We evaluate the precision of $\hat{\theta}_{logit}$ using a confidence interval based on the delta method. Using a first order Taylor series approximation, the $\log_2 \hat{\theta}_{logit}$ can approximately be expanded as

$$\log_2 \hat{\theta}_{logit} = -\frac{\hat{\beta}_1}{\hat{\beta}_2} \approx -\frac{\beta_1}{\beta_2} - \frac{(\hat{\beta}_1 - \beta_1)}{\beta_2} + \frac{\beta_1(\hat{\beta}_2 - \beta_2)}{\beta_2^2}, \tag{9}$$

which yields an approximate variance of $\log_2 \hat{\theta}_{logit}$ as

$$\sigma^2(\log_2 \hat{\theta}_{logit}) \approx \frac{\mathrm{Var}(\hat{\beta}_1)}{\beta_2^2} + \frac{\beta_1^2 \mathrm{Var}(\hat{\beta}_2)}{\beta_2^4} - \frac{2\beta_1 \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)}{\beta_2^3}. \tag{10}$$

By obtaining estimates of the variance and covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ through linear approximation in the non-linear least squares method, an estimate of $\sigma^2(\log_2 \hat{\theta}_{logit})$ can be calculated by substituting $\hat{\beta}_1$ and $\hat{\beta}_2$ into (10) (Cox, 1990). We can obtain an approximate $1-\alpha$ confidence interval for $\log_2 \theta_{logit}$ as

$$\log_2 \hat{\theta}_{logit} \pm z_{1-\alpha/2}\ \hat{\sigma}(\log_2 \hat{\theta}_{logit}), \tag{11}$$

which yields the confidence interval for θ_{logit} as

$$\exp(\log_2 \hat{\theta}_{logit} \pm z_{1-\alpha/2}\ \hat{\sigma}(\log_2 \hat{\theta}_{logit})), \tag{12}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

A log-time regression method and a linear regression method

The models for a test substance and a negative control using a log-time regression method and a linear regression method are, respectively, described as

$$\begin{cases} y'_{ij} &= \beta_4 + \beta_5 \log_2 t_i + e_{ij}, \\ y'_{Cj} &= \beta_3 + e_{Cj}, \end{cases} \tag{13}$$

$$\begin{cases} y'_{ij} &= \beta_6 + \beta_7 t_i + e_{ij}, \\ y'_{Cj} &= \beta_3 + e_{Cj}. \end{cases} \tag{14}$$

Since the absorbance corresponding to ET50 is $\beta_3/2$, the $\log_2$ ET50 obtained from a log-time regression method, $\log_2 \theta_{log}$, and ET50 obtained from each method, θ_{log} and θ_{lin}, are defined by

$$\log_2 \theta_{log} = \frac{1}{\beta_5}\left(\frac{\beta_3}{2} - \beta_4\right), \tag{15}$$

$$\theta_{log} = 2^{(\beta_3/2-\beta_4)/\beta_5}, \tag{16}$$

$$\theta_{lin} = \frac{1}{\beta_7}\left(\frac{\beta_3}{2} - \beta_6\right). \tag{17}$$

We use the ordinary least squares method to estimate parameters, β_3, β_4, β_5, β_6 and β_7, in which Q_{log} and Q_{lin} defined as follows are minimized:

$$Q_{log} = \sum_i \sum_j \left(y'_{ij} - (\beta_4 + \beta_5 \log_2 t_i)\right)^2 + \sum_j \left(y'_{Cj} - \beta_3\right)^2, \tag{18}$$

$$Q_{lin} = \sum_i \sum_j \left(y'_{ij} - (\beta_6 + \beta_7 t_i)\right)^2 + \sum_j \left(y'_{Cj} - \beta_3\right)^2. \tag{19}$$

ET50 estimate are obtained by substituting acquired estimates, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_6$ and $\hat{\beta}_7$, into the definition of ET50 given by (16) and (17), respectively. Using a first order Taylor series approximation, the $\log_2 \hat{\theta}_{log}$ and $\hat{\theta}_{lin}$ can approximately be expanded as

$$\begin{aligned} \log_2 \hat{\theta}_{log} &= \frac{1}{\hat{\beta}_5}\left(\frac{\hat{\beta}_3}{2} - \hat{\beta}_4\right) \\ &\approx \frac{1}{\beta_5}\left(\frac{\beta_3}{2} - \beta_4\right) + \frac{(\hat{\beta}_3 - \beta_3)}{2\beta_5} - \frac{(\hat{\beta}_4 - \beta_4)}{\beta_5} - \frac{1}{\beta_5^2}\left(\frac{\beta_3}{2} - \beta_4\right)(\hat{\beta}_5 - \beta_5), \\ \hat{\theta}_{lin} &= \frac{1}{\hat{\beta}_7}\left(\frac{\hat{\beta}_3}{2} - \hat{\beta}_6\right) \end{aligned} \tag{20}$$

$$\approx \frac{1}{\beta_7}\left(\frac{\beta_3}{2}-\beta_6\right)+\frac{(\hat{\beta}_3-\beta_3)}{2\beta_7}-\frac{(\hat{\beta}_6-\beta_6)}{\beta_7}-\frac{1}{\beta_7^2}\left(\frac{\beta_3}{2}-\beta_6\right)(\hat{\beta}_7-\beta_7). \quad (21)$$

