The Twenty-Seventh International Symposium on Artificial Life and Robotics 2022 (AROB 27th 2022),
The Seventh International Symposium on BioComplexity 2022 (ISBC 7th 2022),
The Fifth International Symposium on Swarm Behavior and Bio-Inspired Robotics 2022 (SWARM 5th 2022)
ONLINE, January 25-27, 2022

An effect of the exclusion criteria on the distribution of blood test values

Rina. Kagawa[1], Masanori Shiro[2]

[1]University of Tsukuba, Japan
[2]Advanced industrial science and technology, Japan
(Tel: 81-29-861-4189)

[1]shiro@aist.go.jp

Abstract: Our objective is to clarify the most accurate distributions of the blood test items commonly used in health checkups. In this study, we used three data sets and assumed the LogNormal distribution with three parameters. We defined the distances between the distributions and tested whether the setting of the exclusion criteria or the Modified Box-Cox transformation had a greater effect on the shape of the distribution. From this analysis, it was found that the setting of the exclusion criteria had an important influence on the shape of the distribution of blood test values.

Keywords: exclusion criteria, log-normal distribution, blood test values

1 INTRODUCTION

1.1 Background

The increasing demand for personalized health care, the accumulation of medical data based on the worldwide spread of electronic health records, and many data-intensive genomic studies in precision medicine have led to the expectation that quantitative evaluation of human disease states is possible based on each patient's context [1]. However, this has not yet been achieved at a sufficiently low cost [2]. One of the hurdles arises from the observed data itself. For example, in clinical practice, vital sign measurements and laboratory tests are conducted at irregular intervals [3]; this causes difficulties in developing a methodology for evaluating patients' conditions based on these data [4].

Although the references for comparison are necessary for quantitative evaluation of the observed values, it has not been fully established how the references should be determined. Our goal is to establish the references for comparison with observed values. We consider that the expected value of the distribution obtained from the healthy population is the best reference for comparison because the expected value could represent the healthiest condition and because the deviation from the expected value could be one of the computable measures used to evaluate changes in patient condition.

Our study will focus on laboratory tests to evaluate patients' disease status. We consider that a reference for comparison must be set for each laboratory test item. The measured values of laboratory test items must be evaluated for deviation, including mean absolute deviation and standard deviation from the reference. However, the distributions obtained from the healthy population using most laboratory test items have not been published; therefore, we need to assume some distribution. For calculation of the deviations at low cost, a parametric distribution is expected because the expected and integral values of the parametric distributions can be calculated quickly with high accuracy if the parameters of the distribution can be accurately estimated.

The objective of this study is to clarify the parametric distribution and the parameters of the reference for comparison for each laboratory test item. However, only two endpoints of RI and one median (we call these three values as *the three values*) have been published for some laboratory test items [5][6]. For a few laboratory test items, histograms have been also published [7], but the expected value cannot be calculated accurately from the histogram because the widths of the histogram's bins are arbitrarily determined. Moreover, these histograms and previous studies showed asymmetric distributions of many laboratory test values from multiple hospitals [8]. Since the expected value and the median are different in the asymmetric distribution, the expected value of the distribution is difficult to calculate based only on the three values of the asymmetric distribution. Therefore, the deviation of the observed value from the asymmetric distribution cannot be accurately calculated.

1.2 Results previous study clarified and purpose of this study

We have shown in our previous study [9] that many laboratory test values follow the Log Normal distribution with three parameters (hereinafter, we call that LN3). The density function of LN3 is shown in Eq.1.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma(x-d)}\exp\left(-\frac{(\ln(x-d)-\mu)^2}{2\sigma^2}\right) \quad \text{(Eq. 1)}$$

Furthermore, our previous study has confirmed that there were differences between the lognormal distributions estimated based on the data collected in actual hospitals or all over Japan and those based on the previous research. Previous studies have pointed out that the RIs of the distribution differ depending on gender [5][6][10]. However, from our previous results, it is inferred that the difference between the properties of the four datasets also

affected the shape of the distributions as much as or more than the gender difference.

Therefore, in this study, we quantified the differences between the estimated distributions based on the data-sets and discussed the cause of the differences.

