Hiroshima J. Med. Sci.
Vol. 55, No. 1, 17~28, March, 2006
HIJM 55–3

A Semi-parametric Method for Describing the Age-specific Distribution of Clinical Measurements in Cross-sectional Study

Mohd. Muzibur RAHMAN[1)], Kenichi SATOH[2)], Keiko OTANI[2)], Masayuki KAMBE[3)] and Megu OHTAKI[2)]

1) Graduate School of Biomedical Sciences, Hiroshima University, 1–2–3 Kasumi, Minami-ku, Hiroshima 734–8551, Japan
2) Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, 1–2–3 Kasumi, Minami-ku, Hiroshima 734–8551
3) Department of Clinical Laboratory Medicine, Division of Medical Intelligence and Informatics, Programs for Applied Biomedicine, Graduate School of Biomedical Sciences, Hiroshima University, 1–2–3 Kasumi, Minami-ku, Hiroshima 734–8551

ABSTRACT

Age-specific distribution of clinical measurements in cross-sectional study is described in this paper. Since the distribution of measurements usually varies with age, a model with an age-dependent structure is needed. We propose here a statistical method for describing the age-specific distribution using an extension of the power-normal-model. The age-dependent parameters are to be estimated through a nonparametric smoothing technique based on the local likelihood method. As a consequence, we can compute a smoothed percentile curve of measurements with reference to age. Several kinds of clinical measurements are analyzed to determine the proposed method.

Key words: *Age-dependent measurement, Box-Cox transformation, Cross-sectional study, Percentile estimation, Local likelihood method, Power-normal-model*

In a prognosis study, information on the percentile of measurement is helpful for the physician to screen target diseases. This requires to know the percent points of the distribution of clinical measurements beforehand. In order to estimate the percent points of distribution of measurements, either the normal distribution model or lognormal distribution model has been applied so far. It is known that the lognormal model often performs well for a left skewed distribution of data. It should be noted, however, that some measurements can not be fitted by these models, and an appropriate alternative model is needed. One effective solution may be to use the Power-normal distribution model[3)], which is based on the Box-Cox[1)] transformation for normalization. This can be described as follows: Given an original measurement of taking a positive value, the transformed value $y^{(\lambda)}$ is defined by

$$y^{(\lambda)}=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda}+\lambda, & (\lambda\neq 0),\\ \log y, & (\lambda=0),\end{cases} \qquad (1)$$

where λ is the shape parameter, often termed the "power parameter". Formula (1) of Box-Cox transformation[5)] is modified to a certain extent from the original version to satisfy the identity equation $y^{(\lambda)}=y$, when $\lambda=1$. The power normal model is based on the hypothesis that the transformed variable has a normal distribution (approximately). Both the normal model and lognormal model can be regarded as special cases of the power normal model.

To illustrate the problem, a set of histogram and box-plot of the RBC (Red Blood Cell) measurements at the age of about 40 years is depicted in Fig. 1 (the data source is Hiroshima University Hospital), where (a), (b) and (c) are the plots for logarithmic transformed measurements, the original one and an optimized power transformed one using the Box-Cox model, respectively. The distribution of original measurements is skewed to the right with a long left side tail. The computed values of the skewness are –1.15, –0.74 and –0.43 for the logarithmic transformation, the original one and the optimized Box-Cox power transformation, respectively. Thus, the skewness for the logarithmic transformation (a) is larger compared to the original one (b), while the distribution is not skewed (approximately) for the optimized Box-Cox power transformation (c). It should also be noticed that the distribution of measurements usually

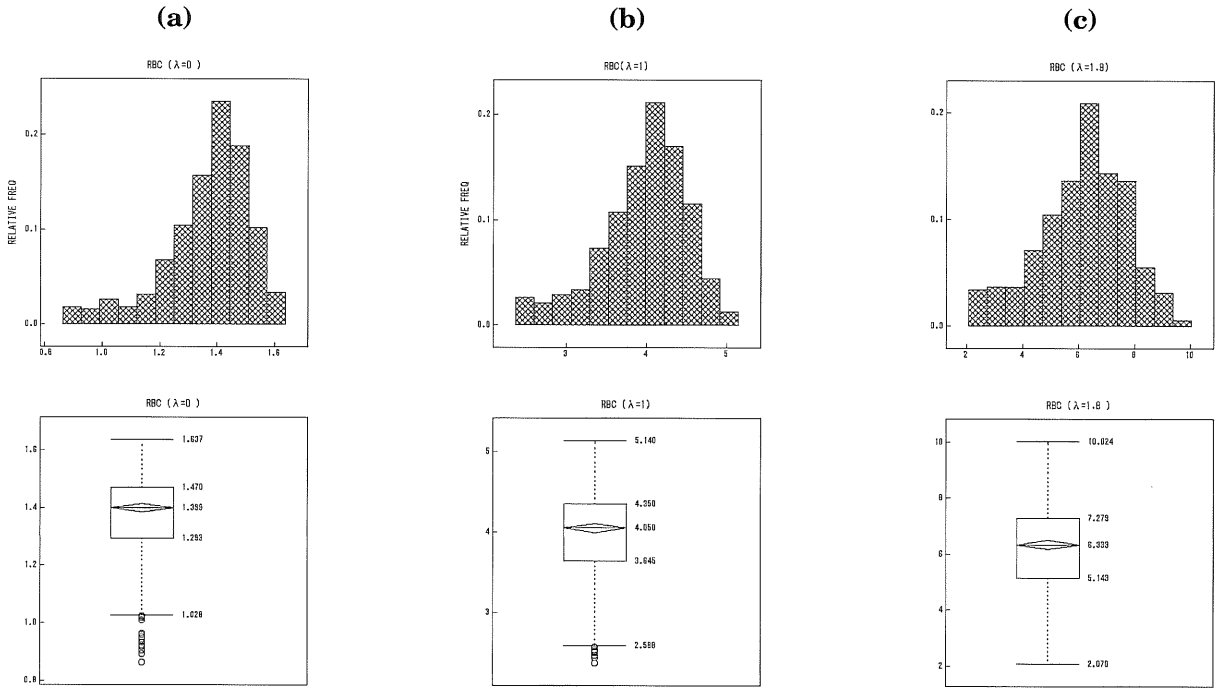

Fig. 1. The histogram and box-plot of the RBC (Red Blood Cell) measurements are shown at the top and bottom, respectively.
(a), (b) and (c) are the plots for logarithmic transformed measurements (λ=0), original one (λ=1) and optimized power transformed one using the Box-Cox model (λ=1.8), respectively.

varies with age so that it is necessary to use a model with an age-dependent structure[10]. Since the original version of the Box-Cox transformation consists of a single power parameter, all the objectives may not be fitted well simultaneously, which has been pointed out by many researchers[3,4,8,9]. So far, in practical applications, statistical information has been commonly obtained by dividing the objects into appropriate strata.

