$\gamma = K/(m_1 + 1)$ and let G_1 be the cumulative distribution function of g_1. Under the similar conditions of Theorem 5 described in Lin *and others* (2006), the approximation of $P_j(K)$ is obtained as

$$P*_j(K) = \Pr(\theta_j \geqslant \overline{G}_1^{-1}(\gamma)|Z_{1j} = 1, \mathbf{y})\Pr(Z_{1j} = 1|\mathbf{y}), \tag{3.4}$$

where

$$\overline{G}_1(t) = \frac{\sum_{j=1}^{m} \Pr(\theta_j \leqslant t|z_{1j} = 1, y_j)\Pr(Z_{1j} = 1|y_j)}{\sum_{j=1}^{m} \Pr(Z_{1j} = 1|y_j)}.$$

$P*_j(K)$ corresponds to the tail-area posterior probability (TPP) of θ_j. Proof of this approximation is provided in the supplementary material available at *Biostatistics* online. Here, we denote this rule as the "TPP" method. The quantity m_1 can be replaced by its estimator $\sum_j \Pr(Z_{1j} = 1|y_j)$ (McLachlan *and others*, 2004).

4. Simulation studies

We conducted a series of simulation studies to assess the performance of our proposed methods. Details of the simulations are presented in Section A of the supplementary material available at *Biostatistics* online.

In summary, the sensitivity and root mean squared error (RMSE) of all the proposed methods were better than those of the other methods. As was expected, the TPP method had the greatest sensitivity, the RPM method had the lowest RMSE values, and the posterior probability of differentially expressed (PPDE) (3.1) ranking had the lowest false-positive rate. The posterior mean under unimodal hierarchical model (PM_U) method, without invoking mixture models, had a lower sensitivity and a larger RMSE, compared with the PM method. The fold change had comparable sensitivity and RMSE values for large sample sizes, but not for small sample sizes. The fold change had very large false-positive rates even when the sample size was large because it does not guard against selecting null genes.

5. Application to a breast cancer study

We illustrate the proposed methods using the data set from a breast cancer clinical study (Wang *and others*, 2005). The data are available from the NCBI GEO database (GSE2034). This study was a large Affymetrix-based gene expression profiling study of 286 untreated patients with lymph node–negative primary breast cancer and analyzed estrogen receptor positive– and negative–patients separately. Here, we restrict our attention to the estrogen receptor positive patients. In this study, out of 22 283 genes, 60 genes were selected on the basis of statistical significance for predicting the risk of relapse. We considered comparison of patients who were relapse-free at 5 years (good prognostic group) and the other patients (poor prognostic group). During the follow-up period, of the 204 estrogen receptor positive patients, 138 patients were relapse-free at 5 years, while 66 developed distant metastasis.

The hyperparameters were estimated as $\hat{\pi}_0 = 0.769$, $\hat{\pi}_1 = 0.055$, $\hat{\pi}_2 = 0.176$, $\hat{\mu}_1 = 0.182$, $\hat{\tau}_1^2 = 0.021^2$, $\hat{\mu}_2 = -0.149$ and $\hat{\tau}_2^2 = 0.014^2$. The σ_j^2 were estimated on a gene-by-gene basis assuming a common variance between the 2 prognostic groups. Figure 1 represents comparison between the RPM statistic, which was shown to yield good gene ranking in the simulation studies in Section 4, and the other statistics for overexpressed genes in the poor prognostic group. Similar trends were observed in underexpressed genes for the poor prognostic group. Figure 1(a) indicates substantial discrepancy in gene ranking between the RPM statistic and the fold change. In Figure 1(b), top-ranked genes based on the RPM statistics had the smallest P-values, but low-ranked genes using the RPM statistic could also have very small P-values. Figures 1(c) and (d) indicate good agreement of gene ranking among the PM, RPM, and TPP statistics, especially, for the greatest values of these statistics (i.e. for top-ranked genes). For

