residual, dose-dependency [regression ($P < 0.01$)] linearity (model fitness), and comparison with control [Dunnett's test ($P < 0.05$)].

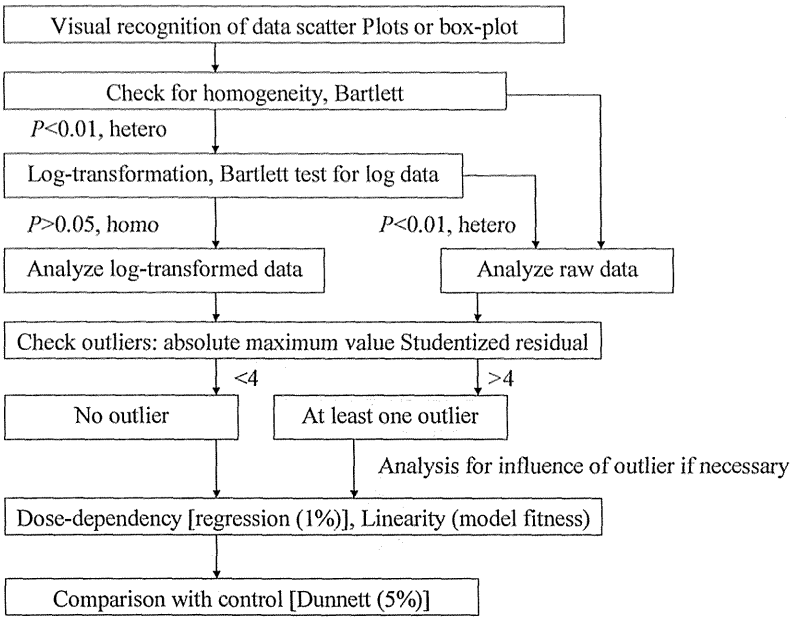

Fig. 4: Improvement Decision Tree Proposed by Hamada et al.

2.1.6. In 2000

A Simple decision tree (Fig. 5) was proposed by Kobayashi et al. [14]. This decision tree traces Bartlett's test, Dunnett's test, and Steel's test.

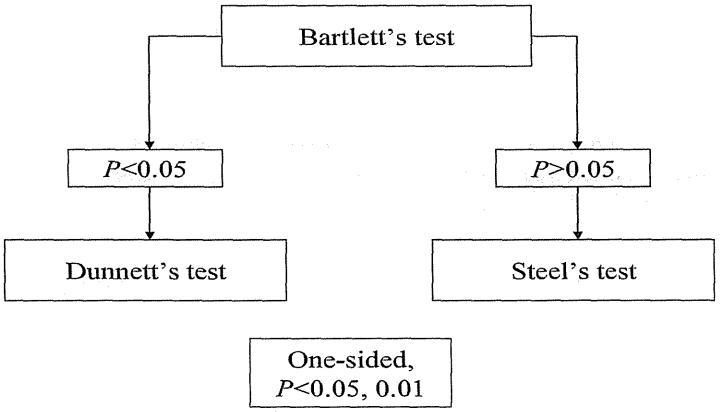

Fig. 5: A Simple Decision Tree Proposed by Kobayashi et al.

In the year 2000, another decision tree (Fig. 6) was proposed by the Japan Pharmaceutical Manufacturers Association (JPMA) working group [15, 16]. This tool is containing the Williams' test and Steel's test. The feature of this decision tree is to adopt Williams' test that assumes the dose related trend. The Bartlett test is not used.

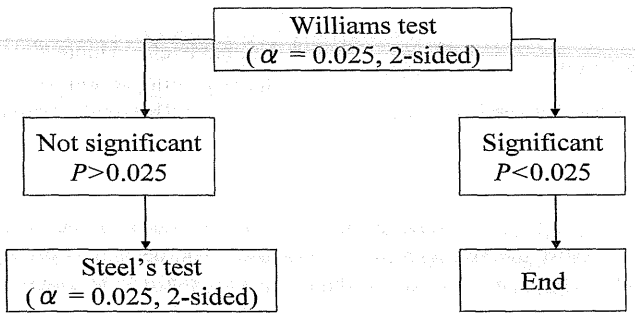

Fig. 6: Decision Tree Proposed by JPMA.