In this paper, we propose an alternative statistical method for describing the age-specific distribution of clinical measurements using an extension of the power normal model with age-dependent parameters, which are to be estimated through a nonparametric smoothing technique.

MATERIALS AND METHODS

The data were collected from individuals, who underwent blood examinations in Hiroshima University Hospital during the period of 1st to 30th September, 2005. All the available data were collected during the above period, among them the total number of males was 5616 and females was 5088. Forty five different types of laboratory tests for clinical prognosis were performed. All the individuals were not required to undergo all the 45 investigations for their prognosis. Thus, the sample sizes of the laboratory tests are different from each other. We analyzed both the male and female data. In this paper we used only the female data for illustration. Age dependency of the measurements seems to be more complicated in female data compared to male data. The minimum and maximum age of the blood examinees were 0 and 99 years respectively.

METHOD FOR ESTIMATING THE AGE-SPECIFIC DISTRIBUTION OF MEASUREMENTS

Let $\{(y_i,t_i)\,|\,i=1,\ldots,N\}$ be a set of independent cross-sectional observations, where y_i denotes the measurement at age t_i for the ith individual. We assume that each measurement is a sample drawn from a population of power normal distribution having the following density,

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y^{(\lambda)}-\mu)^2}{2\sigma^2}\right)y^{\lambda-1}, \tag{2}$$

where the $y^{(\lambda)}$ is given in (1) and is distributed as normal with mean μ and variance σ^2. Using the local likelihood method[6] with the model (2), we can estimate the unknown parameters μ,σ^2 and λ as a function of age, and the steps of estimation are stated as follows.

Step 1. For fixed age t_0, calculate the local weights using the normal kernel function with a bandwidth h, which are specified as:

$$w_i(t_0) = K_h(t_i, t_0) \equiv \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(t_i - t_0)^2}{2h^2}\right),$$

for $i=1, \ldots, N$.

Step 2. For given power parameter λ, make a transformation as follows:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda}-1}{\lambda} + \lambda, & \lambda \neq 0, \\ \log(y_i), & \lambda = 0, \end{cases}$$

for $i=1, \ldots, N$.

Step 3. Calculate the local weighted mean and variance, which are given by

$$\hat{\mu}(\lambda, t_0) = \frac{\sum_{i=1}^{N} w_i(t_0) y_i^{(\lambda)}}{\sum_{i=1}^{N} w_i(t_0)} \quad \text{and}$$

$$\hat{\sigma}^2(\lambda, t_0) = \frac{\sum_{i=1}^{N} w_i(t_0) \{y_i^{(\lambda)}\}^2}{\sum_{i=1}^{N} w_i(t_0)} - \hat{\mu}(\lambda, t_0)^2,$$

respectively. Note that $\hat{\mu}(\lambda, t_0)$ is well-known as the Nadaraya-Watson estimator[7,11].

Step 4. Obtain numerically the estimate of power parameter λ of the Box-Cox transformation for fixed age t_0, so that $\hat{\lambda}(t_0)$ attains the maximum value of the local likelihood function:

$$l(\lambda, t_0) = -\frac{1}{2} \log\{2\pi\hat{\sigma}^2(\lambda, t_0) + 1\} + (\lambda - 1) \frac{\sum_{i=1}^{N} w_i(t_0) \log(y_i)}{\sum_{i=1}^{N} w_i(t_0)}.$$

That is, $\hat{\lambda}(t_0) = \arg\max_{\lambda} l(\lambda, t_0)$. The likelihood function is based on the weighted arithmetic mean after logarithmic transformation.

Step 5. Estimate the α-percent points of the distribution of original measurement y for fixed age t_0, which is specified as:

$$y_\alpha(t_0) = \begin{cases} \{\hat{\lambda}(y_\alpha^{(\hat{\lambda})} - \hat{\lambda}) + 1\}^{\frac{1}{\hat{\lambda}}}, & \hat{\lambda} \neq 0, \\ \exp(y_\alpha^{(\hat{\lambda})}), & \hat{\lambda} = 0, \end{cases}$$

where $y_\alpha^{(\lambda)}$ is the α-percent points of the transformed distribution defined in step 2.

For implementation of the above algorithm, we adopted the value of bandwidth h="range of measurement"×0.1 regardless of measurements in this work. Here, the value of bandwidth h has not been optimized. We selected the value 0.1 as trial and error and also for simplification or simple representation. The value of h will be optimized in future. This has extended the way of estimating the α-percent points to the distribution of the original measurement y for the successive fixed age over the age interval. Subsequently these formed the smoothed percentile curves of the original measurements y over age t.

APPLICATION TO REAL DATA ANALYSIS

The proposed local smoothing method was applied to the following 7 clinical tests; the measurements of Red Blood Cell (million/cu.mm), Phosphorus (mg/dl), Lymphocyte (%), Alkaline Phosphate (IU/liter), Zinc Sulfate Turbidity (U), Magnesium (mg/dl) and C-Reactive Protein (mg/dl). The results are shown in Figs. 2–8.

