, NHANES III and FMMP, as well as their respective healthy subgroups, are given in Table 4. The RIW was calculated using the traditional nonparametric 95% RIs and our previously described robust estimator. For the reasons described above, we chose the logarithm of the ratio of RIWs as the measure of the effect of outlier removal.

The log-ratios of the RIW with no outlier detection to the RIW with outlier detection (× 100%) were calculated for each set of 396 comparisons. The percentages of observations removed from the data sets were similar, ~1.6%. Note that this is more than double the theoretical value of 0.7% for a normal distribution. As seen in Table 4, the log-ratios were >0 for all comparisons, i.e., the anti-logarithm of the ratio was >1. The outlier detection and removal method was effective in reducing the RIW in both the healthy groups and those with disease. For these groups, the average changes for the nonparametric and robust methods were ~10% and 15%, respectively.

To determine whether the ratio was attributable to only a few analytes, the effect of outlier detection on the RIW of each analyte was reviewed. A summary of these analyses is presented in Table S4 of the data supplement. The results were similar to those in Table 4, except that they are given for each analyte. It is clear that the robust

Table 4. Means and SE of log-ratios (× 100%) of no outlier to outlier removal and percentage of observations removed.

		95% RI[a]	
		Nonparametric	Robust
NHANES III			
Total	Mean	10.5	16.9
(1.7% removal)	SE	0.9	1.1
Healthy	Mean	9.8	14.6
(1.5% removal)	SE	1.0	1.1
FMMP			
Total	Mean	9.4	12.9
(1.7% removal)	SE	0.9	0.9
Healthy	Mean	9.8	14.6
(1.6% removal)	SE	1.0	1.1

[a] For each study (FMMP and NHANES) and group (Total and Healthy), the 396 RIs (nonparametric and robust separately) were computed with and without outlier removal. For each of the 396 cases, the RIW without removal was divided by the RIW with removal. The natural logarithm was taken for these ratios and multiplied by 100%. Entries are the means and SE computed across each of these 396 cases.

approach was more affected by outliers than the nonparametric approach. This is because the robust approach was designed to capture information from the tails of the underlying distribution, which may be heavy-tailed *(10)*. Thus, when such information is contaminated, the robust approach will greatly benefit from outlier detection. Lastly, it is also noteworthy that the RIWs for glucose were the most affected by outlier detection and the subsequent removal of observations. This applies to the total groups and their respective healthy subgroups.

The means and SE of the log-ratios of the total to healthy RIWs were calculated. For both the NHANES III and FMMP groups, the RIW was larger for the total group compared with that of the healthy, as was expected. In other words, the inclusion of nonhealthy individuals increased the width of the RI. For the nonparametric estimator, the log-ratio of the RIW was larger for the NHANES III data, on average, by ~10% (SE = 0.7%). For the FMMP data, the log-ratio of the RIW was larger, on average, by ~4.5% (SE = 0.5%). For the robust estimator, the values were 9% (SE = 0.7%) for the NHANES III and 3.6% (SE = 0.4%) for the FMMP group.

The effect of including nonhealthy individuals was also determined by examining the log-ratio of the RIW of total to healthy for each analyte. These results are given in Table 5. We found that the nonparametric RIW was more affected by the presence of the nonhealthy than was the robust RIW. To determine which analytes were most affected, we considered a log-ratio ≥10% indicative of a large effect of a nonhealthy group on the RIW. Therefore, based on this reasonable, although ad hoc, cutoff of 10%, the analytes most affected for the FMMP group were GGT, glucose, and monocytes. For the NHANES III group, the most affected analytes were albumin, AST, BUN, creatinine, GGT, glucose, hematocrit, hemoglobin, MCH, MCV, monocytes, RBC distribution width, and uric acid.

Table 5. Mean log-ratios (× 100%) of total to healthy by analyte.