Then approximate variances of $\log_2\hat{\theta}_{log}$ and $\hat{\theta}_{lin}$ are given by

$$\begin{aligned}\sigma^2(\log_2\hat{\theta}_{log}) &\approx \frac{\mathrm{Var}(\hat{\beta}_3)}{4\beta_5^2}+\frac{\mathrm{Var}(\hat{\beta}_4)}{\beta_5^2}+\frac{\mathrm{Var}(\hat{\beta}_5)}{\beta_5^4}\left(\frac{\beta_3}{2}-\beta_4\right)^2 \\ &\quad +\frac{2\mathrm{Cov}(\hat{\beta}_4,\hat{\beta}_5)}{\beta_5^3}\left(\frac{\beta_3}{2}-\beta_4\right), \quad (22)\\ \sigma^2(\hat{\theta}_{lin}) &\approx \frac{\mathrm{Var}(\hat{\beta}_3)}{4\beta_7^2}+\frac{\mathrm{Var}(\hat{\beta}_6)}{\beta_7^2}+\frac{\mathrm{Var}(\hat{\beta}_7)}{\beta_7^4}\left(\frac{\beta_3}{2}-\beta_6\right)^2 \\ &\quad +\frac{2\mathrm{Cov}(\hat{\beta}_6,\hat{\beta}_7)}{\beta_7^3}\left(\frac{\beta_3}{2}-\beta_6\right). \quad (23)\end{aligned}$$

Estimates of $\sigma^2(\log_2\hat{\theta}_{log})$ and $\sigma^2(\hat{\theta}_{lin})$ can be calculated by substituting $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_6$ and $\hat{\beta}_7$, into (22) and (23). We can obtain an approximate $1-\alpha$ confidence interval for $\log_2\theta_{log}$ as

$$\log_2\hat{\theta}_{log} \pm z_{1-\alpha/2}\,\hat{\sigma}(\log_2\hat{\theta}_{log}), \quad (24)$$

which yields the confidence interval for θ_{log} as

$$\exp(\log_2\hat{\theta}_{log} \pm z_{1-\alpha/2}\,\hat{\sigma}(\log_2\hat{\theta}_{log})). \quad (25)$$

Similarly, we can obtain an approximate $1-\alpha$ confidence interval for θ_{lin} as

$$\hat{\theta}_{lin} \pm z_{1-\alpha/2}\,\hat{\sigma}(\hat{\theta}_{lin}). \quad (26)$$

A two-stage method

Parameter estimates and/or variance covariance matrices occasionally cannot be obtained from the logistic regression method due to the small sample sizes. We consider a two-

stage method in which the log-time regression method is alternatively applied if the logistic regression method cannot construct a confidence interval for ET50.

Design of the simulation study

We evaluate the performance of each estimation method through a Monte-Carlo simulation involving the following steps under the similar conditions to the validation study for TESTSKIN^{TM} (2002) and Vitrolife-Skin^{TM} (2004).

Step1. Specify true ET50 value between 4 and 18 hours assuming a mild test substance. The time point for measurement is allocated as $(t_1, t_2, t_3, t_4) = (2, 4, 8, 16)$ in 4-point design and is allocated as $(t_1, t_2, t_3, t_4, t_5) = (2, 4, 8, 16, 24)$ in 5-point design.

Step2. Generate virtual data for a test substance from the logistic curve on the time-response defined by

$$y'_{ij} = \frac{\beta_3}{1 + \exp(\beta_1 + \beta_2 \log_2 t_i)} + e_{ij}, \tag{27}$$

where $\beta_2 = 2.0$, $\beta_3 = 1.0$ and e_{ij} is mutually distributed as a normal distribution $N(0,\ 0.1^2)$. Since the ET50 value, θ_{logit}, is a function of β_1 and β_2, β_1 is determined from β_2 and θ_{logit}. Figure 1 shows the assumed time-response curves of model (27).

Step3. Generate virtual data for a negative control from $y'_{Cj} = \beta_3 + e_{Cj}$, where $\beta_3 = 1.0$ and e_{Cj} is mutually distributed as a normal distribution $N(0,\ 0.1^2)$.

Step4. Estimate ET50 and construct a confidence interval for ET50 using each estimation method.