Red Blood Cell (RBC)

Fig. 2 (a) shows the smoothed local skewness of the distribution over age. The symbol ξ denotes the local skewness. The distribution is skewed to the left when ξ is positive and skewed to the right when ξ is negative. The distribution shows bilateral symmetry when ξ is zero. The distribution of RBC is skewed to the right with a long left side tail almost over age. Fig. 2 (b) shows the maximum likelihood estimates of power parameter λ. The value of λ varies with the corresponding value of ξ over age. When λ takes the value zero, the logarithmic transformation is available, otherwise the power transformation is available. When $\lambda=1$, the identical transformation is available. For this measurement (RBC), the value of λ is maximum at the age of about 40 years. The power transformation ($\lambda=1.8$) is appropriate around the age of 40 years. Fig. 2 (c) shows the mean trend of measurements and the smoothed percentile curves over age. In Fig. 2 (c), the solid curve represents the smoothed mean trend of measurements over age. The dotted curves represent the 5% and 95% points, the dash dotted curves represent 10% and 90% points and the dash curves represent 20% and 80% points. The width between the same kinds of curves represents the variance of measurements. It shows that the variance is not consistent but changes with age. The variance is relatively large in the younger age group and small around the age of 45 years. The original measurements of RBC are superimposed for reference.

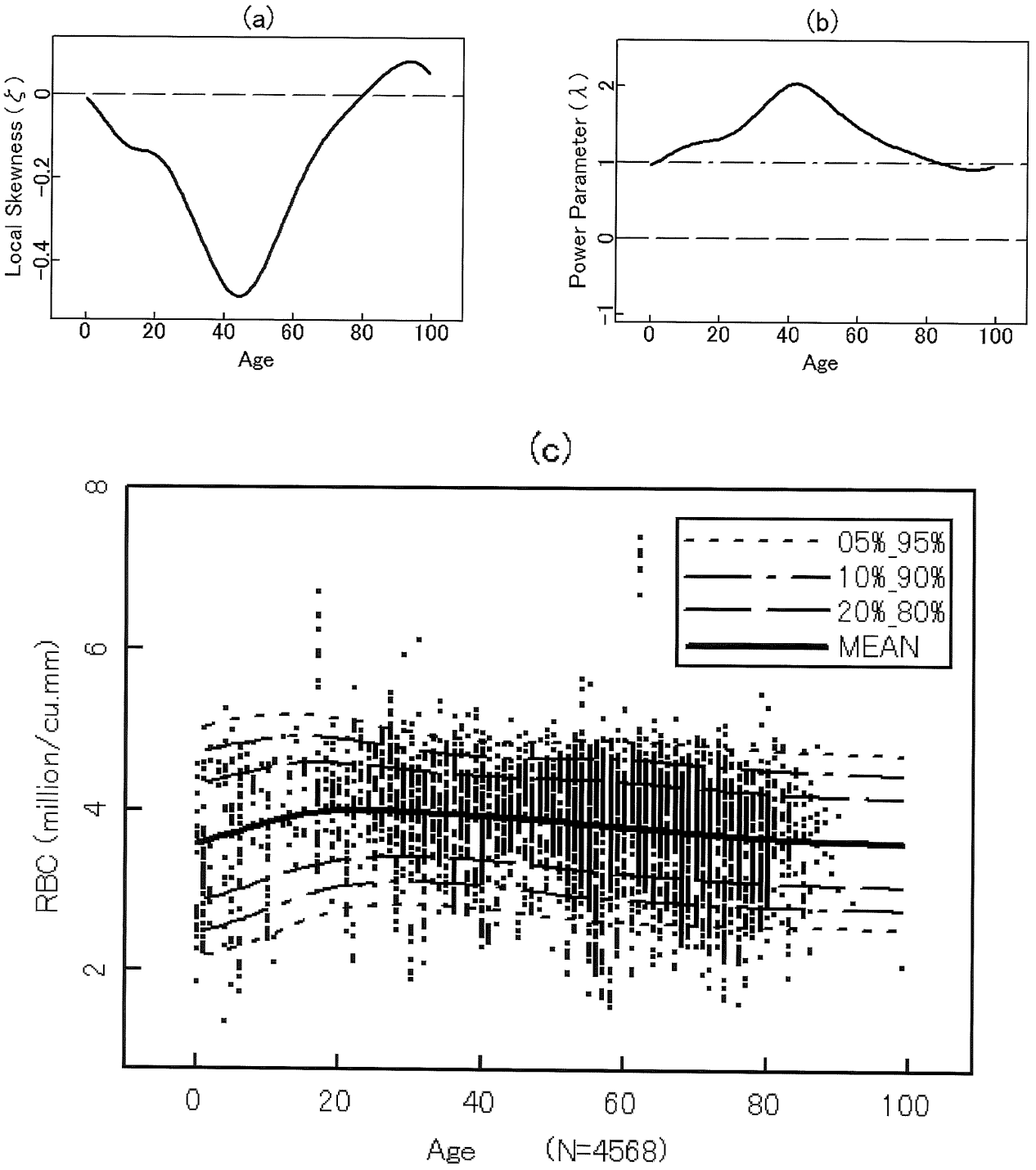

Fig. 2

(a) The smoothed local skewness of the distribution of Red Blood Cell (RBC) over age. The dash line shows the symmetry.

(b) The smoothed maximum likelihood estimates of power parameter over age. The dash dotted line shows the identical transformation and the dash line shows the logarithmic transformation.

(c) The solid curve shows the smoothed mean trend of the measurements over age. The dotted curves, the dash dotted curves and the dash curves show 5% and 95% points, 10% and 90% points and 20% and 80% points, respectively. The original measurements of RBC are superimposed for reference.

