Teppei Yamashita, Kiyoshi Ichihara* and Ayaho Miyamoto

A novel weighted cumulative delta-check method for highly sensitive detection of specimen mix-up in the clinical laboratory

Abstract

Background: We sought to detect specimen mix-up by developing a new cumulative delta-check method applicable to a mixture of test items with heterogeneous units and distribution patterns.

Methods: The distributions of all test results were successfully made Gaussian using power transformation. Values were then standardized into z-score (zx) based on reference interval (RI) so that limits of RI take zx=±1.96. To find a weight for summing absolute value of delta between current and previous zx (Dz), we evaluated the distribution of Dz. Its central portion was always regarded as Gaussian despite the presence of symmetrical long tails. Thus, an adjusted SD (aSD) representing the center was estimated with an iterative method. By setting $1/aSD^2$ as a weight factor, we computed a weighted mean of Dz as an index for specimen mix-up (wCDI).

Results: The performance of wCDI was evaluated, using a model laboratory database consisting of 32 basic test items, by a simulation study generating artificial cases of mix-up. When wCDI was computed from three commonly ordered test sets consisting of 6–9 items each, its diagnostic efficiency in detecting the artificial cases was 0.937–0.967 expressed as area under ROC curves (AUC). When the performance of wCDI was evaluated simply by the number of test items (p) included in the computation, AUC gradually increased from 0.944 (p=5) to 0.976 (p=8). However, when $p \geq 10$, AUC stayed at approximately 0.98.

Conclusions: wCDI was proven to be highly effective in uncovering cases of specimen mix-up. The diagnostic efficiency of wCDI depends only on the number of test items included in the computation.

Keywords: data-mining; delta-check method; laboratory information system; modified Box-Cox power transformation; quality management; within-individual differences.

***Corresponding author: Kiyoshi Ichihara**, MD, PhD, Faculty of Health Sciences, Department of Clinical Laboratory Sciences, Yamaguchi University Graduate School of Medicine, Minami-Kogushi 1-1-1, Ube, 755-8505 Japan, Phone: +81 836 222884, Fax: +81 836 355213, E-mail: ichihara@yamaguchi-u.ac.jp

Teppei Yamashita: Graduate School of Health Care Science, Jikei Institute, Osaka, Japan

Teppei Yamashita and Ayaho Miyamoto: Department of Environmental Science and Engineering, Yamaguchi University, Ube, Japan

Introduction

Laboratory automation has advanced greatly in parallel with ever increasing number of orders to clinical laboratories. However, processing of specimens remains dependent on manual work, and specimen mix-ups can occur through mislabeling and patient misidentification [1, 2]. In fact, the incidence is reported to be not negligible [3, 4], although incidents are usually detected by the clinician. However, such incidents often lead to mistrust towards the clinical laboratory. Therefore, the laboratory must make an all-out effort to detect such errors before reporting test results. A simple strategy is to automatically retest the specimen when any test result exceeds a certain threshold or when results of associated test items are discordant. Due to the high prevalence of extreme values, such protective measures increase the cost of running a laboratory.

Naturally, the only plausible measure of identifying specimen mix-up is to evaluate consistency of the current test results with the previous results. As a basic function of the laboratory information system (LIS), automatic comparison with previous results is made, and a large difference (delta) from the previous one is marked to arouse suspicion. However, interpretation for a set of deltas is usually not straight-forward, requiring knowledge on inherent variability of each test item.

Several schemes have been reported to automatically judge possible specimen mix-up. They are based on either summation of deltas of simultaneously measured test items [5] or discriminant function analysis of a set of

deltas for selected [6] or all test items [7]. These methods, however, are not expected to work properly because the distribution patterns of test results differ greatly from one test item to another [8]. A delta from non-Gaussian skewed distribution tends to exert more influence in the analysis. Furthermore, biological variability of test results differs greatly among test items. For example, glucose or triglyceride shows large fluctuations even among healthy individuals depending on the sampling conditions. Therefore, analysis of deltas should be made in consideration of heterogeneous distribution patterns and differences in biological variability.

We have developed a new delta-check method that has overcome both problems by the normalization of the distributions through power transformation and by a weighted summation of deltas based on biological variability. It also features exclusion of influential data points in deriving biological variability using an iterative procedure [9] and uniform expression of all test results by z-score. This method was designated as the weighted cumulative delta-check (wCDC) method. In this report, we describe the theoretical formulation and demonstrate the performance of the wCDC method with a simulation generating artificially mixed-up cases in a model LIS database.

Materials and methods

Theoretical formulation of the wCDC method

Normalization of distribution patterns

We have reported previously that test results from healthy individuals did not follow Gaussian distribution in most analytics, but their distributions can be transformed into Gaussian using the following modified Box-Cox power transformation formula [8, 10].

$$X^T = \frac{(x-a)^p - 1}{p} \quad (p \neq 0)$$

$$X^T = \log(x-a) \quad (p=0)$$

where x and X^T represent test values before and after transformation, and p and a designate power and the origin of transformation, respectively.