	FMMP		NHANES	
Analytes	**Nonparametric**	**Robust**	**Nonparametric**	**Robust**
Albumin	−0.6	0.5	10.8	9.1
ALT[a]	5.5	5.2	3.5	7.1
AP	4.6	3.5	8.5	10.1
AST	5.2	4.9	16.6	14.5
BUN	−0.2	1.4	14.7	13.3
Calcium	−1.8	−2.1	1.1	1.3
Chloride	−2.9	−3.2	8.4	7.6
CR	4.3	5.7	16.6	12.8
CR-URIN	2.7	2.0	3.7	2.9
GGT	13.2	11.7	38.2	38.8
Glucose	14.7	10.3	21.0	17.2
GRAN	8.6	7.6	9.6	8.0
HCT	4.0	4.2	10.4	10.3
Hb	4.4	4.3	10.1	9.6
K^+	−0.4	2.1	5.7	6.3
LDH	6.3	5.5	8.9	7.4
LYMPH	−0.1	2.5	7.5	6.7
MCH	5.6	4.4	11.3	9.5
MCH CON	2.5	0.6	0.4	−0.8
MCV	3.1	1.5	10.4	9.0
MONO	10.8	4.4	10.1	8.6
MPV	2.0	3.6	4.7	3.7
OSMO	5.0	4.4	7.4	8.1
Platelets	2.2	3.1	8.7	7.2
PO_4	5.1	3.7	5.6	4.7
RBC	1.1	2.0	7.8	6.8
RBCDW	4.8	5.4	16.0	14.5
Sodium	7.9	3.7	6.6	5.6
TBILIRUB	1.6	2.1	8.0	5.2
TCO_2	3.7	0.8	7.4	5.8
TPROTEIN	5.9	3.6	5.3	4.7
URIC	5.4	4.0	14.0	13.6
WBC	7.3	7.1	6.5	6.5

[a] ALP, alanine aminotransferase; AP, alkaline phosphatase; CR, creatinine; CR-URIN, urinary creatinine; GRAN, granulocytes; HCT, hematocrit; Hb, hemoglobin; LDH, lactate dehydrogenase; LYMPH, lymphocytes; MCH CO, MCH concentrate; MONO, monocytes; MPV, mean platelet volume; OSMO, osmolality; RBCDW, RBC distribution width; TBILIRUB, total bilirubin; TCO_2, total CO_2; TPROTEIN, total protein; URIC, uric acid; WBC, white blood cells.

Discussion

The IFCC *(5)* and NCCLS *(6)* protocols describe the use of healthy individuals to determine a RI. The problem is that health status is difficult to validate. This is similar to the problem of trying to prove the null hypothesis in statistics. We proposed in this study the use of outlier detection to help validate healthy individuals for inclusion in the RI calculation. The proposed outlier detection scheme consisted of two parts: (*a*) all of the data are transformed to fit a normal distribution, using the Box–Cox family of transformations; and (*b*) a robust outlier detection technique is used based on the transformed data. A standard normal-theory outlier detection is not used because the transformation may not be entirely successful. The use of a robust outlier detection scheme after a transformation to normality provides protection against the possible perturbation of the parameter estimates for the transformation. Thus, the outliers that may cause erroneous transformation will nevertheless be identified because of the outlier resistance of the robust detection scheme. The reverse of these two steps was not chosen because the robust outlier detection scheme would identify too many outliers if the underlying population were homogeneous but skewed.

The simulation study demonstrated that the use of outlier removal when no outliers exist would lead to minimal loss of accuracy of the endpoint estimators of the RI. However, if there are as a few as 5% outliers (not detectable by the Dixon range test), then the gains in accuracy may be enormous, depending on the underlying distribution of the analyte. The above study was based on the minimal sample size of 120 as recommended by the IFCC. A similar simulation was performed for sample size of 60, and similar results were obtained.

With a random sample from a normal distribution, the outlier detection scheme will, theoretically, identify ~0.7% of the observations as outliers. We found that in two healthy groups, the percentage of identified outliers, by analyte, ranged from ~0% to 4.2% (Table S3 in the data supplement). The weighted average (because of different sample sizes) for all 33 analytes was ~1.6%, or a little more than twice that of an ideal, normal group (Table 4). Further verification of the outlier detection scheme comes from the glucose data. For this analyte, 3.8–6.2% of the values were identified as outliers, consistent with an estimate of a diabetes prevalence of 6% in the US population *(13)*. It is also true that both of the groups would contain individuals in a prediabetic state.

In general, the effect of the removal of 1.6% of the observations decreased the RIW by ~10–15%. This effect varied among the 33 different analytes. For glucose, the effect was striking. When we used outlier detection, the reduction in the 95% RIW for both groups was ~70–90%. (For the 90% RIW, results were not reported; however, the effect was smaller but substantial, ranging from 40% to 60%.)

A limitation of our outlier detection scheme is that it will not work for sample groups that are significantly different from the intended reference population. Specifically, the resulting RI will not work, in most cases, for pediatric data obtained from inpatients and applied to the population of healthy children. However, every method of RI estimation will suffer in this case. Another limitation of our approach is that by removing outliers we may increase the false-positive rate from the nominal value of 5% (for 95% RIs). However, this is balanced by the fact that for most analytes the approach will improve the ability to detect disease as well as remove inappropriate values. For example, individuals of different races may be combined to estimate a single RI. If all of the data are used, this may produce a RI inappropriate for the relevant population.