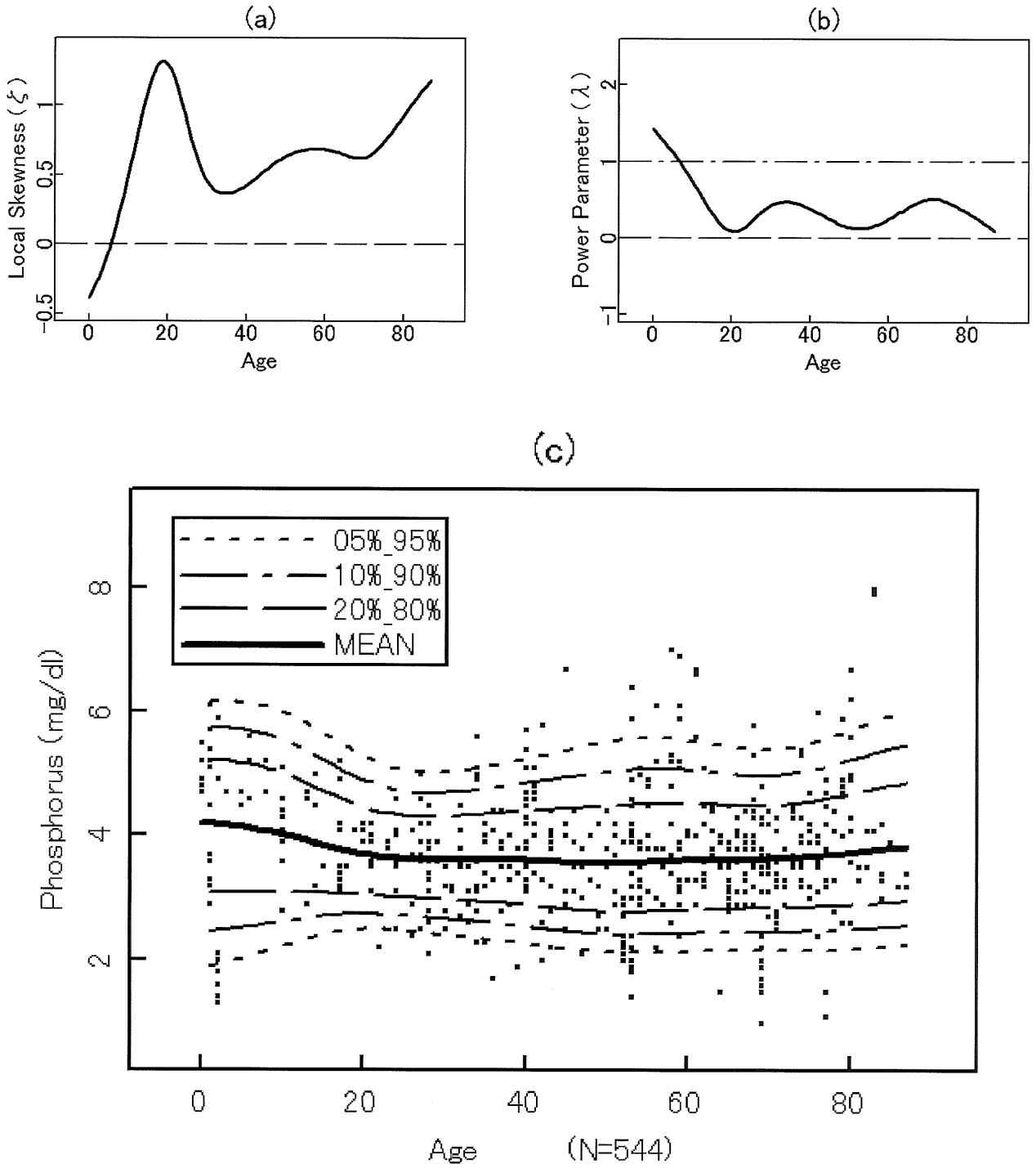

Fig. 3

(a) The smoothed local skewness of the distribution of Phosphorus (P) over age. The dash line shows the symmetry.

(b) The smoothed maximum likelihood estimates of power parameter over age. The dash dotted line shows the identical transformation and the dash line shows the logarithmic transformation.

(c) The solid curve shows the smoothed mean trend of the measurements over age. The dotted curves, the dash dotted curves and the dash curves show 5% and 95% points, 10% and 90% points and 20% and 80% points, respectively. The original measurements of P are superimposed for reference.