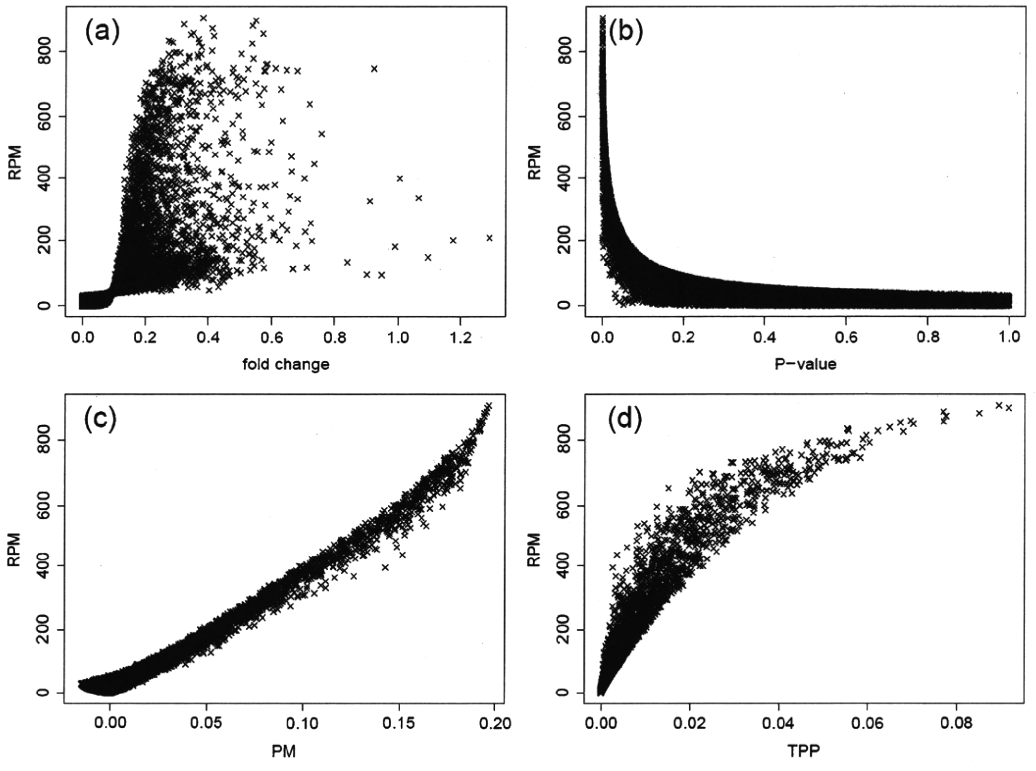


Fig. 1. Comparison of the RPM statistic with other statistics for the breast cancer data set. The 4 panels show scatter plots of the RPM statistic versus the fold change (a), the P-values from two-sample t-tests (b), the PM statistic (c), and the TPP statistic with $K = 30$ (d), for overexpressed genes in the poor prognostic group.

example, out of top 30 genes based on the RPM statistic, 27 and 21 genes were also selected in top 30 genes based on the PM and TPP statistics, respectively. The discrepancy in gene ranking between the proposed methods and fold change in Figure 1(a) would reflect very high false-positive rates for fold change as we found in our simulation study.

We also investigated overlap of top genes between the proposed methods and the 60 genes reported in the original paper (see Section B of the supplementary material available at *Biostatistics* online). There were 7 overlaps when the top 30 genes were selected by the RPM method for each of overexpression and underexpression in the poor prognostic group. 26 (43%) of the 60 genes reported in the original paper were among the top 1% of genes according to their RPM values. As indicated by Figure 1(b), there were fewer overlaps between the proposed methods and the t statistic. Because the sensitivity of the proposed methods was higher, gene ranking based on the proposed methods would be more reliable.

6. DISCUSSION

In microarray studies, prioritizing or ranking genes is an important statistical task. Because the number of simultaneous comparisons can go into the tens of thousands, gene ranking can suffer from a lack of accuracy. Sharing information across genes and incorporating the null/nonnull mixture structure are expected to be effective in improving accuracy. As seen in our simulations, the proposed PM, RPM, and TPP methods showed higher sensitivity and lower error of rank estimation for differential genes with

the greatest effects, compared to conventional methods. In addition, these methods had low false-positive rates. Although the results of our resampling studies indicated that such good performance can be compromised by violating the model assumptions, including independence among genes and the normality of gene expressions, our ranking methods, especially the RPM method, were still accurate compared to the other methods. Violation of normality can be handled by using other parametric distributions, as noted in Section 2. The greater accuracy of the PM method, compared to the PM_U method, shows that incorporating the null mixture component is effective in improving ranking accuracy. The PPDE ranking had the lowest false-positive rate, as was expected because of its theoretical optimality (Berger, 1985; McLachlan *and others*, 2006). These results are reasonable because a ranking method performs well for the criterion or loss function in gene ranking from which it is derived. Ranking methods should be selected according to the criterion of interest in gene ranking, that is, depending on the probability of nondifferential expression or the magnitude of differential expression.