Target data in LIS, however, contain a large number of extreme values that apparently affect estimation of p and a. Therefore, before power transformation, we truncated 1% of the data on each tail of the distribution. The maximum likelihood estimation method was used for fitting p and a. As the two estimators are dependent on each other, we adopted an algorithm to estimate just p by this method and then set a at a location corresponding to mean (M) − 4×SD of the transformed data, that is, $a=p\times(M-4SD+1)^{(1/p)}+a_0$, where a_0 represents the previous a. The initial value for a was set at $x_{min}-(Me-x_{min})/10$ where Me represents median and x_{min} represents the smallest observed value. After adjusting a, p was again estimated iteratively until both parameters stabilized. In consideration of gender-dependent difference in distributions of test results, the transformation was done separately for male and female in all the test items.

Uniform expression of test results

To make results of any test item comparable and unaffected by measurement units, all the transformed test results were standardized to a uniform scale on the basis of reference interval (RI) as explained below.

First, the lower and upper limits of the RI (LL, UL), were transformed to LL^T and UL^T by the power transformation:

$$LL^T = \frac{(LL-a)^p - 1}{p}$$

$$UL^T = \frac{(UL-a)^p - 1}{p}$$

Assuming the RI was determined parametrically after power transformation with the same p and a, mean (M^T) and SD (SD^T) of RI under the transformed scale were computed as follows:

$$M^T = \frac{UL^T + LL^T}{2}$$

$$SD^T = \frac{UL^T - LL^T}{3.92}$$

Using M^T and SD^T, transformed test result X^T was converted to z_x (z-score) with the following formula.

$$z_x = \frac{X^T - M^T}{SD^T}$$

This conversion to the uniform scale was done separately for male and female using the gender-specific RIs.

These flows of data processing are illustrated in Figure 1.

Derivation of SD representing within-individual differences

As the next step, to derive within-individual difference in two successive measurements, we consecutively scanned an individual result, from current to the past, and retrieved an immediate past result of the same patient one or more days apart, if any. The difference of the two (D_z) was computed after power transformation and standardization as $z_{curr}-z_{prev}$, and the distribution of D_z for all records was examined item by item. We tried to use the SD of D_z as the index of within-individual difference. However, distributions of D_z were always symmetrical but had long tails, and extreme values in the tails had strong influence in computing SD. Therefore, we adopted the iterative truncation and correction (ITC) method [9] to obtain an unbiased mean and SD unaffected by extreme values in the periphery of the distribution. The ITC method was originally developed for computing means in a setting of external quality assurance surveys, in which we often observe a cluster of extreme values in the periph-

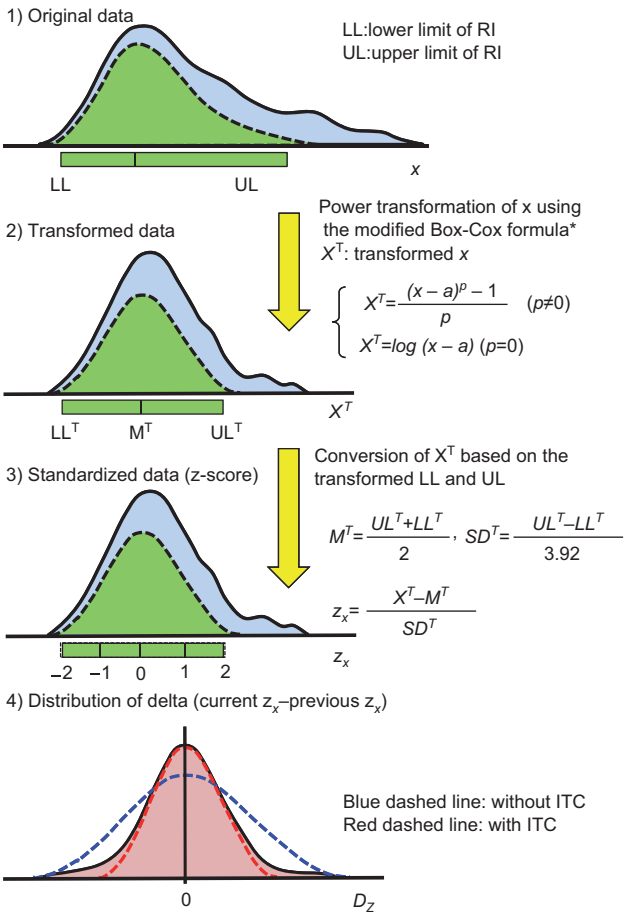

Figure 1 Schematic flow of data processing required to implement wCDC method.
Light blue frequency distribution curve area represents all results for given test item retrieved from the LIS. Green-colored band below the curve represents reference interval (RI) for the test item. Inner green curve area drawn just above RI bar illustrates imaginary distribution of reference values used for computing the RI.

ery of peer group distributions. ITC involves an iterative process of truncation of large blocks of data on both tails of the distribution (outside $M\pm k\times SD$) followed by correction of M and SD according to the truncation coefficient (k). This adjustment is valid only when we can assume Gaussian distribution in the central portion (within $M\pm k\times SD$). Using this principle, we derived adjusted SD (aSD) for the distribution of D_z.