Step5. Iterate 10,000 times Step 2 through Step 4, and calculate a proportion of estimable cases, a bias in estimates, and a coverage probability in which each interval contains the true ET50 values. A bias in estimates is defined as the median of the difference of the estimate of ET50 and the true ET50 value.

The reason we assume a mild test substance in Step 1 is that estimating ET50 for clearly strong or weak substances is not essential.

Results

We report the results of the Monte-Carlo simulation study in Tables 1 through 3 and present the corresponding scatter plots in Figures 2 through 4 to compare the performance of each estimation method. In these tables and figures, the left side shows the results in 4-point design and the right side shows those in 5-point design.

The characteristics of each estimation method are summarized below.

A logistic regression method

- The proportion of estimable cases decreases as low as 85% with the increase of true ET50 values in 4-point design, whereas it is almost 100% in 5-point design.
- The bias in estimates is negligible in both 4- and 5-point designs.

- The coverage probability is always below the nominal confidence level of 95% and as low as 88% in some cases.

According to the above mentioned results, the logistic regression method is appropriate in 5-point design, whereas another method should complementarily be used in addition to the logistic regression method in 4-point design.

Examples of simulated data that yielded feasible and unfeasible estimates, when the true ET50 value is 14 hours, are shown in Figure 5 together with the true and estimated time-response curves. This figure suggests that we tend to encounter difficulty in obtaining confidence intervals when we do not have measurements on time points around ET50.

A log-time regression method

- The proportion of estimable cases is almost 100% in both 4- and 5-point designs.
- ET50 estimates tend to be greater than the true ET50 value with the increase of true ET50 values in 4-point design.
- The bias in estimates is within 2 hours and, therefore, negligible in 5-point design.
- Although the coverage probability tends to be greater with the increase of true ET50 values in both 4- and 5-point designs, the discrepancy from the nominal confidence level of 95% is within $\pm 5\%$.

These results suggest caution in rarely adopting 4-point design because estimates tend to be great when the true ET50 value is great, although no remarkable defects appear in 5-point design.

A linear regression method

- The proportion of estimable cases is almost 100% in both 4- and 5-point designs.
- The estimates of ET50 tend to be great in both 4- and 5-point designs, when the true ET50 value is small.
- The coverage probability is seriously low since it is as low as 50% in 4-point design or 40% in 5-point design in the worst cases.

These results suggest that the linear regression method should not be adopted due to low coverage probabilities irrespective of design. The 5-point design is more disadvantageous than the 4-point design because the time points for obtaining measurements in 5-point design included $T_5 = 24$ hour in our simulation setting. Actually, the measurement at the 24 hour point leads to a smaller value of the gradient than the expected value.

Two-stage method

- The proportion of estimable cases is almost 100% in both 4- and 5-point designs.
- The estimates of ET50 are on average almost the same as the true ET50 value in both 4 and 5-point designs.
- Although the coverage probability tends to be greater with the increase of true ET50 values in both 4- and 5-point designs, the discrepancy from the nominal confidence level of 95% is within $\pm 5\%$.

These results indicate that the two-stage method is reasonable for obtaining a confidence interval for ET50, although it should be slightly adjusted so as to keep the coverage probability near the nominal confidence level.

Discussion

We recommend using the two-stage method for obtaining a confidence interval for ET50. However, further investigations are necessary to extend the conclusion to any case of the design and analysis of experiments using 3D skin models, since the adopted simulation conditions are adaptable only for the real validation studies of TESTSKIN and Vitrolife-Skin. When the use of refined statistical software such as SAS or R is difficult, we recommend using the log-time regression method with 5-point design although the biased estimates within 2 hours are occasionally obtained.

The condition where the proportion of estimable cases in the application of logistic regression method in 4-point design realizes values below 100% depends on the number of time points, the positioning of time points, and the scale of measurement errors (Sozu et al., 2005, Sozu et al., 2006). Properly setting these conditions considering the convenience of workers is important and further studies are necessary to address this issue.

The results of this research would promote the use of 3D skin models through the achievement of adequate and quantitative evaluations of skin irritation of test substances.

Acknowledgment

The authors would like to thank the editor and two anonymous referees for helpful comments that greatly improved this article. This research was supported in part by a grant from the Japanese Society of Alternative to Animal Experiments.

References

Botham, P. M. (2004) The validation of in vitro methods for skin irritation, *Toxicol. Letters*, 149, 387–390.

Cox, C. (1990) Fieller's theorem, the likelihood and the delta method. *Biometrics*, 46: 709–718.

Kojima, H. (2005a) Present and future of alternative to skin irritation testing in Japan, *Fragrance Journal*, 33(2), 43–52 (in Japanese).