2 Materials and Methods for estimation of the distributions

Materials

This study used four datasets as in our previous study [9], and the laboratory test items included in each dataset are shown in Table 1. The abbreviations for laboratory test items are drawn from a previous study [11]. The results for males and females were used as different data for each laboratory test item. Unless otherwise stated, exclusion criteria were not defined for these datasets.

1. University of Tsukuba Hospital (UTH)

Laboratory test values were obtained for medical check-ups for 518 males (M) and 512 females (F) over the age of 20 (average age = 60.72) who had medical check-ups between January 1, 2017 and December 31, 2018 at University of Tsukuba Hospital. For persons who had multiple medical check-ups, the values from the first check-up for each laboratory test were used for the analysis.

2. National Health and Nutrition Survey (NHNS)

This is a health screening project in Japan, and the histograms of some laboratory tests were published [7]. We used data from 2013 to 2019, excluding 2016 because most of those data were missing, according to the data source policy [7]. A total of 7,632 males and 10,796 females over the age of 20 were targeted. The values of the histogram boundaries for each laboratory test item were the same for every year. For easy comparison with the other datasets, we obtained one histogram by averaging the frequencies contained in each bin of the annual original histogram, which is normalized for each laboratory test item.

3. Japanese Committee for Clinical Laboratory Standards (JCCLS)

This dataset contains the three values of laboratory tests of Japanese individuals (2,733 males and 3,612 females, ages 18–72) [6][11]. Exclusion criteria were based on parameters including body mass index, alcohol intake, and so on, designed to recruit healthy subjects [6] because RIs are generally calculated based on a healthy population.

The method used to create this dataset was as follows [6][11]. The histograms obtained via measurements were converted into normal distributions using Modified Box-Cox transformation [12], and the values corresponding to 5% and 95% intervals were obtained. The RIs were then obtained by inverse transformation of the value. The density function $h(x)$ we explicitly expressed is shown in Eq. 2.

$$h(x) = \frac{(x-\mu+4\sigma)^{p-1}}{\sqrt{2\pi}\sigma} exp\left(-\frac{\left(\frac{(x-\mu+4\sigma)^{p}-1}{p} - m\right)^{2}}{2\sigma^{2}} \right) \quad \text{(Eq. 2)}$$

In Eq. 2, because the shift of x without expected value μ was fixed as 4σ, some types of histograms might not be fitted to $h(x)$ appropriately. The errors caused by these transformations would affect published RIs.

4. Ichihara+ 2013 (Ichihara)

The three values for each laboratory test item of Japanese individuals (2,082 in total for males and females, ages 20–65) are shown in Ichihara et al. [5]. Similar to JCCLS, RIs were calculated on this dataset using Modified Box-Cox transformation. The exclusion criteria were almost the same as in JCCLS.

Previous work in which the median of each laboratory test item has not been published [13] were excluded from our experiments. A national medical study conducted in Japan [14] was also excluded because the data published in this study were biased and the distribution parameters cannot be estimated.

Table 1– Laboratory test items

Dataset	Items
UTH	BRC, Hb, Ht, TP, Alb, AST, ALT, γGT, ALP[1], AMY, LD, TC, TG, HDL-C, LDL-C, Cr, UA, Na, K, Cl, Ca, HbA1c, Glu,
NAHS	AST, ALT, γGT, TC, TG, HDL-C, LDL-C, Cr, UA,
JCCLS	BRC, Hb, Ht, TP, Alb, AST, ALT, γGT, ALP, AMY, LD, TC, TG, HDL-C, LDL-C, CK, Cr, UA, Na, K, Cl, Ca, HbA1c, Glu, IgG, IgA, IgM, C3, C4
Ichihara	Alb, AST, ALT, γGT, AMY, LD, TC, TG, HDL-C, LDL-C, Cr, UA, Na, K, Cl, CK, Ca, IgG, IgA, IgM, C3, C4

Methods for estimation of the distributions

Prior to the experiments, we used LN3 to estimate the distribution parameter sets of CUTH[2], JCCLS, and Ichihara for each of the 58 laboratory test items stratified by gender. For CUTH and NAHS, the sum of the squared errors for each laboratory test item was calculated between the value at the center of each column of the histogram and for the

[1] Please see the supplemental data for the details of ALP data.