Computation of weighted cumulative delta-check index

We assumed that the larger the aSD of a given test item, the less effective the item to be used for distinguishing specimen mix-up. Therefore, we used the inverse of aSD^2 as a weight, w, in the summation of D_z.

$$w=\frac{1}{aSD^2}$$

Thus, a new index to indicate possible specimen mix-up, named weighted cumulative delta index (wCDI), was derived by the following formula,

$$wCDI=\frac{\sum_{i=1}^{k} w_i\,|D_{zi}|}{\sum_{i=1}^{k} w_i}$$

where we assume that there are k-test items simultaneously measured in current and previous records. An absolute difference of the i-th test item (i=1, 2, ..., k) between the two measurements, $|D_{zi}|$, is multiplied by the weight, w_i, and summed for k items, and then divided by the sum of the weights.

Procedures for validation

Data source

A model database retrieved in 1998 from a large clinical laboratory and made totally anonymous for use in the practicum of a laboratory informatics course was used for the validation study. The database was composed of 171,547 records (inpatient 79,307; outpatient 92,240) consisting of 22,677 unique IDs representing a period of 1 year. The test items used for the evaluation were the 32 most commonly measured items including white blood cell (WBC), red blood cell (RBC), hemoglobin (Hgb), hematocrit (Hct), platelet (PLT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), total protein (TP), albumin (ALB), sodium (Na), potassium (K), chloride (Cl), calcium (Ca), inorganic phosphate (IP), blood urea nitrogen (BUN), creatinine (CRE), uric acid (UA), total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), triglyceride (TG), glucose (GLU), aspartate aminotransferase (AST), alanine aminotransferase (ALT), lactate dehydrogenase (LDH), alkaline phosphatase (ALP), γ-glutamyltransferase (GGT), amylase (AMY), choline esterase (CHE), activated partial thromboplastin time (APTT), and prothrombin time (PT).

The average number of test items measured per order was 17.4. The minimum number of simultaneously measured items valid for computing wCDI was set to 5. Thus, a total of 137,134 paired records were used for the validation study.

Tests for normality of distribution

Goodness-of-fit to Gaussian distribution was done by the following two methods: 1) Skewness and kurtosis [11]. Skewness (Sk) represents a degree of asymmetry in distribution: Gaussian distribution gives Sk=0.0, and a distribution skewed toward lower and upper tails gives Sk<0.0 and Sk>0.0, respectively. Kurtosis (Kt) represents the peakedness of distribution: Gaussian distribution gives Kt=0.0, and a steeper distribution such as a logarithmic Gaussian distribution gives Kt>0.0. We regarded fulfillment of both −0.3<Sk<0.3 and −0.3<Kt<0.3 as Gaussian distributions.

As computation of Sk and Kt is severely influenced by the presence of extreme values in tails of distribution, we applied a nonparametric truncation procedure before computing Sk and Kt by

excluding data points located outside the following lower and upper extreme limits (eLL, eUL).

$$eLL = Q1 - 3.0 \times (Me - Q1)$$

$$eUL = Q3 + 3.0 \times (Q3 - Me)$$

where Me, Q1, and Q3 represent the median and first and third quartiles of a given distribution, respectively.

2) χ^2-test for normality. On the basis of M and SD of the distribution, test values were partitioned into eight segments by setting seven boundary values between –1.6SD and 1.6SD. Goodness-of-fit to Gaussian distribution can be evaluated from observed and expected frequencies (Oi and Ei) for each segment (i=1,..., 8) as follows [12]:

$$\chi^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i}$$

This method was used to test for normality of a distribution after truncation by the ITC method. However, the size of the data we dealt with was so huge that statistical testing of normality is too sensitive. Therefore, we modified the testing by repeatedly resampling a subset of the original dataset for 100 times and computed the average of the χ^2 values. We arbitrarily set the data size for resampling as 200.

Diagnostic evaluation of wCDI

To evaluate performance of the wCDC method, we conducted a simulation study using the model database. We computed wCDI consecutively for each record, from current to the past for the entire dataset. These values constituted 'natural' wCDI for the control group. To obtain cases of specimen mix-up, we randomly created pairs of unmatched records among those of the same day and computed wCDI. These values represented wCDI for the 'artificial' group. Then, performance of the wCDC method in detecting the artificial group was evaluated in two parts according to the combination of test items included in computing wCDI. In part one, evaluation was limited to those wCDI that were computed for three commonly ordered test sets consisting of six to nine items each. In part two, the evaluation was made according to the number of test items included in computing wCDI, disregarding the combination of test items. A cut-off value to distinguish the natural and artificial groups was determined based on receiver-operating characteristic (ROC) curve analysis [13]. An overall degree of differentiation was expressed as the area under the ROC curve (AUC).