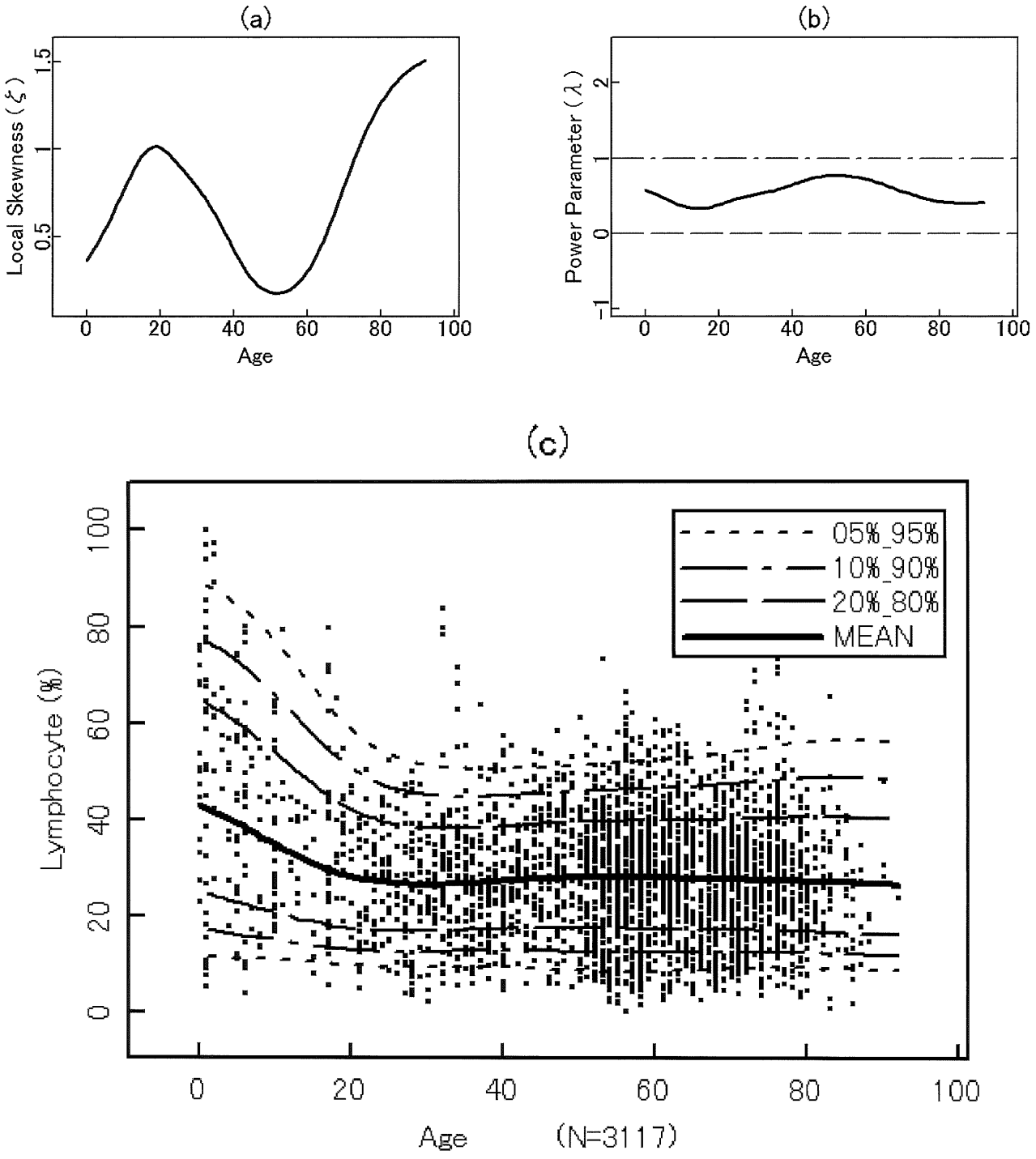

Fig. 4

(a) The smoothed local skewness of the distribution of Lymphocytes (LY) over age.

(b) The smoothed maximum likelihood estimates of power parameter over age. The dash dotted line shows the identical transformation and the dash line shows the logarithmic transformation.

(c) The solid curve shows the smoothed mean trend of measurements over age. The dotted curves, the dash dotted curves and the dash curves show 5% and 95% points, 10% and 90% points and 20% and 80% points, respectively. The original measurements of LY are superimposed for reference.

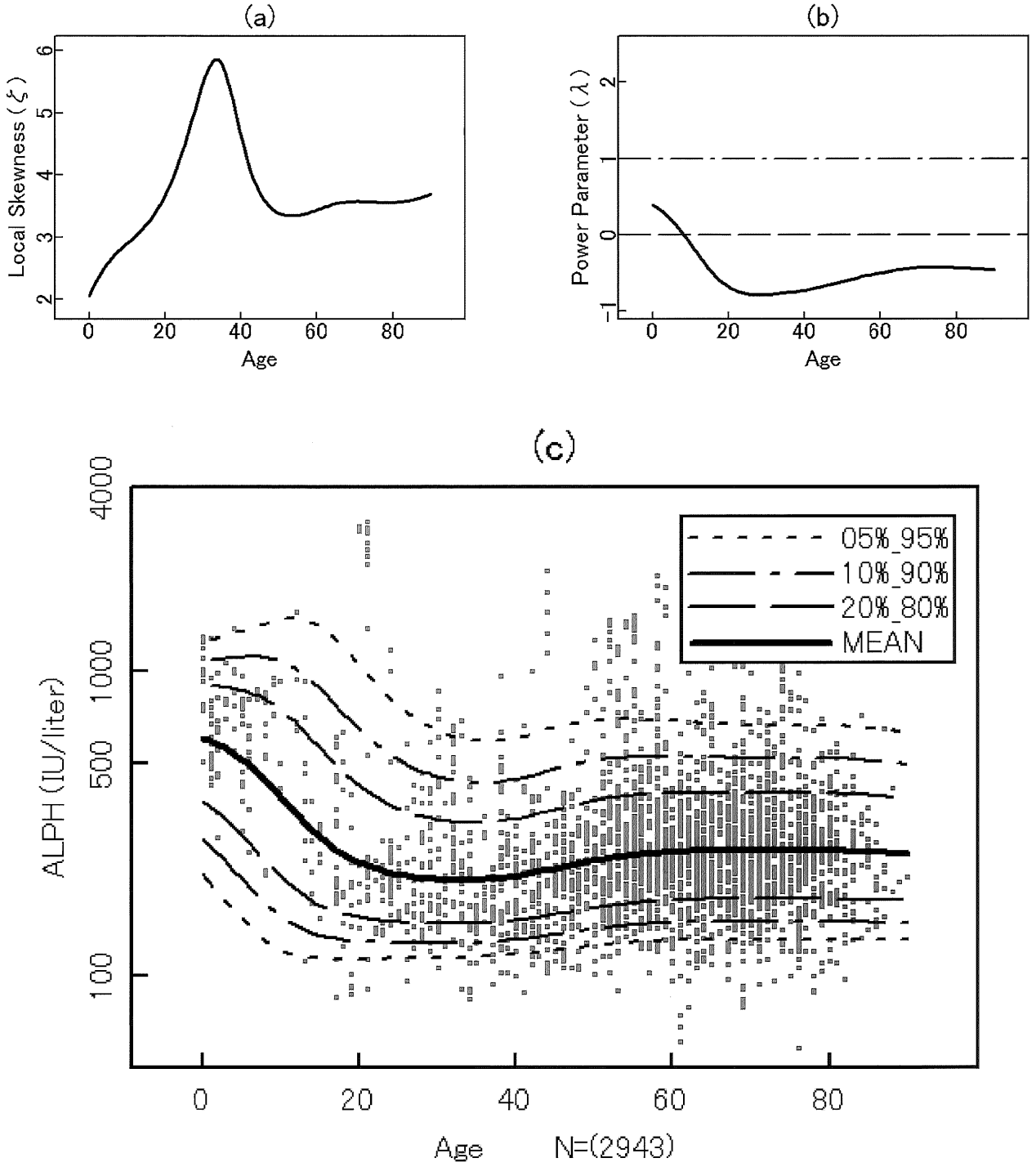

Fig. 5

(a) The smoothed local skewness of the distribution of Alkaline Phosphate (ALPH) over age.

(b) The smoothed maximum likelihood estimates of power parameter over age. The dash dotted line shows the identical transformation and the dash line shows the logarithmic transformation.

(c) The solid curve shows the smoothed mean trend of measurements over age. The dotted curves, the dash dotted curves and the dash curves show the 5% and 95% points, 10% and 90% points and 20% and 80% points, respectively. The original measurements of ALPH are superimposed for reference.