The problem of including unhealthy individuals in an estimate of the RI is a real one. Few laboratories can obtain well-defined healthy individuals in the numbers required to meet IFCC and NCCLS standards *(5, 6)*. There are no publications as to how much of an effect this may have on the RI estimate. In this work, we showed that even in the most carefully defined healthy population, outlier detection is required to obtain a better estimate of the RI. To perform our analysis we considered only comparisons after outlier removal. Having the NHANES III and FMMP groups, which have well-defined healthy and nonhealthy subgroups, allowed us to examine the effect of including nonhealthy individuals in our RI estimates. We have shown that, as expected for almost every analyte, the RIW was wider for the total group than for the healthiest group. The difference in RIW is narrowed if robust estimators rather than nonparametric estimators are used.

We observed that the FMMP total RIW was only ~4% wider than that of the healthy subgroup, compared with 9% for the NHANES III group. A well-defined healthy group is required to determine RIs *(5, 6)*. We have quantified the effects of using a sample that includes nonhealthy individuals in the computation of RIs. These effects can be substantial and vary by analyte (see Table 5). For example, GGT and glucose are the analytes most sensitive to the inclusion of nonhealthy individuals in the calculation of their respective RIs.

Several compilations of RIs have been reported. Specifically, there is one for the geriatric population *(14)*. Reference values for the same analytes vary widely among the different reporting laboratories. Some of these differences may be the result of method bias. However, in cases where the same methods are used, we believe that our estimates can serve as guidelines for which analytes may be most affected by inclusion of the nonhealthy.

RIs by definition do not include the entire range of healthy values. Thus the problem becomes how should extreme values be omitted before the calculation. This can be done either by a priori knowledge of known strata or by the use of outlier detection. For example, including individuals who exercise extremely hard, such as marathon runners, is not appropriate for describing the usual healthy population. If histories were taken, as they should be, these individuals would be put in their own stratum and not be included in the RI determination. When histories are not known, the outlier detection scheme described here should flag such individuals, and if possible, their medical histories may be evaluated for unusual activities, such as extreme exercise. The decision of which method to use to omit values depends on the knowledge (stratification), or lack thereof (outlier detection), of the individuals in the sample.

In conclusion, laboratories do not usually have specimens from a well-defined healthy group. Instead, specimens are obtained from a variety of sources, depending on availability, e.g., laboratory workers. Compounding this problem is the fact that laboratories do not have physician-determined health status to serve as a prescreener. [See Sasse *(15)* for a general discussion on the RI problem.] Therefore, our proposed outlier detection scheme will be useful in practice because it reduces the effect of unhealthy individuals as well as those who may be inappropriate for the test group. We provide estimates of the effects of including a nonhealthy group on various analytes. This should be useful in the assessment of RIs.

References

1. Centers for Disease Control and Prevention. Third National Health and Nutrition Examination Survey (NHANES III), 1988–1994 [CD-ROM]. Hyattsville, MD: National Center for Health Statistic Data Dissemination Branch Centers for Disease Control and Prevention.
2. Copeland BE, Pesce AJ. The medical heritage concept: a model for assuring comparable laboratory results in long-term longitudinal studies. Ann Clin Lab Sci 1992;22:110–24.

3. Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. Clin Chem 1998;44:622–31.

4. Horn PS, Pesce AJ, Copeland BE. Reference interval computation using robust, parametric, and nonparametric analyses. Clin Chem 1999;45:2284–5.

5. International Federation of Clinical Chemistry. 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.

6. Sasse EA, Aziz KJ, Harris EK, Krishnamurthy S, Lee HT Jr, Rutland A, Seamonds B. How to define and determine reference intervals in the clinical laboratory; approved guideline. NCCLS document C28-A, Vol. 15, No. 4. Wayne, PA: NCCLS, 1995.

7. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc 1964;B26:211–52.

8. Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley, 1977:506pp.

9. Hoaglin DC, Iglewicz B, Tukey JW. Performance of some resistant rules for outlier labeling. J Am Stat Assoc 1986;81:991–9.

10. Horn PS. A biweight prediction interval for random samples. J Am Stat Assoc 1988;83:249–56.

11. Horn PS. Robust quantile estimators for skewed populations. Biometrika 1990;77:631–6.

12. Harrel FE, Davis CE. A new distribution-free quantile estimator. Biometrika 1982;69:635–70.

13. NIH. National Institute of Diabetes and Digestive and Kidney Diseases website. http://www.niddk.nih.gov/index.htm (Accessed July 2, 2001).

14. Faulkner WR, Meites S. Geriatric clinical chemistry: reference values. Washington: AACC Press, 1994:386–91.

15. Sasse EA. Reference intervals. In: Kaplan LA, Pesce AJ, eds. Clinical chemistry: theory analysis and correlation. St. Louis, MO: CV Mosby, 1996:365–81.