Results

Performance of the power transformation

Effectiveness of the power transformation to normalize patients' test results is shown in Table 1. The test items examined here were all known to have skewed distribution with |Sk|>0.3. The effect of excluding 1% of the data on each tail of the distribution before transformation was examined. The Sk and Kt values were computed for three cases: Case 1 for distributions of the original dataset and Case 2 and Case 3 for distributions after power transformation without and with exclusion of extreme values, respectively. For the sake of space, the analytical results

		1) Original		2) Box-Cox transformation without truncation					3) Box-Cox transformation with truncation			
Name	n[a]	Sk	Kt	p	a	Sk	Kt	n[b]	p	a	Sk	Kt
WBC	75610	0.63[c]	−0.23	0.108	0.2	0.05	0.03	74192	0.316	1.5	0.10	0.04
TP	45272	−0.42[c]	0.04	2.636	0.4	0.01	0.29	44486	2.051	1.4	−0.11	0.29
ALB	36068	−0.43[c]	0.44[d]	1.446	0.6	−0.23	0.67[d]	35464	0.863	1.3	−0.52[c]	0.31[d]
Na	48079	−0.62[c]	−0.04	2.443	118.1	−0.43[c]	0.34[d]	47387	2.550	118.1	−0.43[c]	0.34[d]
Ca	22241	−0.50[c]	0.08	1.393	3.8	−0.40[c]	0.19	21827	2.832	3.8	−0.19	0.34[d]
BUN	72673	0.91[c]	−0.70[d]	0.032	0.6	0.32[c]	−0.03	71412	0.000	5.7	0.07	0.09
CRE	73352	1.13[c]	−1.43[d]	0.562	0.0	0.94[c]	−1.05[d]	71914	0.000	0.3	0.52[c]	−0.35[d]
HDL-C	13035	0.48[c]	0.03	0.169	2.9	0.08	0.23	12786	0.289	21.9	−0.08	0.22
TG	23240	0.98[c]	−0.52[d]	0.024	4.8	0.22	0.25	22799	0.086	40.1	−0.12	0.14
GLU	20736	1.15[c]	−0.92[d]	0.041	23.1	0.74[c]	−0.17	20346	0.016	66.1	0.32[c]	0.21
AST	49325	1.25[c]	−1.33[d]	0.520	1.0	0.94[c]	−0.67[d]	48479	0.028	9.9	0.18	−0.03
ALT	49300	1.38[c]	−1.54[d]	0.021	2.5	0.38[c]	0.04	48455	0.000	5.9	0.11	0.18
LDH	40649	0.71[c]	−0.33[d]	0.023	33.4	0.26	0.05	39837	0.141	103.5	0.03	0.05
ALP	40155	1.02[c]	−0.91[d]	0.026	15.5	0.45[c]	−0.10	39359	0.000	103.5	0.09	0.03
GGT	57020	1.60[c]	−2.11[d]	−0.004	4.7	0.42[c]	0.19	55893	0.000	3.3	0.50[c]	0.16
AMY	18881	0.71[c]	−0.21	0.026	1.8	0.00	0.13	18532	0.200	22.7	−0.09	0.05
APTT	16954	1.06[c]	−0.74[d]	0.026	15.2	0.46[c]	0.06	16623	0.070	21.3	0.06	0.12

Table 1 Effectiveness of Gaussian transformation by modified Box-Cox formula.
[a]Number of original dataset; [b]Number of truncated dataset; Sk, Skewness ([c]|Sk|>0.3); Kt, Kurtosis ([d]|Kt|>0.3).

are shown only for the male dataset. However, the almost identical results were obtained in the female dataset.

From the values of Sk and Kt, it is evident that the original distributions (Case 1) deviate severely from the Gaussian form. Meanwhile, comparison of performance between Cases 2 and 3 showed that prominent increase in the success rate of Gaussian transformation occurred in the latter case. The estimated *p* did not differ much between Case 2 and Case 3, but the estimated *a* did change appreciably, implying that aberrant/unrealistic values often exist in the lower tail of distribution of test values, such as zero, in the LIS, and the exclusion procedure was effective in removing them.

Derivation of within-individual differences

The magnitude of within-individual differences was computed as the SD of the distribution of D_z for each test item. Computed SD of D_z for the 32 test items with or without using the ITC method are listed in Table 2, which clearly shows that the SD by ITC method (aSD) is obviously smaller than the SD without the ITC method. Results of the χ^2-test clearly indicate that the central portion of the distribution of D_z can be regarded as Gaussian. The analytical results in the table are again shown only for the male dataset and those for the female dataset were omitted.

Six typical distributions of D_z for ALB, ALP, Ca, Cl, GGT, and GLU are illustrated in Figure 2. Each has clear peak at D_z=0.0 and shows very smooth symmetrical distribution but has long tails. Theoretical Gaussian curves were drawn over the histograms by use of the original M and SD or the adjusted M and aSD. Nearly perfect fitting of the latter curve to the histogram again indicates that the central portion of the distribution is Gaussian and that the ICT method is very effective in deriving SD unaffected by extreme values in the periphery. As all data were both transformed and standardized, aSD of any test item is now mutually comparable and indicates the magnitude of within-individual difference. Smaller values of aSD in the ascending order were observed for the following test items: TP, MCV, ALP, MCH, ChE, MCHC, and GGT (Table 2).