The ranking accuracy of fold change would be asymptotically optimal because of the good performance in sensitivity and RMSE of gene ranking under large sample scenarios ($n/2 = 80$) in the simulations described in the supplementary material available at *Biostatistics* online. Further, the RMSE of fold change was lower than the proposed empirical Bayes methods under large sample scenarios in the resampling exercise in the supplementary material available at *Biostatistics* online. A similarly good performance using conventional methods, compared to those of empirical Bayes methods, has also been seen with large samples in other experiments (Greenland, 1993). In small sample scenarios, however, optimality was largely violated. Further, with respect to false-positive rates, the ranking of the fold change did not perform well, even with large samples. Although the MAQC project (MAQC Consortium, 2006) reported good reproducibility of ranking based on fold change, "accuracy" and "reproducibility" are different concepts as remarked by Witten and Tibshirani (in preparation). Hence, ranking via the fold change is not recommended for general use when the objective is to prevent false-positive detection.

For general practical use, the proposed 3 methods should be used according to the purpose of analysis. However, for the 3 proposed methods, the sensitivities in detecting differential genes with the greatest effects as well as the false-positive rates were comparable. Accordingly, we would recommend the RPM method.

As noted in Section 2, attempts to obtain stable estimates of the hyperparameters of interest, that is, ξ_1 and ξ_2 in the distribution of the effect sizes θ_j, include the use of reasonable estimates of the mixing proportions such as treating π_0 as fixed quantities, invoking reasonable constraints or placing prior distributions on the mixing proportions in the EM algorithm. Comparison of these approaches is outside the scope of this paper, but it is an important subject for future research.

The R code for gene ranking is available in the supplementary material at *Biostatistics* online.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors would like to thank Tomonori Oura for many valuable comments and advice on an earlier draft of this article, and the editor for helpful comments and suggestions. *Conflict of Interest:* None declared.

REFERENCES

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

CARLIN, B. P. AND LOUIS, T. A. (2009). *Bayesian Methods for Data Analysis*. New York: Chapman & Hall.

CHOE, S. E., BOUTROS, M., MICHELSON, A. M., CHURCH, G. M. AND HALFON, M. S. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology* **6**, R16.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science* **23**, 1–47.

EFRON, B., TIBSHIRANI, R., STOREY, J. AND TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

GOTTARDO, R., PANNUCCI, J. A., KUSKE, C. R. AND BRETTIN, T. (2003). Statistical analysis of microarray data: a Bayesian approach. *Biostatistics* **4**, 597–620.

GREENLAND, S. (1993). Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. *Statistics in Medicine* **12**, 717–736.

GUO, L., LOBENHOFER, E. K., WANG, C., SHIPPY, R., HARRIS, S. C., ZHANG, L., MEI, N., CHEN, T., HERMAN, D., GOODSAID, F. M. *and others* (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology* **24**, 1151–1161.

LAIRD, N. M. AND LOUIS, T. A. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics* **14**, 29–46.

LIN, R., LOUIS, T. A., PADDOCK, S. M. AND RIDGEWAY, G. (2006). Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis* **1**, 915–946.

LO, K. AND GOTTARDO, R. (2007). Flexible empirical Bayes models for differential gene expression. *Bioinformatics* **23**, 328–335.

LÖNNSTEDT, I. AND SPEED, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

LOUIS, T. A. AND SHEN, W. (1999). Innovations in Bayes and empirical Bayes methods: estimating parameters, populations and ranks. *Statistics in Medicine* **18**, 2493–2505.

MAQC CONSORTIUM (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**, 1151–1161.

MATSUI, S., ZENG, S., YAMANAKA, T. AND SHAUGHNESSY, J. (2008). Sample size calculations based on ranking and selection in microarray experiments. *Biometrics* **64**, 217–226.

MCLACHLAN, G. J., BEAN, R. W. AND JONES, L. B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**, 1608–1615.

MCLACHLAN, G. J., DO, K.-A. AND AMBROISE, C. (2004). *Analyzing Microarray Gene Expression Data.* Hoboken, NJ: Wiley.

NEWTON, M. A. AND KENDSIORSKI, C. M. (2003). Parametric empirical Bayes methods for microarrays. In: Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. (editors), *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer, pp. 254–271.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. AND TSUI, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Biostatistics* **8**: 37–52.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. AND AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176.