[2] We apply the label *CUTH* to the converted distribution of UTH using kernel density estimation.

points at the same location on the estimated distribution; therefore, the parameter set of the distribution was determined using the Nelder-Mead method to minimize the sum of the squared errors. For JCCLS and Ichihara, the three values were provided for each laboratory test item, allowing for a unique determination of the parameter set for LN3. The parameter set of each statistical distribution that minimized the sum of the squared errors was determined using the Nelder-Mead method.

Two experiments were accomplished based on the estimated parameter sets for each laboratory test item of each dataset.

The computing environment used was R 3.6.3, Perl 5.30, Python 3.7.2, and SciPy 0.6.6 with Ubuntu 20.04 LTS.

3 Experiment 1: Comparison of expected values between datasets for each laboratory test item

Our previous work [9] pointed out that the estimated distributions of CUTH and NAHS were similar, and those of Ichihara and JCCLS were similar, moreover, the estimated distributions of the former group (those of CUTH or NAHS) and those of the latter group (those of Ichihara or JCCLS) were not similar.

To clarify the cause of the differences of the estimated distributions based on the datasets, we focused on two features common to Ichihara and JCCLS: (1) setting the similar exclusion criteria and (2) using the Modified Box-Cox transformation. For Ichihara and JCCLS, RIs were estimated by the inverse transformation of the Modified Box-Cox transformation after converting the measured original data to a normal distribution [5][6]. Ichihara and JCCLS were based on the common exclusion criteria to recruit healthy subjects, and these exclusion criteria have not been applied for CUTH. If this has a dominant influence on the estimation of distributions, the shape of the estimated distributions of Ichihara and JCCLS would change to the shape that is interpreted as healthier than CUTH's shape of estimated distributions. For example, the peak position of Ichihara's AST (M) would be smaller than that of CUTH's AST (M). If not, that tendency would not be observed.

In this study, we investigate the possibility that exclusion criteria of datasets is the cause of the differences of the estimated distributions based on the datasets.

3.1 Methods

To test the hypothesis that exclusion criteria of datasetsare the cause of the differences of the estimated distributions based on the datasets, We focused on the expected value of the distribution for our examination. If the exclusion criteria are the main cause of changes in distributions, then the direction of change in expectations compared to CUTH for JCCLS and Ichihara would be consistent with the direction of change in expectations that physicians would expect using their medical knowledge of the populations in each data set for many laboratory test items. We determined the direction whether the expected value (Eq. 3) of the LN3 distribution estimated from Ichihara or JCCLS is larger or smaller than the expected value from CUTH. We call this direction of difference from the expected value 'DFE'.

$$exp\left(\mu + \left(\frac{\sigma^2}{2}\right)\right) + d \qquad \text{(Eq. 3)}$$

Moreover, one of the authors, a physician, predicted DFEs based on common clinical knowledge and experiments. The DFE calculated based on our estimated distributions and the DFE predicted by one physician were compared for each laboratory test item.

3.2 Results

Among the 46 laboratory test items by gender, the DFE calculated based on our estimated distributions and the DFE predicted by one physician agreed for all 42 items but four items: AST (F), Hb (F), Ht (F), and ALP (M) (see Table 2). From this result, it was shown that the direction of difference in the expected value of the estimated distribution of JCCLS and Ichihara compared to those of CUTH tends to be due to the exclusion criteria.

	Consistency between the DFEs and the physician's prediction of the direction of the change	
laboratory test items	*JCCLS*	*Ichihara*
ALB (F/M), ALP (F), ALT (F/M), AMY (F/M), AST (M), CHO (F/M), CRE (F/M), Ca (F/M), Cl (F/M), Fasting blood sugar (F/M), HT (F), Hb (F), HbA1c (F/M), HdL-C (F/M), K (F/M), LD (F/M), LdL-C (F/M), Na (F/M), Number of red blood cells (F/M), TG (F/M), TP (F/M), UA (F/M), gamma-GTP (F/M)	✔	✔
AST (F)	✔	X
Hb (F), Ht (F), ALP (M)	X	✔

Table 2: Results of consistency between the DFEs and the physician's prediction of the direction of the change

4 Experiment 2: Distances between the estimated distributions of each dataset

The results of Experiment 1 suggested that the differences in the shapes of the estimated distributions were due to the exclusion criteria. Therefore, in Experiment 2, we quantitatively evaluated the difference in the shape of the estimated distribution corresponding to the exclusion criteria.