Name	Without ITC method			With ITC method		
	Mean	SD	χ^2	Mean	aSD	χ^2
WBC	0.03	1.46	36.31[a]	0.07	0.77	9.95
RBC	0.05	1.09	23.99[a]	0.03	0.71	9.78
Hgb	0.07	1.28	25.86[a]	0.04	0.81	10.26
Hct	0.05	1.09	21.90[a]	0.02	0.73	9.80
PLT	−0.02	0.92	42.14[a]	0.00	0.46	10.74
MCH	0.01	0.37	20.89[a]	0.01	0.25	10.08
MCHC	0.01	0.38	12.05[a]	0.01	0.33	10.37
MCV	0.00	0.51	60.47[a]	−0.01	0.21	9.86
TP	0.02	1.06	107.13[a]	−0.02	0.20	12.41[a]
ALB	0.05	1.10	21.90[a]	−0.02	0.75	9.50
Na	−0.02	0.95	16.73[a]	−0.01	0.73	10.08
K	−0.03	1.36	15.28[a]	−0.04	1.03	10.80
Cl	0.00	1.04	14.76[a]	0.00	0.83	9.95
Ca	0.02	1.33	64.08[a]	−0.01	0.66	10.08
IP	0.06	1.88	73.92[a]	0.00	0.82	9.82
BUN	0.00	1.17	40.24[a]	0.00	0.61	10.52
CRE	0.01	1.04	54.67[a]	0.01	0.51	9.97
UA	0.03	1.11	72.94[a]	0.00	0.51	10.32
TC	0.01	0.65	14.97[a]	0.00	0.51	10.92
HDL-C	−0.01	0.60	15.43[a]	0.00	0.47	10.92
LDL-C	0.02	0.78	16.74[a]	0.01	0.59	10.48
TG	0.01	0.79	19.01[a]	0.01	0.55	10.12
GLU	0.04	1.76	60.80[a]	0.02	0.74	10.42
AST	0.02	1.08	37.83[a]	0.04	0.57	9.93
ALT	0.00	0.82	36.37[a]	0.05	0.45	10.10
LDH	0.04	0.91	44.28[a]	0.04	0.46	9.42[a]
ALP	0.00	0.50	62.23[a]	0.02	0.23	9.58
GGT	0.00	0.78	59.82[a]	0.05	0.34	11.32
AMY	−0.01	1.22	74.66[a]	0.00	0.46	10.14
CHE	0.01	0.42	25.89[a]	0.00	0.26	10.07
PT(%)	0.03	1.46	30.06[a]	−0.01	0.86	10.54
APTT	−0.04	0.75	47.60[a]	0.00	0.35	9.52

Table 2 Effectiveness of the iterative truncation and correction (ITC) method in adjusting SD.
[a]$p<0.05$, $\chi^2_{(df=5,\ p=0.05)}=11.07$.

Diagnostic performance of wCDI

Fixed combination of test items

The performance of wCDI in identifying specimen mix-up was investigated by artificially generating cases of mix-up. Although wCDI can be computed for any combination of test items, we examined three commonly ordered test sets: Set 1 (WBC, RBC, Hb, Ht, PLT, MCV, MCH, MCHC), Set 2 (TP, Alb, BUN, CRE, UA, Na, K, Cl, Ca), and Set 3 (ALT, AST, LDH, ALP, GGT, TP).

Results of the simulation study are shown in Table 3A. The accuracy of distinguishing two groups by wCDI, expressed as AUC (4th column), were as high as 0.937–0.967. Sensitivity of correctly detecting artificial cases was determined using a cut-off value of wCDI that gives a false-positive (FP) rate of 5.0%, 7.5%, or 10%. The sensitivities were 63.2%–84.8%, 74.0%–89.2%, and 80.7%–91.6%, respectively (5th–7th columns). The same analysis was done for a special case when wCDI values without weight (or equal weight regardless of aSD) were used for the detection. As expected, it resulted in poor accuracy (Table 3B).

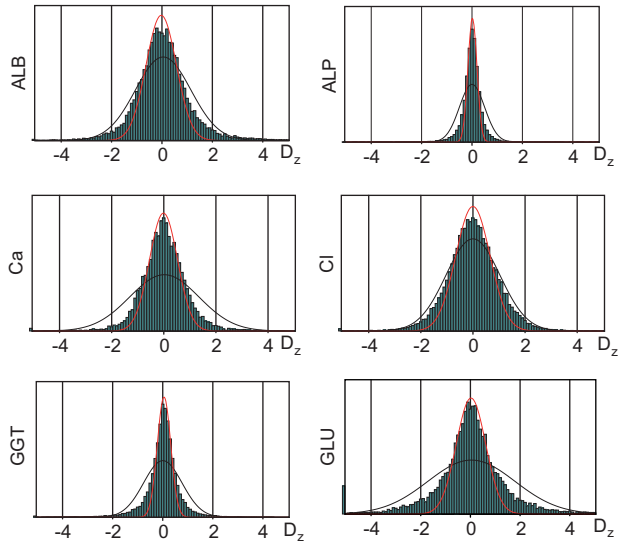

Figure 2 Examples of distribution of deltas (within-individual differences between current and previous z scores).
Theoretical Gaussian curves were drawn over the histogram by use of the original mean and SD of the distribution (curve drawn in black color) and by use of the adjusted mean and SD (aSD) based on the ICT method (curve drawn in red color). The analysis was done separately for male and female, and this figure was made for the male dataset.