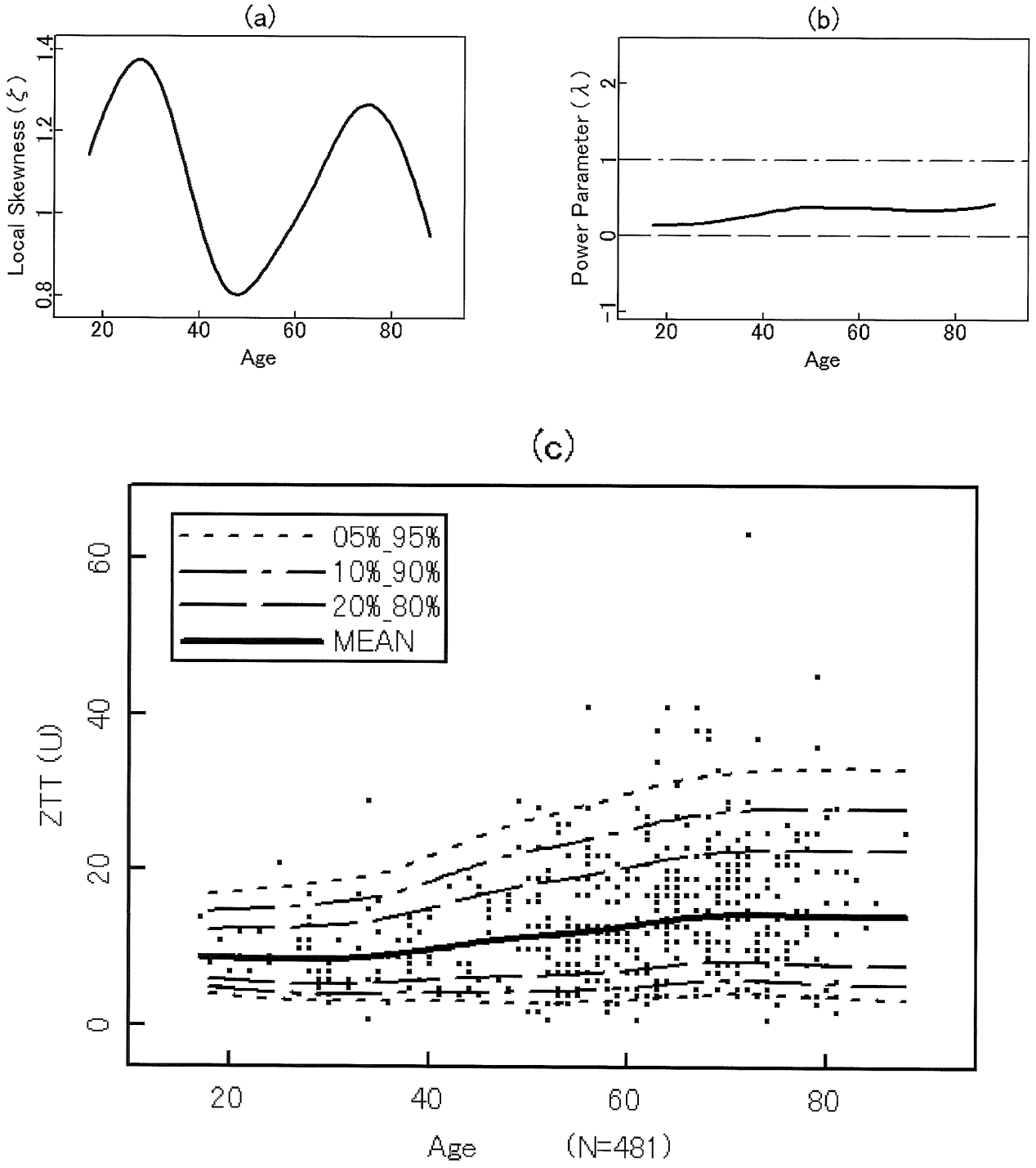

Fig. 6

(a) The smoothed local skewness of the distribution of Zinc Sulfate Turbidity (ZTT) over age.

(b) The smoothed maximum likelihood estimates of power parameter over age. The dash dotted line shows the identical transformation and the dash line shows the logarithmic transformation.

(c) The solid curve shows the smoothed mean trend of measurements over age. The dotted curves, the dash dotted curves and the dash curves show 5% and 95% points, 10% and 90% points and 20% and 80% points, respectively. The original measurements of ZTT are superimposed for reference.