4.1 Methods

For any pair of estimated distributions, we defined the distance as the sum of the absolute values of the differences on the PDFs of the two distributions. Specifically, we calculated the sum of the squared differences taken at 10,000 points at the same location on each of the distributions.

4.2 Results

Figure 1 shows the distribution of distances when the distance between Ichihara and JCCLS and the distance between CUTH and NAHS are defined as *Near* and the other distances[3] are defined as *Far* for each laboratory test item. The horizontal axis represents distance, and the vertical axis represents PDF, or the frequency of distances. The sample size for kernel density estimation was 62 (Near) and 114 (Far). The average values of PDF were 0.335 (Far) and 0.179 (Near), and the null hypothesis that the average values of PDF of Far and Near were equal has been rejected ($p < .001$; effect size = 0.767).

Therefore, it is quantitatively shown that the shapes of the four datasets can be divided into two groups—CUTH and NAHS, and JCCLS and Ichihara—which are similar to each other.

The small peak at around 1.0 for Near is due to the fact that the RIs of γGT in Ichihara are different from the RIs of γ GT in JCCLS. Small peaks were observed around 0.5 for Near and Far, but this is difficult to explain rationally. As shown in Figure 1, the distances of the many pairs of the two distributions were less than 0.8. Since the distance is the sum of absolute difference of 10,000 points, a difference in the PDF per point is considered to be within $8x10^{-5}$ on average. Therefore, the difference of the PDF per point should be assumed to be approximately $8x10^{-5}$ on average when the exclusion or inclusion criteria for population selection used to obtain the three values, that is required to estimate the distribution by LN3, would be unknown. The influence of the difference in the shape of the distribution on the value of two endpoints of RI depends on the shape of the distribution. Therefore, it is not generally possible to estimate how two endpoints of RI change from the shape of the distribution.

[3] The distance between Ichihara and CUTH, the distance between Ichihara and NAHS, the distance between JCCLS and CUTH, and the distance between JCCLS and NAHS.

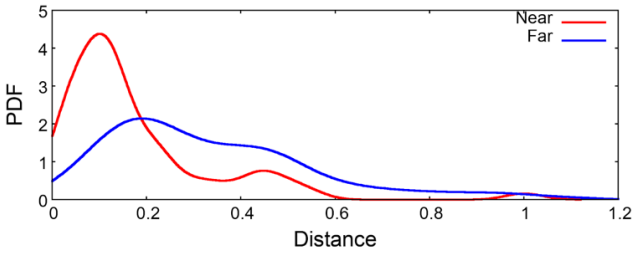

Figure 1– Distribution of distance as the sum of absolute values of difference between the estimated distributions

5 Discussion

We calculated the distance between the estimated distributions using the sum of the absolute value of errors, but the distance could also be estimated with other measures, such as symmetrical Kullback-Leibler divergence. Our results showed that the use of statistical distributions other than lognormal distribution should not be precluded. Our results also showed that even when the original data were examined by gender, the influence of the criteria for selecting the population, which affects the shape of the distribution of laboratory test values, cannot be ignored. Therefore, the distribution should be estimated based on well-defined inclusion or exclusion criteria. However, the effect of a smaller sample size on shape when determining the distribution is not negligible. One way to solve this dilemma is to agree on one well-defined criterion. However, it is difficult to reach a broad consensus on one criterion—even for a specific purpose, such as recruiting a healthy subject. Data accumulation is progressing rapidly, and the distribution can be determined stably by taking a large sample size without defining exclusion or inclusion criteria.

Limitations and future work

The data used in this study were acquired exclusively in Japan. Another limitation is that the four datasets did not feature the same population age. For one laboratory test item, when the median value of the distribution is unknown and only RIs are published, our method would not be applicable to the item. Even if the three values of the distribution of a laboratory test are given, when two or three of them are close in value, it is not possible to determine with much clarity the three parameters of the distribution [15]. A future study could determine what types of two-parameter distribution are appropriate for such laboratory test items. Another future study could better clarify the effect of the criteria of population selection on the distributions. For example, we will try to estimate the distribution obtained from inpatients and provide detailed analysis, including the stratification by diseases. In other future work, we will estimate the shapes of the simultaneous distributions of pairs of laboratory test items.