Arbitrary combination of test items

We also evaluated the performance of wCDC methods for arbitrary combination of test items. The same simulation study was performed by artificially generating cases of mix-up. Performance was evaluated simply by stratifying wCDI for 'natural' and 'artificial' cases by the number of test items included in the computation. Figure 3 shows how the AUC (Figure 3A), cut-off value, and sensitivity of detection (by setting FP rate at 5%, 7.5%, or 10%) change with the number of test items k (=5–20) used in computing wCDI (Figure 3B and C). AUC and sensitivity increased proportionately for k≤10 but remained almost unchanged for k>10, and the cut-off value decreased until k=10 and remained unchanged for k>10. Therefore, for k>10, cut-off values can be set approximately at 0.90, 0.83, and 0.75, respectively, for FP rates of 5%, 7.5%, and 10%. To determine the effect of a weighting factor in computing wCDI, we also evaluated the results for wCDI with equal weighting and show the corresponding results in broken lines. It is evident that performances are always poor without weighing in the computation.

From these results, we found that it was not necessary to set an individual cut-off value for wCDI for each combination of test items. Rather, we can set the cut-off value to judge wCDI according to the number of test items included in the computation.

Discussion

There have been various attempts to detect possible cases of specimen mix-up in routine clinical laboratory data by use of information techniques. However, real clinical laboratory data are very heterogeneous and contain a number of extreme data. Therefore, simple statistical analysis is of no use in uncovering cases of mix-up. We coped with this problem by applying a series of techniques for data analyses.

First, we converted the distribution of patients' test values into Gaussian with a modified Box-Cox power transformation formula after excluding 1% of extreme values on both ends of the distribution. We proved that this method was very effective in bringing the distribution very close to Gaussian. Although we have found that almost all laboratory test results from healthy individuals can be converted to Gaussian by the Box-Cox method [8, 10], it is of great interest to find that patient test values

	Data size		AUC	Sensitivity, % (cut-off value)		
	Artificial	Natural		FP=5%	FP=7.5%	FP=10%
(A) Fixed test item (with weight)						
Set 1 (WBC, RBC, Hb, Ht, PLT, MCV, MCH, MCHC)	126,211	105,691	0.967	84.8% (0.82)	89.2% (0.73)	91.6% (0.67)
Set 2 (TP, Alb, Na, K, Cl, Ca, BUN, CRE, UA)	4087	10,205	0.953	74.8% (1.13)	81.7% (1.02)	86.2% (0.94)
Set 3 (AST, ALT, LDH, ALP, GGT, TP)	36,327	39,255	0.937	63.2% (1.07)	74.0% (0.92)	80.7% (0.82)
(B) Fixed test item (without weight)						
Set 1 (WBC, RBC, Hb, Ht, PLT, MCV, MCH, MCHC)	126,211	105,691	0.952	76.2% (1.21)	82.5% (1.07)	86.2% (0.98)
Set 2 (TP, Alb, Na, K, Cl, Ca, BUN, CRE, UA)	4087	10,205	0.944	70.0% (1.17)	77.8% (1.06)	83.5% (0.98)
Set 3 (AST, ALT, LDH, ALP, GGT, TP)	36,327	39,255	0.933	61.2% (1.37)	71.8% (1.19)	78.8% (1.07)

Table 3 Performance of the wCDC method when applied to common test sets.
FP, False-positive.