Kojima, H., Shiraishi, A., Andoh, Y., Okazaki, Y., Ozawa, N., Kawabata, R., Kadono, K., Sozu, T., Suzuki, T., Tabawa, A., Nakano, H., Morikawa, N., Hori, M., Yamashita, K., and Yoshimura, I. (2005b) Validation study for Vitrolife-Skin^{TM}, a three-dimensional cultured human skin model, I, as an alternative to skin irritation testing using ET50 protocol, *ALTEX*, 22(Spl), 160.

Kojima, H., Shiraishi, A., Andoh, Y., Okazaki, Y., Ozawa, N., Kawabata, R., Kadono, K., Sozu, T., Suzuki, T., Tabawa, A., Nakano, H., Morikawa, N., Hori, M., Yamashita, K., and Yoshimura, I. (2005c) Validation study for Vitrolife-Skin^{TM}, a three-dimensional cultured human skin model, II, as an alternative to skin irritation testing using post-incubation (PI) protocol, *ALTEX*, 22(Spl), 161.

Morikawa, N., Morota, K., Morita, S., Kojima, H., Nakata, S., and Konishi, H. (2002) Prediction of human skin irritancy using a cultured human skin model: comparison of chemical application procedures and development of a novel chemical application procedure using the Vitrolife-Skin™ model, *AATEX*, 9(1), 1–10.

Morikawa, N., Morota, K., Suzuki, M., Kojima, H., Nakata, S., and Konishi, H. (2005) Experimental study on a novel chemical application procedure for in vitro skin corrosivity testing using the Vitrolife-Skin™ human skin model, *AATEX*, 11(1), 68–78.

Morota, K., Morikawa, N., Morita, S., kojima, H., and Konishi, H. (1998) Development and evaluation of the cultured skin model, *Tiss. Cult. Res. Commun.*, 17, 87–93.

Omori, T., Saijo, K., Kato, M., Itagaki, H., Hayashi, M., Miyazaki, S., Ohno, T., Sugawara, H., Teramoto, N., Tanaka, N., Wakuri, S., and Yoshimura, I. (1998) Validation study on five cytotoxicity assays by JSAAE II Statistical analysis, *AATEX*, 5, 39–58.

Shiraishi, A., Hyodo, Y., Sozu, T., Hamada, C., and Yoshimura, I. (2005) A statistical method for estimating ET50 under the condition of small volume of data. *ALTEX*, 22(Spl), 166.

Shiraishi, A., Hyodo, Y., Sozu, T., Hamada, C., and Yoshimura, I. (2006) A statistical method for estimating ET50 using small size data. *AATEX*, 11(Spl), 269.

Sonoda, I., Kojima, H., Sato, A., Terasawa, M., Goda, M., Hori, M., Okamoto, H., Mizuno, M., Imai, N., Takei, M., Uetake, N., Goto, M., Kawabata, R., Sasaki, Y., Ukawa, K., Ozawa, N., Suzuki, T., Usami, M., Kasahara, T., Goto, K., Torishima, H., Takahashi,

H., Ishibashi, T., Morikawa, N., and Yoshimura, I. (2002) A prevalidation study for three-dimensional cultured human skin models as alternatives to skin irritation testing. *AATEX*, 8(3-4), 91–106.

Sozu, T., Takanuma, M., Shiraishi, A., Hamada, C., and Yoshimura, I. (2005) Statistical considerations for positioning time points in ET50 estimation using three dimensional human skin model, *ALTEX*, 22(Spl), 166.

Sozu, T., Takanuma, M., Shiraishi, A., Hamada, C., and Yoshimura, I. (2006) Statistical considerations for allocation and the number of time points in a ET50 estimation using three dimensional human skin model, *AATEX*, 11(Spl), 268.

Tables and Figures

Table 1: Results of proportion of estimable cases (%) for each estimation method.

	4-point design				5-point design			
	Estimation method				Estimation method			
True ET50 values	Logistic	Log-time	Linear	Two-stage	Logistic	Log-time	Linear	Two-stage
4	99.2	100.0	100.0	100.0	99.1	100.0	100.0	100.0
5	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0
6	99.8	100.0	100.0	100.0	99.7	100.0	100.0	100.0
7	99.0	100.0	100.0	100.0	99.0	100.0	100.0	100.0
8	99.1	100.0	100.0	100.0	99.3	100.0	100.0	100.0
9	99.8	100.0	100.0	100.0	99.8	100.0	100.0	100.0
10	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
11	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
12	99.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0
13	98.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
14	97.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
15	94.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
16	92.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
17	88.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0
18	85.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 2: Results of bias in estimates for each estimation method.