6 Conclusion

We clarified that the differences in the shapes of the estimated distributions were due to exclusion criteria. We also clarified the differences between estimated distributions based on the three values when no criteria is published. The results of this research are expected to be incorporated into several clinical tests. We believe that the results will be the foundation for constructing references of quantitative evaluation of human disease states in the future.

Ethics

This research was approved by the Ethics Committee of the University of Tsukuba Hospital (permission number: R1-080).

Acknowledgements

The present study was supported in part by the Japan Science and Technology Agency (JST)-Mirai Program Grant Number JPMJMI19G8, the Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research (Nos. JP18H06363, JP19K19347), R&D Center for Frontiers of MIRAI in Policy and Technology, and TIA nano KAKEHASHI Tsukuba Innovation Grants. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Ms. Emiko Nishida, Ms. Mika Sumimoto, and Ms. Noriko Ohkubo have greatly contributed to the manual data processing.

References

[1] L.A. Simmons, M.A. Dinan, T.J. Robinson, and R. Snyderman, Personalized medicine is more than genomic medicine: confusion over terminology impedes progress towards personalized healthcare, *Pers Med* **9** (2012), 85-91.

[2] Kasztura, M., Richard, A., Bempong, N. E., Loncar, D., & Flahault, A. Cost-effectiveness of precision medicine: a scoping review. *International journal of public health* **64** (2019), 1261-1271.

[3] T.A. Lasko, J.C. Denny, and M.A. Levy, Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data, *PloS One* **8** (2013), e66341.

[4] X. Li, D. Zhu, and P. Levy, Predicting clinical outcomes with patient stratification via deep mixture neural networks, *AMIA Jt Summits Transl Sci Proc* **2020** (2020), 367-376.

[5] K. Ichihara, F. Ceriotti, T.H. Tam, S. Sueyoshi, P.M. Poon, M.L. Thong, *et al*., The Asian project for collaborative derivation of reference intervals:(1) strategy and major results of standardized analytes, *Clin Chem Lab Med* **51** (2013), 1429-1442.

[6] K. Ichihara, Y. Yamamoto, T. Hotta, S. Hosogaya, H. Miyachi, Y. Itoh, *et al*., Collaborative derivation of reference intervals for major clinical laboratory tests in Japan, *Ann Clin Biochem* **53** (2016), 347-356.

[7] Ministry of Health, Labour and Welfare, The National Health and Nutrition Survey. Available from: https://www.mhlw.go.jp/bunya/kenkou/kenkou_eiyou_chousa.html [cited 2021 Mar 27] (in Japanese)

[8] M. Shiro, T. Kohro, and R. Kagawa, Development of pseudo datasets of blood test for nonprofit use, *37th Jt. Conf. Med. Inform*. **37** (2017) S152-S155

[9] Masanori Shiro, Rina Kagawa: Validity of lognormal distribution in analyzing laboratory test values to quantitatively evaluate patient's context of disease status, MEDINFO2021, Studies in health technology and informatics (in press)

[10] Japanese Committee for Clinical Laboratory Standards, Shared reference range of major laboratory test items in Japan (2020), https://www.jccls.org/wp-content/uploads/2020/02/2020013103.pdf [cited 2021 Mar 27] (in Japanese)

[11] Japanese Committee for Clinical Laboratory Standards, Shared reference range of major laboratory test items in Japan, (2019) https://www.jccls.org/wp-content/uploads/2020/11/public_20190222.pdf [cited 2021 Mar 27] (in Japanese)

[12]Ichihara, K., & Kawai, T. (1996). 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. Journal of clinical laboratory analysis, 10(2), 110-117.

[13] M. Yamakado, K. Ichihara, Y. Matsumoto, Y. Ishikawa, K. Kato, Y. Komatsubara, *et al*., Derivation of gender and age-specific reference intervals from fully normal Japanese individuals and the implications for health screening, *Clin Chim Acta* **447** (2016), 105-114.

[14] Ministry of Health, Labour and Welfare, NDB open data, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000177182.html [cited 2021 Mar 27] (in Japanese)

[15] Schmittroth, F. (1979). A method for data evaluation with lognormal distributions. *Nuclear Science and Engineering*, *72*(1), 19-34.

Address for correspondence
Masanori Shiro is the corresponding author on this report.
The email address is shiro@aist.go.jp.