SHEN, W. AND LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Series B* **60**, 455–471.

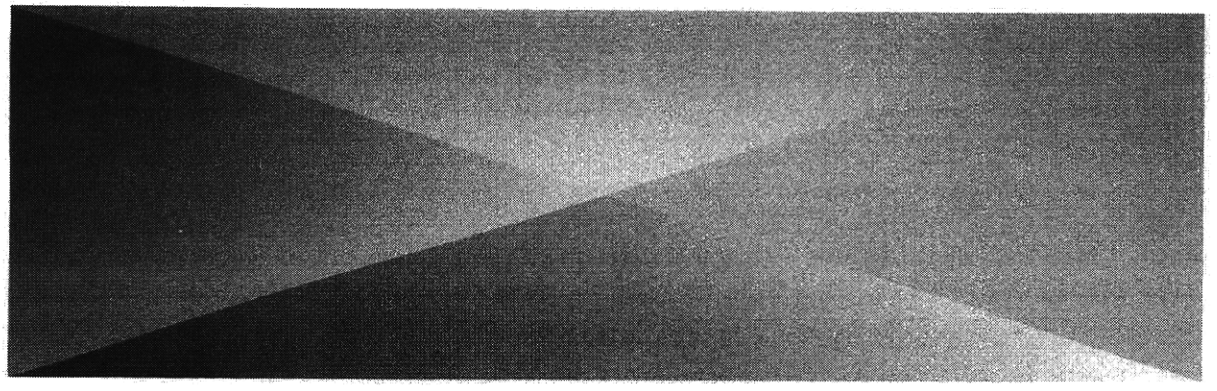
STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

WANG, Y., KLIJN, J. G., ZHANG, Y., SIEUWERTS, A. M., LOOK, M. P., YANG, F., TALANTOV, D., TIMMERMANS, M., MEIJER-VAN GELDER, M. E., YU, J. *and others* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679.

[*Received June 27, 2009; revised September 18, 2009; accepted for publication October 20, 2009*]

RCommanderによる
データ解析

大森　崇・阪田真己子・宿久　洋　著



共立出版

symmetrical. Log transformation is known to stabilize variance, normalize positively skewed distribution. This presentation discusses the advantages, limitations and methods suggested in the literature on log transformation.

Poster Presentation Abstracts

MU/IBS/06: Prediction of the toxicity score by in vitro test: An application of Bayesian statistical calibration

Takashi Omori
Department of Biostatistics, Kyoto University School of Public Health, Japan
Email: omori.t@ky8.ecs.kyoto-u.ac.jp

Objective(s) of the study : Chemical safety has become a major concern in the maintenance of human life. In the current circumstances, new testing methods are required to take animal ethics into consideration. Therefore, alternatives to animal tests are required. We consider an alternative test in which several reference chemicals are used for assessing an unknown chemical. We assume that the test method was validated through inter-laboratory studies. To obtain more information from an alternative test for the assessment of unknown chemicals, a reasonable approach would be to predict the score of the target animal test, thereby necessitating the use of statistical calibration. We propose the use of simple Bayesian statistical calibration, in which data from validation studies are used to elicit the prior distribution.

Methodology : Statistical calibration is typically performed in 2 stages: the calibration stage and the prediction stage. In the calibration stage, pairs of sample scores of the target animal test and the alternative test are required to estimate the regression function. In this stage of the proposed method, prior knowledge, which is obtained from validation studies, is also used for estimation. In the prediction stage, we calculated the unknown score by considering score of the alternative test of the unknown chemical.

Results and conclusion : We applied the proposed method to predict the maximum average score of the Draize eye-irritation test by using data from multi-laboratory studies on a cytotoxicity test performed in the 1990s. The predicted values were slightly less accurate.

Although further investigations are required, the direct prediction approach by the proposed method is expected to be useful for the assessment of chemicals.

MU/IBS/11: The design of three-arm non-inferiority trial including placebo with assay sensitivity

Eisuke Hida and Toshiro Tango
Email: e-hida@niph.go.jp
Department of Technology Assessment and Biostatistics,National Institute of Public Health, Japan

Three-arm trials including the experimental treatment, an active reference treatment and a placebo are recommended in guidelines of the ICH and EMEA as a useful approach to assessment of assay sensitivity. Statistical test procedures of three-arm non-inferiority

The validation studies conducted by JSAAE

Takashi Omori

(DoshishaUniversity)

The Japanese Society of Alternative to Animal Experiments (JSAAE) conducted validation studies for development alternatives, and I have participated mainly as a biostatistician in several studies. Here, I review some of my experiences and practical statistical issues.