True ET50 values	4-point design				5-point design			
	Estimation method				Estimation method			
	Logistic	Log-time	Linear	Two-stage	Logistic	Log-time	Linear	Two-stage
4	−0.01	0.32	1.41	0.00	−0.01	0.46	1.16	0.00
5	−0.02	0.20	1.88	−0.02	−0.01	0.51	2.55	−0.01
6	−0.01	−0.02	1.90	−0.01	0.00	0.34	3.06	0.00
7	−0.01	−0.27	1.71	−0.01	0.00	0.06	3.15	0.01
8	−0.02	−0.48	1.44	−0.02	−0.01	−0.26	3.03	−0.01
9	−0.05	−0.59	1.15	−0.05	−0.02	−0.60	2.78	−0.02
10	−0.06	−0.54	0.88	−0.06	−0.02	−0.90	2.47	−0.02
11	−0.05	−0.27	0.67	−0.05	−0.02	−1.15	2.15	−0.02
12	−0.04	0.31	0.48	−0.03	−0.02	−1.32	1.82	−0.02
13	−0.02	1.30	0.39	−0.01	−0.01	−1.39	1.50	−0.01
14	−0.04	2.84	0.39	0.01	−0.01	−1.31	1.22	−0.01
15	−0.08	5.18	0.49	0.02	−0.02	−1.07	0.97	−0.02
16	−0.15	8.67	0.71	0.02	−0.02	−0.57	0.78	−0.02
17	−0.22	13.80	1.05	0.08	−0.03	0.20	0.64	−0.03
18	−0.22	21.38	1.55	0.31	−0.05	1.32	0.55	−0.05

Table 3: Results of coverage probability for each estimation method.

True ET50 values	4-point design				5-point design			
	Estimation method				Estimation method			
	Logistic	Log-time	Linear	Two-stage	Logistic	Log-time	Linear	Two-stage
4	88.1	90.6	93.2	88.1	86.5	91.3	98.7	86.6
5	88.4	91.3	71.6	88.4	87.0	88.0	86.1	87.0
6	89.0	92.5	56.2	89.0	87.9	90.6	63.7	87.9
7	88.9	91.7	53.1	89.0	88.0	93.3	46.8	88.1
8	90.1	91.6	59.6	90.2	89.5	92.4	40.1	89.6
9	90.0	92.9	70.7	90.0	90.1	90.6	40.1	90.1
10	89.9	95.1	81.7	89.9	90.4	89.4	45.8	90.4
11	89.7	96.9	88.2	89.7	90.5	89.3	54.2	90.5
12	90.0	98.2	91.6	90.0	90.4	90.2	64.9	90.4
13	90.3	98.6	93.2	90.4	89.8	91.9	74.3	89.8
14	91.4	98.8	93.9	91.7	89.1	94.0	81.7	89.1
15	93.2	98.4	94.2	93.6	89.6	96.1	87.5	89.6
16	94.1	97.8	94.9	94.5	90.6	97.7	90.4	90.6
17	93.2	96.7	95.2	94.0	91.2	98.5	92.1	91.3
18	91.5	95.4	95.7	92.7	92.0	99.2	93.2	92.0

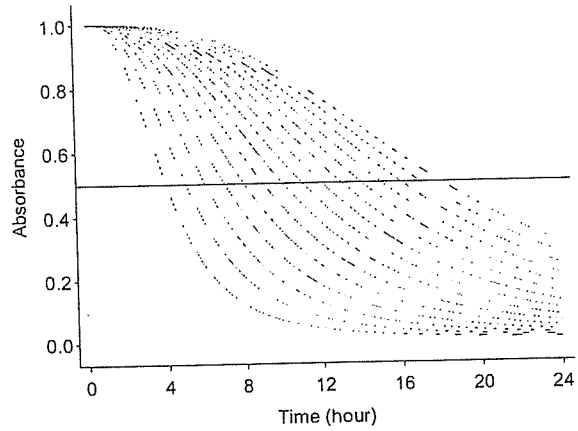
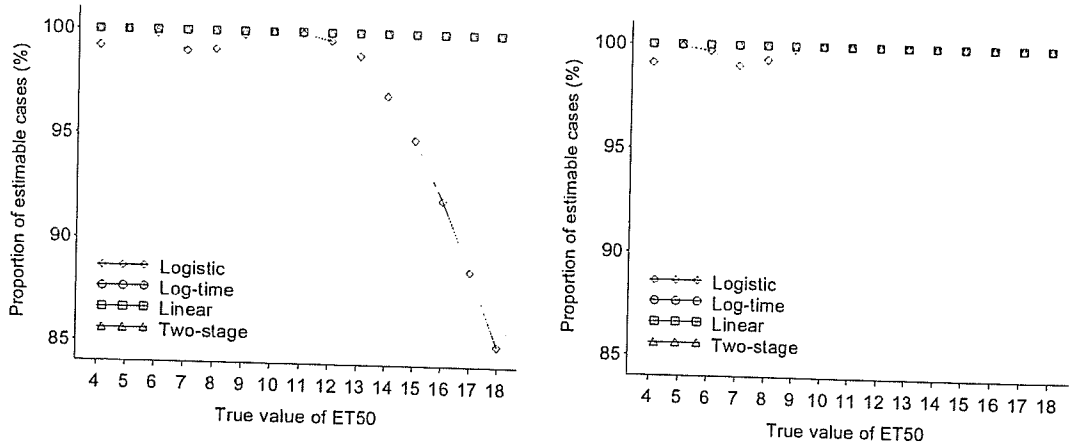
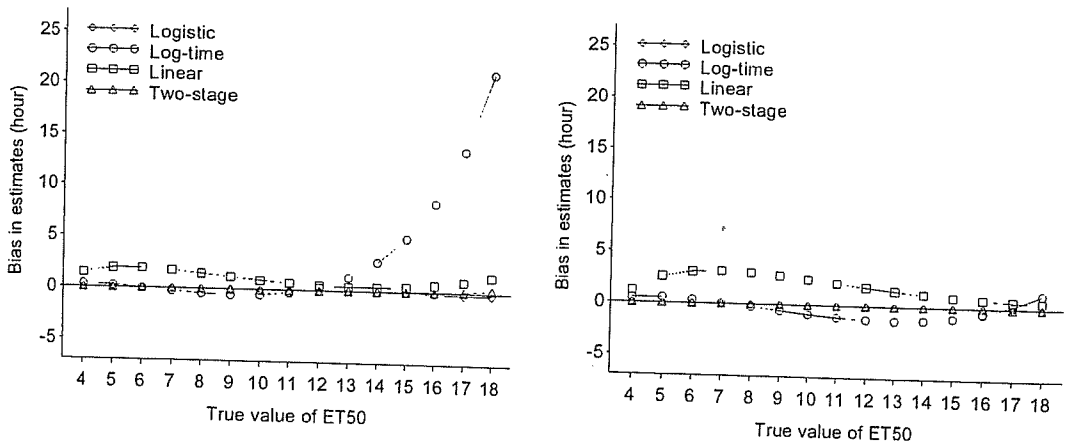