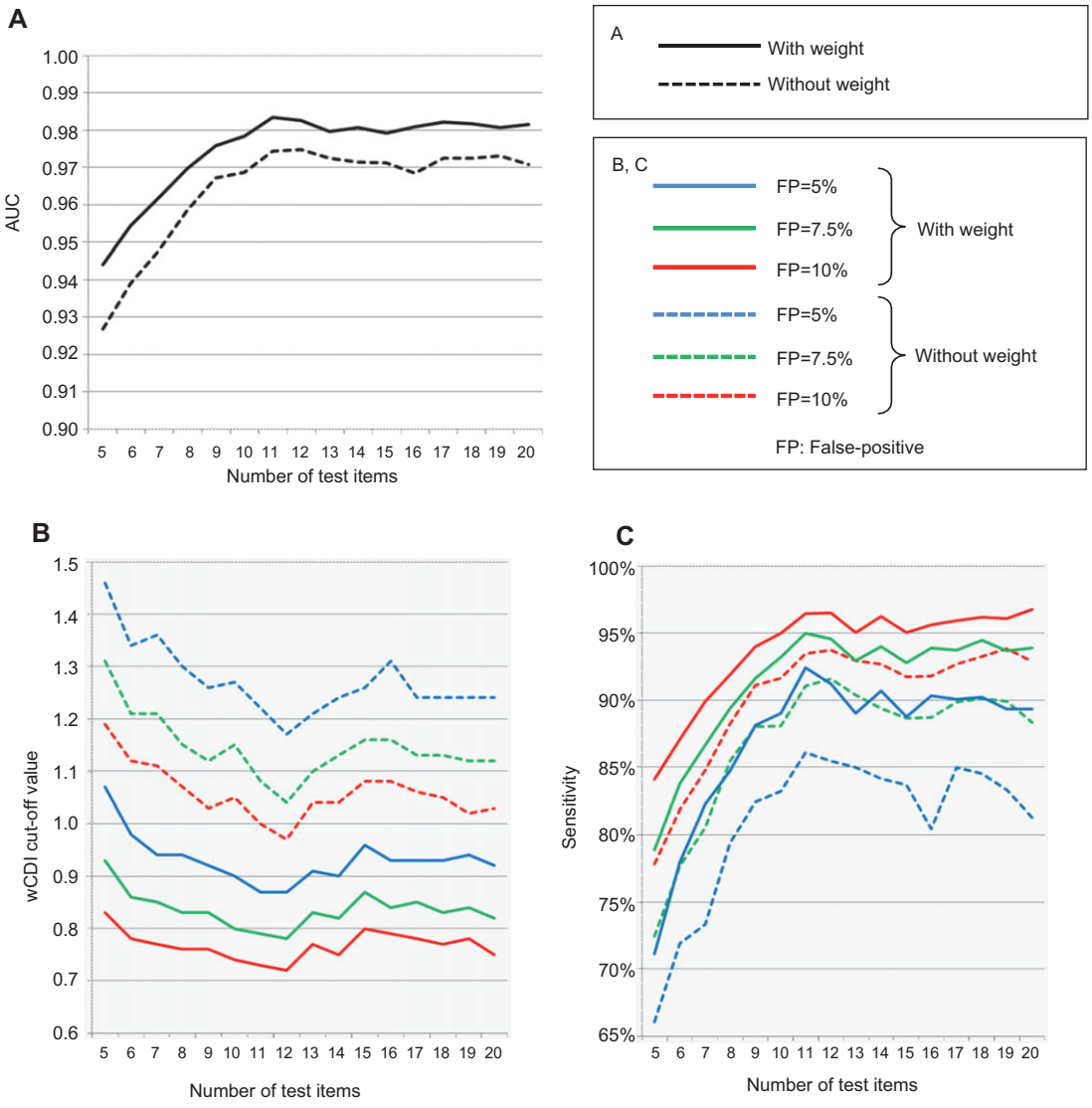

Figure 3 Performance of the wCDC method in relation to number of test items.
(A) Effect of weighting (solid line) versus non-weighting (dashed line) on AUC for distinguishing cases of mix-up. (B) and (C) Cut-off values and sensitivities, respectively, corresponding to false-positive rates of 5% (blue line), 7.5% (green line), and 10% (red line). Dashed lines represent wCDI computed without weighting.

can be also converted to Gaussian for almost all test items simply by prior exclusion of highly extreme values. This implies that the distribution becomes symmetrical, and balanced treatment of abnormal values on lower and higher sides is possible.

In testing for normality by use of Sk and Kt after applying power transformation, we had to truncate data in the tails of the transformed distributions because both parameters are very sensitive to extreme values with cubic and fourth-power terms of deviation from mean, respectively, in the formulae. Tukey's procedure is conventionally used to truncate data outsides ($Q1–1.5\times IQR$ and $Q3+1.5\times IQR$) where IQR represents interquartile range ($Q3–Q1$). However, it assumes symmetrical distribution in the truncation. We overcame this problem by adopting cut-off values reflecting the asymmetry, $Q1–3.0\times(Me–Q1)$ and $Q3+3.0\times(Q3–Me)$, for the lower and upper sides, respectively. They correspond to 0.7 and 99.3 percentile points in the case of Gaussian distribution. We believe it is essential to use our formulae in dealing with laboratory test results that sometimes show highly skewed distribution.

Another step we took was conversion of the transformed value into a z-score to deal with test results uniformly regardless of measurement units. In this conversion, we used a special approach to standardize the value on the basis of the gender-specific RI because we believed

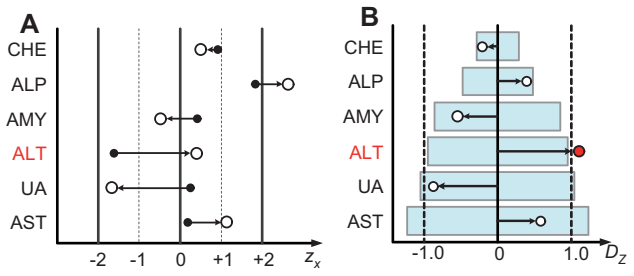

Figure 4 Two-way displays showing differences between current and previous test results provided by the wCDC system when the wCDI value exceeds a certain limit.
(A) Clinical implication view shown in the uniform scale using z-scores. (B) Degree of deviation between the two measurements in reference to the expected variability (aSD) of each test item.

that the standardized value (z-score) should be clinically interpretable by scaling the reference limits (LL, UL) as –1.96 and 1.96. In fact, the merit of scaling z-score based on RI was that we can graphically display two sets of test results for current and past tests as shown in Figure 4A, and users can interpret clinical significance of observed difference in z-scores. This graphical display was built as a part of a system implementing the new delta-check method. In parallel, the system offers a view showing degrees of difference in reference to expected variability (aSD of delta) of each test item (Figure 4B). This way, the system provides information regarding both clinical and analytical implications of the observed difference, thus facilitating final judgment of whether data deviating highly from previous results can be regarded as a case of mix-up.