In the 1990s, JSAAE conducted a large-scale validation study for searching several candidates for the Draize eye irritation test. In this study, over 1500 electronic data files were collected. Unfortunately, we found that these data files needed correction because many files were not prepared according to our expectations. We began data cleansing and requested the person in charge of the experimental laboratory to send the corrected files; however, the work delayed the study. Thereafter, when the JSAAE began the validation study, data management, that is, the method of collection of appropriate data, became one of the important issues.

In 2003, the JSAAE conducted a validation study to evaluate an alternative test for phototoxicity testing. The evaluated alternative test was a battery of tests, including red blood cell hemolysis assay and yeast growth inhibition assay. In the validation study, pre-experiments with a positive control chemical were made mandatory by the management team before formal experiments were conducted in all the experimental laboratories, although submission of the p[re-experimental data files before conducting the formal experiments was not required. After the experiments were completed, we collected both the pre-experimental data files and the formal experimental data files. On the basis of 2 graphs of positive control chemicals for pre-experiment and formal-experiment, it was suggested that the standard operating procedure might be altered at a laboratory and might be complied with at another laboratory. This graphical display indicated the need for checking the transferability.

In 2005, the JSAAE planned to conduct a validation study of the LLNA-DA (See Yahashita's abstract). In this assay, the stimulation index (SI), which is a ratio of the ATP amount for a chemical group and it for vehicle control group, is used to assess a chemical. According to the original LLNA, the SI is the primary measure for judgment of toxicity. However, some researchers prefer using statistical tests, because judgment based on the SI does not take variation of data into consideration. In this study, we prepared the SI plot with confidence interval as the error bar. This graph showed the results of the approximate statistical tests on the SI scale and successfully provided evidence for interlaboratory reliability of the LLNA-DA assay.

Statistical works beings the planning stages of the validation study and discussion with the members of the management team and the experimenters is important for data management and analysis. To obtain successful results from the validation study, both thoroughly examined proposed test method and good relationship and understanding between researchers are needed.

P-13

バリデーション研究に関する統計的支援について

○音泉卓[1]、大森崇[1]

[1]同志社大学文化情報学部

E-mail: bih0173@mail4.doshisha.ac.jp

[目的]

統計学における専門性は，データ解析や解釈の際に必要とされている．統計家は代替法を開発することに従事する研究者が統計解析を行う際に，アドバイザーとしてのみ従事することも可能ではある．しかし， The Organization for Economic Co-operation and Development (OECD) のガイダンス(GD)34 (OECD 2005) はその役割について "a statistical advisor (biostatistician) should be a member of, or consultant to, the validation manager/management team and be involved in all phases of the validation of new and revised tests." と述べている．Short-Time Exposure (STE) 試験は眼刺激性試験の代替法として期待される試験法であり，現在そのバリデーション研究が行われている．我々は統計的データ解析を支援するためにこの研究に参加しており，データの収集，データベースの構築，データ解析，得られた結果の報告を行っている．統計的専門性を提供するために，どのようにバリデーション研究に統計家が関わるかは，研究の状況によって異なるものであるが，そのあり方についての利点や欠点はあまり議論をされてこなかった．ここでは，STE 試験の経験を踏まえて，統計家の3つの支援の在り方を考察することにする．

[方法]

以下の3つの統計的支援としての統計家の関わり方を検討する．(1)すべての段階でバリデーション運営委員会に統計的な相談を受け，助言を行うが，データ解析は行わない統計家．(2) すべての実験が終了した後にデータを解析するために研究に参加する統計家データの収集から解析の報告を行う．(3)役割は(2)と同じであるが，計画段階から研究に参加する．

[結果と考察]

(1)では，統計家はデータ解析に割く労力は少なく，専門的な知識の提供に集中ができる．しかし，しばしば重要となる研究の実情の理解することが，困難となる可能性がある．(2)では，独立なデータ解析が行われるため，結果を保証する上で有用である．しかし，データを得た後の作業は多くなり，結果的に時間を費やすくことになる．このことは,しばしば運営委員会の要求に合わない．(3)では，統計家は計画の段階から研究の詳細を理解することが可能である．しかし，このことは多くの時間を費やすことにもなる．

他のバリデーションと統計家との共同作業は研究を成功させるための一つの大きな要因である．両者の関わり方に関しての利点や欠点が計画の段階から考慮されるべきであろう．

[参考文献]

OECD. Guidance document on the validation and international accetance of new or updated test methods for hazard assessment. 2005.