2.1.7. In 2008 (recent decision tree)

Most of the toxicologists adopt a conservative approach for analyzing the data. The data are examined for homogeneity of variance, and if the variance is homogeneous, parametric tests are used and for heterogeneous variance nonparametric tests are used. Usually, the data are not examined for normality, though it is a fact that for most of the statistical tools, it is also a requirement that the data show normality. If at all the data is examined for normality, it is not vividly explained in most of the books on biostatistics, what nonparametric statistical tools should be used for the data that show non-normality. Shapiro-Wilk's *W* test seems to be more appropriate for testing normality, as this test can be used for the data that shows normal or non-normality by visual examination of the graph. Kobayashi et al. [17] proposed a flow chart describing the statistical tool that may be used for the analysis of the data showing a normal or non-normality (Fig. 7 and 8).

It is important to examine the data for both homogeneity of variance and normality. The disadvantage of Bartlett's test which is widely used to examine for homogeneity of variance, is its hyper sensitivity to heterogeneous data [18]. We propose that when normality of each group is confirmed by Shapiro-Wilk's *W* test, Dunnett's test may be used for further analysis. When the control group or all groups do not show normality, Steel's test of separate type may be used. When normality is not shown by one or two of the dosage groups, the data may be analyzed using Dunnett's test after excluding the group/s that do not show normality. However, the biological relevance of the excluded data has to be carefully scrutinized.

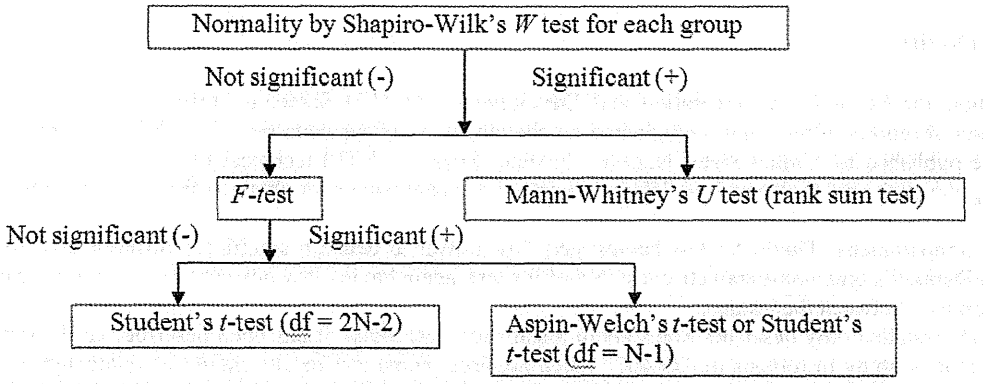

Fig. 7: Flow Chart for Selecting the Statistical Tool When the Data Show a Normality or Non-Normality (Number of Groups = 2).

Figs. 1–4 traces a complex path, whereas Figs. 5–8 a simple path. Statistical tools given in Figs. 1 and 3 were seldom used as of 2010. Statistical tools given in Fig. 1 were used in the 28-day repeated dose toxicity studies of existing chemical substances by the Guideline of the Chemical Substance Control Law [19] in Japan. However, we recommend the statistical tools of Fig. 5 for the analysis of the data obtained from repeated dose administration studies.