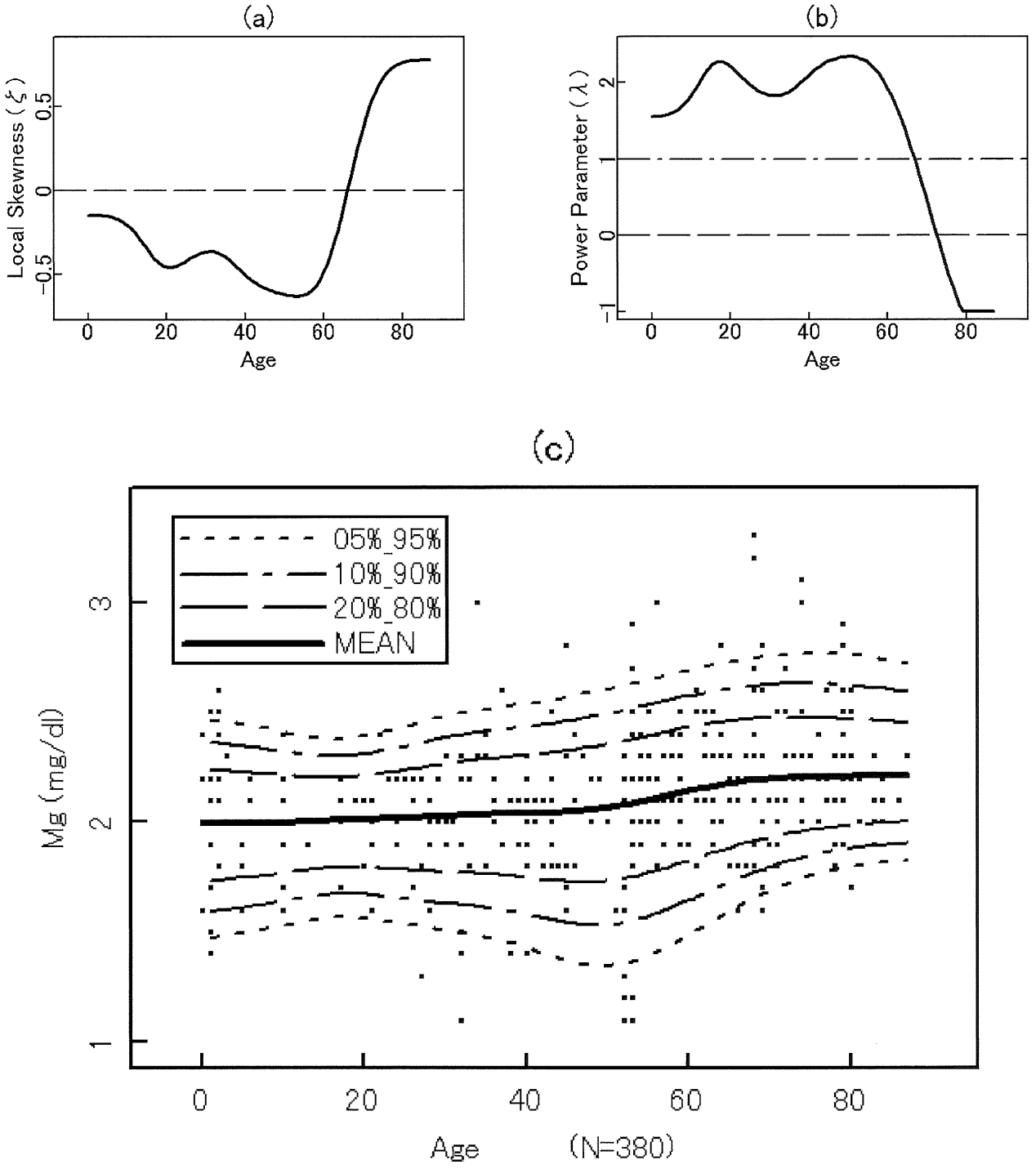

Fig. 7

(a) The smoothed local skewness of the distribution of Magnesium (Mg) over age. The dash line shows the symmetry.

(b) The smoothed maximum likelihood estimates of power parameter over age. The dash dotted line shows the identical transformation and the dash line shows the logarithmic transformation.

(c) The solid curve shows the smoothed mean trend of measurements over age. The dotted curves, the dash dotted curves and the dash curves show 5% and 95% points, 10% and 90% points and 20% and 80% points, respectively. The original measurements of Mg are superimposed for reference.

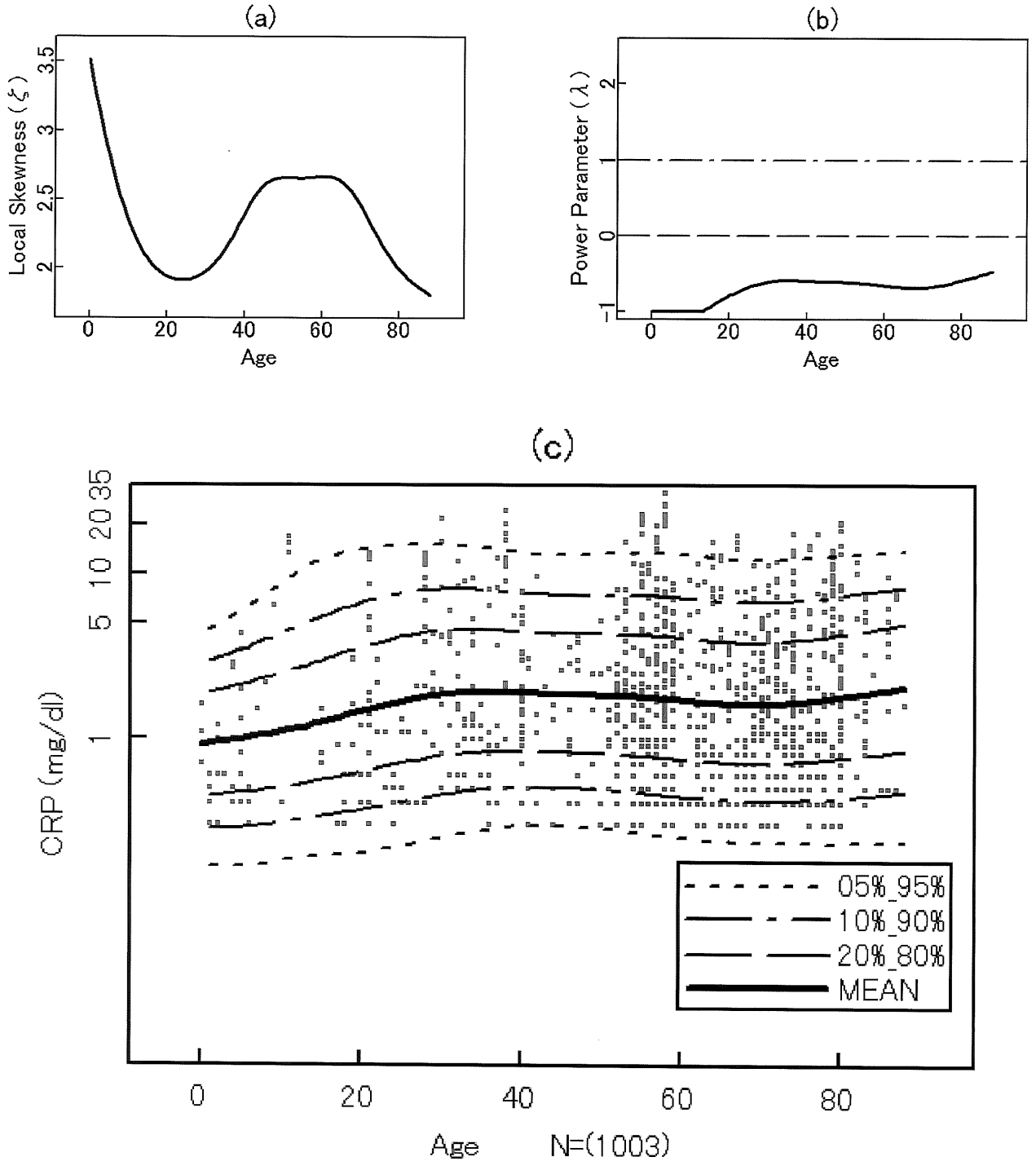

Fig. 8

(a) The smoothed local skewness of the distribution of C-Reactive Protein (CRP) over age.