Fig. 1: Assumed time-response curves based on the logistic regression model.



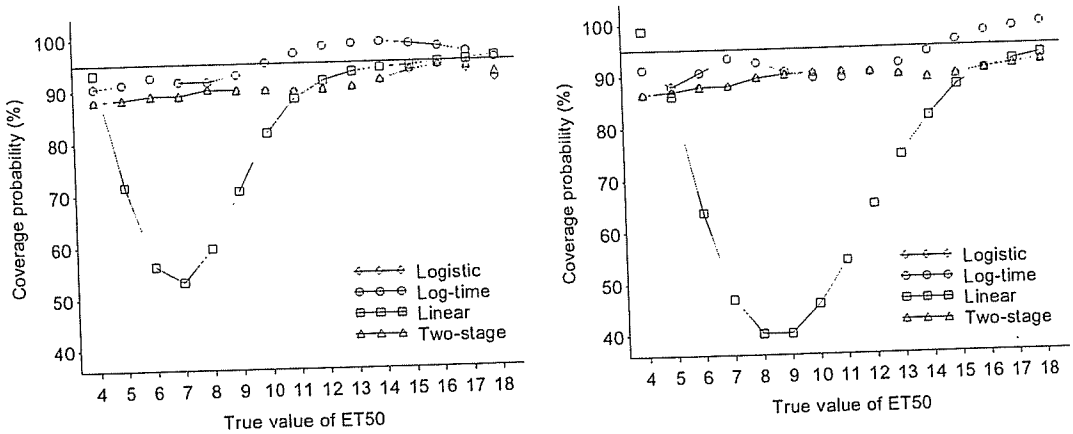
4-point design　　5-point design

Fig. 2: Results of proportion of estimable cases (%) for each estimation method.



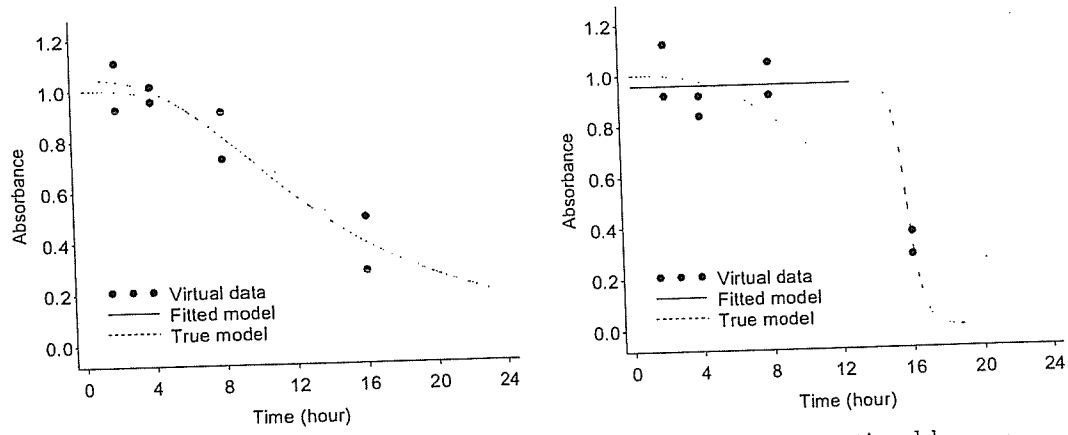
4-point design　　5-point design

Fig. 3: Results of bias in estimates for each estimation method.