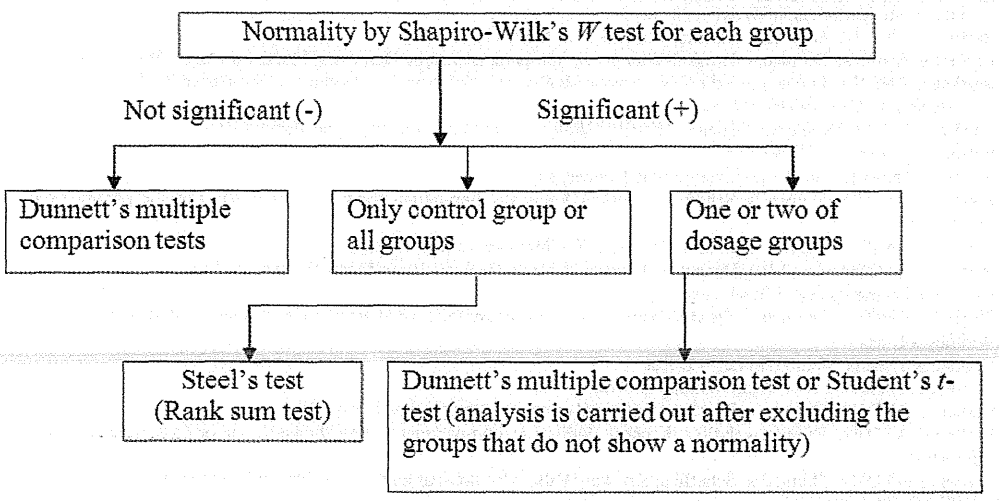

Fig. 8: Flow Chart for Selecting the Statistical Tool When the Data Show a Normality or Non-Normality (Number of Groups > 2).

2.2. Difference in the use of statistical tools for analyzing data obtained from 28-day repeated dose toxicity studies in various test facilities in Japan

2.2.1. In Japan

A total number of 122 numbers of 28-day repeated dose toxicity studies conducted in various test facilities in Japan during the period 1985–2004 were examined [20]. The studies were conducted following the guideline of the CSCL. The number of studies examined of each test facility is given in parenthesis: Food and Drug Safety Center, Kanagawa (22), An-Pyo Center, Shizuoka (22), Mitsubishi Chemical Safety Institute Ltd., Ibaraki (18), Safety Research Institute for Chemical Compounds Co., Ltd., Hokkaido (15), Bozo Research Center Inc., Shizuoka (12), Research Institute for Animal Science in Biochemistry & Toxicology, Kanagawa (11), Panapharm Laboratories, Kumamoto (10), Nihon Bioresearch Inc., Gifu (9), and National Institutes of Health Science, Tokyo (3).

Out of 122 studies examined, 79 studies used statistical tools that follow a complex path (tool numbers; 2, 3, 4, 5, 8, 9, 10, 12, 15, 16, and17) and 43 studies used statistical tools that follow simple path (tool numbers; 1, 6, 7, 11, 13, and14) (Table 1). The statistical tools describing the method of analyze, in the case of three or more groups and two groups were mentioned in six studies, whereas this description was not found in 11 studies. Only eight studies used trend test. In the tool number 10, the significance level of the ANOVA and Kruskal-Wallis's H test was set at $P = 0.10$. For comparing with the control, the tool set the significance level of $P = 0.05$. Tool numbers 13 and 14 did not perform Bartlett's test for testing the homogeneity of variance. Use of one-sided or two-sided test was not indicated in 87 studies. Only one study indicated use of non-parametric test. Student's or Aspin-Welch's t-test and Mann-Whitney's U test were used for analyzing the data of two groups of recovery period [21]. Student's or Aspin-Welch's t-test and Mann-Whitney's U test were used by recovery period in the two groups alone; other tests were used by administration period in the four groups. The difference in these analytical methods seems to depend on an examination enforcement year.

2.2.2. OECD SIDS

Organisation for Economic Co-operation and Development (OECD) Screening Information Data Set (SIDS) is a compilation of reports of the studies conducted on chemicals in various countries. OECD SIDS is more or less similar to the one published by United States National Toxicity Program (NTP) technical report. An extract of OECD SIDS explaining the statistical tools used in different countries for analyzing data obtained from toxicity studies are given in Table 2.

In Japan, nonparametric Dunnett's test having very low power to detect a significant difference is used. This test is similar to Dunnett's type non-parametric test. Scheffé's test, again having low power to detect a significant difference is used to compare between the groups.

The OECD guideline only describes notes about a statistical technique. It has been described in the repeated dose 28-day oral toxicity study in rodents of TG 407. When possible, numerical results should be evaluated by an appropriate and generally acceptable statistical method. Comparisons of the effect along a dose range should avoid the use of multiple t-tests. The statistical methods should be selected during the design of the study. Namely, multiple t-tests are a means of Dunnett's test of an enhanced version of the t-test [22].