P-14

バリデーション研究における感度，特異度，一致度の統計的モデルを用いたアプローチ

○大森 崇

[1]同志社大学文化情報学部

E-mail: tomori@mail.doshisha.ac.jp

[目的]

感度，特異度，一致度は，代替法とその対象となる試験を比較において一致の度合いを表す割合である．これらはバリデーションの関連性評価において重要な指標である．一般にバリデーション研究では，数多くの異なる物質を用いることが要求される．一方，施設で実施できる実験数はさまざまな制約のために制限される．そのため，しばしば複数の施設で部分的，もしくは完全に異なる物質を用いて研究が実施されている．そのような場合のデータの要約方法はまだ十分に検討がなされていない．ここでは，統計的モデルの一つであるロジスティック回帰モデルを用いたアプローチが適用可能であることを示し，数値例を示すことにする．

[方法]

3施設（A，B，C）による仮想的なバリデーション研究を想定する．このバリデーション研究の主な目的は関連性(relevance)の評価である．40の異なる物質を用いたいが，各施設は制約のため30物質しか実験を行うことができないとしよう．データは，対象となる試験の陽性と陰性に基づき分類された頻度の集計を用いることにする．

ロジスティック回帰モデルは異なる条件の割合を推定する際に用いられる．このモデルの特定のコーディングの設定を用いて，感度，特異度，一致割合に対応する指標を直接推定することが可能である．感度はこのモデルの主効果と対象となる試験の陽性に関する因子で，特異度は主効果と対象となる試験の陰性に関する因子で推定できる．一致度は主効果のみのモデルを用いて推定できる．

[結果と考察]

感度と特異度が施設 A で5/15と12/15，施設 B で9/15と15/15，施設 C で8/15と13/15であるデータセットにおいて，これらの平均値により求めた感度，特異度，一致割合はそれぞれ0.6，0.89，0.74となる．一方，ロジスティック回帰モデルを用いたアプローチでは，対応する割合はそれぞれ0.60，0.89，0.79として得られる．

SL-1

Data science on validation studies conducted by JSAAE

○**Takashi Omori**

Doshisha University Faculty of Culture and Information Science
E-mail: tomori@mail.doshisha.ac.jp

[Objective(s)] Scientific evidence is derived from collected data. The responsibilities of a statistical expert include the design of data collection, validating the collected data, construction of databases to analyze data, data analysis, reporting the results, and so on. Some statisticians prefer to use the term "data science," rather than "statistics," as the former seems to better reflect their work. My personal experiences with statistical work in validation studies for the development of alternatives just fits to use this word "data science." Objective of this lecture is to present what we have learned through the data of the validation studies conducted by the Japanese Society for Alternative Animal Experiments (JSAAE).

[Materials and Methods] Statistical work on the following 3 validation studies conducted by JSAAE were presented from the statistical expert point-of-view: (Case 1) a large-scale validation study of cell cytotoxicities to investigate candidates for the Draize eye irritation test; (Case 2) validation of a battery of tests, including red blood cell hemolysis assay and yeast growth inhibition assay, to assess an alternative method for phototoxicity testing; (Case 3) validation of the modification to a local lymph node assay, known as the LLNA-DA.

[Results and Discussion] In case 1, over 1500 electronic data files were collected. Unfortunately, these data files required corrections because many of them were not prepared according to our expectations. Thereafter, data management became an important issue. In case 2, 2 graphs of positive control chemicals for pre-experiment and formal-experiment suggested that, the standard operating procedure might be altered at a laboratory level. This graphical display indicated the need for checking the transferability. In case 3, the plot of stimulation index, which is an important measure of the assay, was prepared with its confidence interval as the error bar. This graph showed the results of the approximate statistical tests on the SI scale and successfully provided evidence for inter-laboratory reliability of the LLNA-DA assay.

These cases indicated that statistical work begins at the planning stages of the validation study. Discussion with the members of the management team and the investigators is important for data management and data analysis. To obtain successful results from the validation study, not only a well-designed test method but also a good relationship and understanding between researchers is required.