4-point design　　5-point design

Fig. 4: Results of coverage probability for each estimation method.



Example of estimable case　　Example of not estimable case

Fig. 5: Examples of simulated data for a test substance with the true and estimated time response curves.

Omori T, AATEX 12(1), 25-31, 2006

ORIGINAL ARTICLE

A Measure Evaluating Relevance of a Validation Study of Alternatives to Animal Testing

Takashi Omori

Department of Biostatistics, Kyoto University, Japan

Abstract

Sensitivity, specificity and accuracy are well known measures for evaluating the relevance of an inter-laboratory validation study for alternative tests. It is not generally discussed that the measures are dependent on two determining factors: a set of chemicals and the number of laboratories. Furthermore, some alternative tests such as these for the phototoxicity test have an "Equivocal" category for judging the toxicity of chemicals. These facts have made it difficult to interpret the value of the measures.

Therefore, in this paper we propose new measures to evaluate the alternatives, which depend on a set of chemicals rather than on both factors, and can treat data which have "Equivocal" category. We also propose their confidence intervals, which are measures of their precision.

Key words: relevance, inter-laboratory validation study, sensitivity, specificity, accuracy, confidence interval

Introduction

Recently, due to an increasing social concern for animal welfare, a lot of alternative animal tests have been proposed, and in order to examine their feasibility and practicality various inter-laboratory validation studies have been conducted (e.g. Ray et al., 1994; Spielmann et al., 1998). Generally, the primary purpose of the validation study is to evaluate both the relevance and reliability of a proposed alternative test from the results of experiments using the alternative test (Balls et al., 1999). Sensitivity, specificity and accuracy are measures to determine the effectiveness of the alternative test when both the alternative and the animal tests have a binary classification for judging toxicity of chemicals, as "Positive" and "Negative". These are well known measures which have been widely used to evaluate the relevance of the alternative test in many validation studies (e.g. Balls et al., 1990; Roy et al., 1994; Spielmann et al., 1998).

However, two points should be taken into consideration concerning the interpretation of the summarized data from validation studies. The first point is that the values in the 2 by 2 table, which summarizes data, depend not only on a selected set of chemicals in the study but also on the number of participant laboratories. The other point is that a category for "Equivocal" produced from some alternative tests such as these for the photoxicity test, which is neither a "Positive" nor "Negative" category, is often provided. For instance, the test guideline of the in vitro 3T3 NRU phototoxicity test states that *'a test substance with a PIF < 2 or an MPE < 0.1 predicts: "no phototoxicity". A PIF >2 and < 5 or an MPE > 0.1 and < 0.15 predicts: "probable phototoxicity" and a PIF > 5 or an MPE > 0.15 predicts: "phototoxicity".'* where the PIF and the MPE are measurements of phototoxicity for the test (OECD, 2004). In this case, since there was a range suggesting similar performance when several cut-off points were examined, the

category "probable phototoxicity" as "Equivocal" was set (Peters and Holzhütter, 2002). Sugiyama, et al (1994) proposed a red blood cell hemolysis assay to predict phototoxicity of chemicals, and they classified photohemolysis into three categories, +, ± and -.

In this paper, we discuss the above two points for the measures, sensitivity, specificity and accuracy, and propose new measures for evaluating the relevance of an inter-laboratory validation study. We also construct an equation for their confidence intervals, which measure their precision of them (Altman, 2000a).

Methods

Definition for sensitivity, specificity and concordance

Table 1 shows a 2 by 2 table. Sensitivity is defined as the proportion of chemicals judged as positive by an alternative test in which the chemicals are identified as positive by an animal test. When data is summarized as in table 1, sensitivity is calculated by a / (a + b). Specificity is defined as the proportion of chemicals judged as negative by the alternative test in which the chemicals are identified as negative by the animal test. The measure is d / (c + d). Accuracy is defined as the proportion of a corresponding number of chemicals by the judgment of the alternative test in which all the chemicals are identified by the animal test. The measure is obtained as (a + d) / (a + b + c + d).

It is rarely noted that the values of these measures depend on the selected set of chemicals. If the toxicity of the selected chemicals in a validation study has only the strongest classes and the weakest classes, the values of these measures would be expected to be higher when the assessed alternative test has a good correlation to the targeted animal test. If the researchers conducting the validation study can select test chemicals before the experiments on the alternative test, they can control the measures. On the other hand, if they choose many middle class chemicals in the study, the measures may show an inferior result compared to our expectation. Even if the chemicals are selected by an external person not directly involved in the study, the values of these are dependent on the selected chemicals. Thus, we should interpret the values of these as conditional proportions dependent on the set of selected chemicals in the study.