(b) The smoothed maximum likelihood estimates of power parameter over age. The dash dotted line shows the identical transformation and the dash line shows the logarithmic transformation.

(c) The solid curve shows the smoothed mean trend of measurements over age. The dotted curves, the dash dotted curves and the dash curves show 5% and 95% points, 10% and 90% points and 20% and 80% points, respectively. The original measurements of CRP are superimposed for reference.

Phosphorus (P)

The results are shown in Fig. 3. The distribution is skewed to the right under the age of 10 and skewed to the left over the age of 10. The identical transformation is appropriate around the age of 10 years.

Lymphocyte (LY)

The results are shown in Fig. 4. The distribution is skewed to the left for almost all ages. It is takes the maximum value around the age of 20 and a minimum around the age of 50. The variance is relatively large in the younger age group. It decreases gradually to the age of 30 years.

Alkaline Phosphate (ALPH)

The results are shown in Fig. 5. The distribution is skewed to the left over age. The logarithmic transformation is appropriate around the age of 10 years. Large variance is observed relatively in the younger age group, which decreases gradually to the age of 30 years. The original scale of measurement is used in the X axis and the log scale is used in the Y axis in Fig. 5 (c).

Zinc Sulfate Turbidity (ZTT)

The results are shown in Fig. 6. The distribution is skewed to left over age. The mean trend and the variance gradually increase with age.

Magnesium (Mg)

The results are shown in Fig. 7. The distribution is skewed to left under the age of 65 years and skewed to the right over the age of 65 years. The identical transformation is available around the age of 65 years. The mean trend increases over the age of 50 years. The comparatively small difference is observed in mean or variance because these estimates are not sensitive to small and large observations of measurement. But the power parameter λ is sensitive to the small and large observations of measurement 7 (b). Another cause of difference may be the small sample size.

C-Reactive Protein (CRP)

The results are shown in Fig. 8. The distribution is skewed to the left. The variance increases with age by the age of 20 years. The original scale of measurement is used in the X axis and the log scale is used in the Y axis in Fig. 8 (c).

DISCUSSION

The proposed method is considered to be versatile for estimating the percent points of distribution of clinical measurements. However, it is also true that the result of estimation may depend on the bandwidth h of the nonparametric smoothing. In this work we adopted tentatively h="range of measurement"×0.1 through several trials and errors. The bandwidth value should be updated in future by a more suitable value for each measurement using some optimization criterion.

We also have to mention that "age" in this paper refers not only the "biological age" but the cohort effect[2)] in part. The changes among the clinical measurements during infancy, childhood and menopause or in elderly persons are known as the biological effect and the changes which occur due to the different time periods of birth or era are known as the cohort effect. Based on a cross-sectional study, it is not possible to separate the effect of the "biological age" and the cohort effect in clinical measurements data. Therefore, the estimates of the percent points of distribution for the measurements by the current method may not be applied directly to the measurements of individuals of other time period because the results are based on a cross-sectional study.

It should be noted that the measurements in these data were from healthy people as well as people who were unhealthy due to different diseases, so that the data may be heterogeneous. That's why, the results of the current analysis has no substantial meaning. If data from healthy individuals are available, our method would be more useful. In order to obtain more detailed and long-range applicable percent points of the clinical measurements, we need further analysis using some more specific longitudinal data consisting of repeated measurements per individual.

ACKNOWLEDGEMENTS

We are very grateful to the editor and anonymous referees, whose comments were invaluable for the revision and enrichment of this paper. We thank Mr. Hideyuki Itaba, Ms. Akemi Matsubara and Mr. Takashi Arase of the Department of Clinical Laboratory Medicine, Hiroshima University, for organizing the data. We also thank Dr. Tetsuji Tonda of the Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, for helping at the different stages of this research project.

(Received December 21, 2005)
(Accepted January 23, 2006)

REFERENCES

1. **Box, G.E. P. and Cox, D.R.** 1964. An analysis of transformations (With discussion). JRSS, B **26:** 211–252.
2. **Fienberg, S.E. and Mason, W.M.** 1978. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *In* K.F. Schuessler (ed.), Sociological Methodology 1979, London Jossey-Bass.
3. **Goto, M., Inoue, T. and Tsuchiya, Y.** 1987. Double power-normal transformation and its per-

formances: an extensive version of Box-Cox transformation. J. Jpn. Stat. Soci. **17:** 149–163.

4. **Goto, M.** 1992. Extensive views of power transformation: some recent developments. Proceedings of the Honolulu Conference on Computational Statistics as Memorial of the Fifth Anniversary of Japanese Society of Computational Statistics, Honolulu, December 1–5.
5. **Hamasaki, T., Isomura, T., Ohtaki, M. and Goto, M.** 1999. Power transformation model and its modifications. Jpn. J. App. Stat. **28:** 179–190.
6. **Hjort, N.L. and Jones, M.C.** 1996. Locally parametric nonparametric density estimation. Annals of Statistics **24:** 1619–1647.
7. **Nadaraya, E.A.** 1964. On estimating regression. Theory of probability and its application **10:** 186–190.
8. **Sakia, R.M.** 1992. The Box-Cox transformation technique: a review. Statistician **41:** 169–178.
9. **Seber, G.A.F. and Wild, C.J.** 1989. Nonlinear Regression. New York: John Wiley and Sons.
10. **Virtanen, A., Kairisto, V. and Uusipaikka, E.** 2004. Parametric methods for estimating covariate-dependent reference limits. J. Clin. Chem. Lab. Med. **42 (7):** 734–738.
11. **Watson, G.S.** 1964. Smooth regression analysis. Sankhya, Series A **26:** 101–116.