The most crucial and challenging issue in establishing the wCDC method was to estimate variability of difference in two successive test results from routine laboratory data that included all kinds of extreme results. Actually, distribution of differences between current and previous values (D_z) showed smooth symmetrical distribution but always had a very long tail on either end. This fact implies that SD computed from the entire range of distribution cannot be used as a measure of within-individual differences. However, we found that the central portion of the distribution was clearly regarded as Gaussian by the limited-range χ^2-test. Therefore, we applied the ICT method [9] to obtain an adjusted SD representing the central portion. The ICT method was originally developed to derive unbiased means (center) of test value distributions in external quality control surveys. We proved that the method is also applicable to derive unbiased SD of a distribution containing a large number of extreme values on either or both tails.

In testing normality of distribution of D_z in its central portion, we needed to use the χ^2-test. However, it is very sensitive to data size. Actually, when all the observed frequencies are uniformly multiplied by the factor of m, χ^2 statistics are simply increased m times although the degree of freedom does not change and the cut-off value remains the same as illustrated in the formulae below:

$$\chi^2=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$$

$$\chi^2=\sum_{i=1}^{k}\frac{(\mathbf{m}\times O_i-\mathbf{m}\times E_i)^2}{\mathbf{m}\times E_i}=\mathbf{m}\times\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$$

This property of the statistical test hinders its use with our large-scale data. Therefore, we adopted an approach to apply the method by repeatedly sampling a small subset of the original dataset and taking an average of the χ^2 values. This way we could objectively judge differences in distribution patterns with or without the ITC method. We believe this modification is appropriate in a practical sense because our purpose was only to demonstrate that greatly improved fitting to the Gaussian distribution of the central part is possible with the ITC method compared with not using the method.

With regard to the general applicability of the new delta-check method, we have evaluated the performance of detecting the artificially mixed-up cases with or without limiting the dataset to those from outpatients. There was no appreciable difference in the performance attributable to a change in the proportion of abnormal results in the database. It implies that the Gaussian transformation makes the magnitude of differences in test results equivalent regardless of the test level used for comparison.

Another important consideration in applying the new delta-check method is allowable limit of time interval between two successive measurements. When the interval is too long, the performance may be affected by age-related changes in test results especially for data from pediatric and aged population.

Therefore, the system now sets the maximum time interval to 1 year. Furthermore, the system automatically provides information about the time interval between the two successive measurements so that the user can interpret the implication of the difference from the time interval. However, the system can refresh a list of the SDs for within-individual differences regularly. Therefore, its performance is not affected by a long-term bias in the analytical system or by a shift in the patient population.

A possible problem with the wCDC method could be that wCDI does not take into consideration of correlations

among test items involved in the calculation. Detection of specimen mix-up can be improved by use of Mahalanobis distance of two sets of data in multivariate space after Gaussian transformation and standardization. However, it requires a fixed set of test items for computation and determining the cut-off value. In contrast, wCDI can be computed flexibly for any combination of test items and uses cut-off values according to the number of test items included in computing wCDI.

Conflict of interest statement

Authors' conflict of interest disclosure: The authors stated that there are no conflicts of interest regarding the publication of this article.
Research funding: None declared.
Employment or leadership: None declared.
Honorarium: None declared.

Received November 4, 2012; accepted January 8, 2013

References

1. Quillen K, Murphy K. Quality improvement to decrease specimen mislabeling in transfusion medicine. Arch Pathol Lab Med 2006;130:1196–8.
2. Morrison AP, Tanasijevic MJ, Goonan EM, Lobo MM, Bates MM, Lipsitz SR, et al. Reduction in specimen labeling errors after implementation of a positive patient identification system in phlebotomy. Am J Clin Pathol 2010;133:870–7.
3. Carraro P, Zago T, Plebani M. Exploring the initial steps of the testing process: frequency and nature of pre-preanalytic errors. Clin Chem 2012;58:638–42.
4. Dunn EJ, Moga PJ. Patient misidentification in laboratory medicine: a qualitative analysis of 227 root cause analysis reports in the Veterans Health Administration. Arch Pathol Lab Med 2010;134:244–55.
5. Nosanchuk JS, Gottmann AW. CUMS and delta checks. A systematic approach to quality control. Am J Clin Pathol 1974;62:707–12.
6. Sheiner LB, Wheeler LA, Moore JK. The performance of delta check methods. Clin Chem 1979;25:2034–7.
7. Iizuka Y, Kume H, Kitamura M. Multivariate delta check method for detecting specimen mix-up. Clin Chem 1982;28: 2244–8.
8. Ichihara K, Boyd JC. IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). An appraisal of statistical procedures used in derivation of reference intervals. Clin Chem Lab Med 2010;48:1537–51.
9. Ichihara K, Kawai T. An iterative method for improved estimation of the mean of peer-group distributions in proficiency testing. Clin Chem Lab Med 2005;43:412–21.
10. Ichihara K, Kawai T. Determination of reference intervals for 13 plasma proteins based on IFCC international reference preparation (CRM470) and NCCLS proposed guideline (C28-P,1992): trial to select reference individuals by results of screening tests and application of maximal likelihood method. J Clin Lab Anal 1996;10:110–7.
11. Sokal RR, Rohlf FJ. The normal probability distribution. In: Biometry, 2nd ed. New York: WH Freeman & Company, 1981: 98–127.
12. Bland M. The chi-squared goodness of fit test. In: An introduction to medical statistics, 2nd ed. New York: Oxford University Press Inc., 1995:244–5.
13. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–77.