Motivated data

Table 2 shows a typical form of data from a validation study. The symbols "P", "E" and "N" in the Table mean "Positive", "Equivocal" and "Negative" to be judged by *In vivo* test or the al-

Table 1. The 2 by 2 table.

		Animal test	
		Positive	Negative
Alternative test	Positive	a	c
	Negative	b	d
		a+b	c+d

Table 2. A motivated example of a inter-laboratory validation study.

Chemical	In vivo	Laboratory					
		a	b	c	d	e	f
A	P	P	E	E	P		
B	P	P	N	P	E		
C	N	P	P	P	P		
D	P	P	E			E	P
E	N	P	P			P	P
F	N	N	P			N	N
G	P			P	P	P	P
H	N			N	E	E	E
I	N			N	N	N	N

Symbols: P, positive; E, equivocal; N, negative

ternative test. This data is from an actual validation study conducted in Japan which has not been published yet. In the study, nine chemicals were tested by six laboratories. In order to meet an increasing demand for assessing test chemicals, the laboratories used the alternative test for as many chemicals as possible. However, due to time and financial constraints, all the laboratories did not experiment applying the alternative test for all the chemicals. In view of animal welfare, data from animal tests is usually obtained from some published articles and/or databases including data from past experiments; animal tests are rarely conducted in validation studies. Therefore there is usually only one result for each chemical. On the other hand, some results for each chemical in an alternative test are obtained from the inter-laboratory study.

When the measures, sensitivity, specificity and accuracy, are calculated, data, as in Table 2, is summarized by a 2 by 2 table, in which a result from a chemical in a laboratory for an alternative test corresponds to a result from using the same chemical in an animal test; total for four cells in the 2 by 2 table is 36 as is the case in Table.

Consideration of two points

Furthermore, in addition to the fact that the measures are a conditional proportion of a set of chemicals, we also have to consider that these depend on the number of laboratories conducting inter-laboratory validation studies. However, when data is summarized by a 2 by 2 table, as in Table 2, distinguishing between the two factors, the set of selected chemicals and the number of participant laboratories is overlooked. Then the interpretation of the value is difficult. For instance, the sensitivity from a laboratory which has examined ten positive chemicals is 100% when all the chemicals are judged positive. The sensitivity from the ten laboratories which examined a positive chemical is also 100% when all laboratories judge positive for the chemical. Should we regard both sensitivities as the same? Some people often use only the values of these measures from different validation studies without taking into consideration these factors, when they compare the alternatives.

The presence of an "Equivocal" category is another difficulty involved in interpreting the measures. Since these measures are based on the assumption that the results of both tests are expressed as binary categories, often data for "Equivocal" is artificially changed: these are eliminated from the numerator; data for "Equivocal" is relabeled as "Positive" (e.g. Sugiyama et al., 1994). The value of the measures depends on which treatment is used.

Proposed methods

We propose similar measures to sensitivity, specificity and accuracy, which take into consideration and deal with the previous two points.

Firstly, we consider the relationship between two factors; chemical and laboratory. Since several laboratories experiment using the alternative test for a same chemical in the inter-laboratory validation study, data from the validation study has a hierarchical structure between two factors. In the proposed methods, the factor of chemical becomes a basic unit.

Suppose y_{ij} is a variable to explain the result from an alternative test, and x_i is a variable to explain the result from an animal test, where subscript i and j mean the i th chemical (i = 1, 2, ..., n) and the j th laboratory (j = 1, 2, ..., m_i) respectively. The variable y_{ij} take 1 for the "Positive" result, 0 for the "Negative" and 0.5 for the "Equivocal", when the alternative test is experimented for the i th chemical in the j th laboratory. The variable x_i is 1 for the "Positive" result of the targeted animal test, and 0 for the "Negative" result. We initially define p_i as a proportion for the number of positive results in the i th chemical for the alternative test, that is

$$p_i = \sum_j y_{ij} / m_i . \qquad (1)$$

As shown the appendix A, we can calculate the variance, $V(p_i)$, based on the assumption of trinomial distribution.

Using p_i, we also define q_i as

$$q_i = x_i p_i + (1 - x_i)(1 - p_i). \qquad (2)$$

Note that q_i is a measure for the reliability of the i th chemical. The alternative test shows good reliability when the value of q_i is close to 1.

Finally, we define three measures which correspond to sensitivity, specificity and accuracy, using p_i, and call these measures *Psn*, *Psp* and *Pac*, respectively;

$$Psn = \sum_i x_i p_i / \sum_i x_i \qquad (3)$$