Table 1: Classification of Number of Studies Based on the Statistical Tools Used for the Analysis of Quantitative Data

Tool no.	Description of statistical tools	Number of studies
1	Dunnett's, Student or Aspin-Welch's t-test	5
2	Bartlett's, ANOVA, Dunnett's, Kruskal-Wallis's H or Steel's test	7
3	Bartlett's, ANOVA, Dunnett's, Kruskal-Wallis's H, non-parametric type Dunnett's, Student or Aspin-Welch's t-test	9
4	Bartlett's, ANOVA, Dunnett's, Scheffé's, Kruskal-Wallis's H, Non-para type Dunnett's, non-parametric type Scheffé's test, Student or Aspin-Welch's t-test	10
5	Bartlett's, NOVA, Dunnett's, Duncan's, Kruskal-Wallis's H or non-parametric type Dunnett's test	9
6	Bartlett's, Dunnett's or Steel's test	20
7	Bartlett's, Dunnett's, or non-parametric type Dunnett's test	10
8	Bartlett's, ANOVA, Dunnett's, Scheffé's, Kruskal-Wallis's H, non-parametric type Dunnett's test or non-parametric type Scheffé's test	23
9	Bartlett's, ANOVA, Dunnett's, Kruskal-Wallis's H or Mann-Whitney's U test	14
10	Bartlett's, ANOVA ($P = 0.10$), Dunnett's, Kruskal-Wallis's H ($P = 0.10$) or Mann-Whitney's U test	1
11	Bartlett's, Dunnett's test or Steel's test	3
12	Bartlett's, ANOVA, Dunnett's, Kruskal-Wallis's H, non-parametric type Dunnett's test, Student's t-test or Mann-Whitney's U test	1
13	Dunnett's, t-test or Mann-Whitney's U test	4
14	Dunnett's, Scheffé's, t- or Mann-Whitney's U test	1
15	Bartlett's, ANOVA, Dunnett's, Kruskal-Wallis's H or non-parametric type Dunnett's test	3
16	Bartlett's, ANOVA, Dunnett's, Jaffé's, Kruskal-Wallis's H, non-parametric type Dunnett's test or non-parametric type Jaffé's test	1
17	Bartlett's, ANOVA, Dunnett's, Scheffé's, Kruskal-Wallis's H, non-parametric type Dunnett's, non-parametric type Scheffé's or Student's t-test	1
	Jonckheere's trend test (not included in the number of tools)	8
	Total	122

Table 2: Statistical Tools Used in Different Countries for Analyzing Data Obtained From Toxicity Studies

(1) Country and year, (2) Cas No., (3) Test substance, (4) Test guideline (TG) No. in OECD or test period, (5) Analytical tools
(1) Belgium, 2002, (2) 144-55-8, (3) Sodium bicarbonate, (4) 32 wk, (5) Dunnett's multiple comparison test, LSD test, Mann-Whitney's *U*-test, and Student's *t*-test
(1) BMU*, 2004, (2) 25321-14-6, (3) Dinitrotoluene, (4) 104 weeks & 52 weeks (5) ANOVA, Bartlett's test, and Dunnett's test
(1) France, 2002, (2) 2432-99-7, (3) 11-aminoundecanoic acid, (4) TG 407, (5) No mentions
(1) France , 2003, (2) 115-11-7, (3) Isobutylene, (4) 105 wk, (5) Cox's method, Tarone's life table, The Poly-k, Dunnett, Williams, Shirley, and Dunn tests
(1) Germany, 2002, (2) 90387-57-8, (3) Formaldehyde, reaction products with sulfonated 1,1'-oxybis [methylbenzene], sodium salts, (4) TG 414, (5) ANOVA, Dunnett test, Healy test, Kruskal-Wallis test, and Dunn test
(1) Germany, 2003, (2) 556-82-1, (3) 3-Methylbut-2-en-1-ol, (4) TG 408, (5) Dunnett test, Kruskal-Wallis test, and Wilcoxon test
(1) Germany, 2003, (2) 947-04-6, (3) Dodecane-12-lactam, (4) TG 408, (5) Levene, 1-ANOVA, Student *t*, Bonferroni's, Scheffé's, and Kruskal-Wallis tests
(1) Germany, 2003, (2) 288-32-4, (3) Imidazole, (4) TG 408, (5) Dunnett's test, Kruskal-Wallis test, and Wilcoxon-test
(1) Germany, 2003, (2) 122-52-1, (3) Triethyl phosphate, (4) TG 421, (5) ANOVA, *F*-test, *t*-test, and Welch-*t*-test
(1) Germany, 2003, (2) 108-39-4, 106-44-5, and 15831-10-4, (3) m/p-Cresole, (4) 28 days, 27 weeks, and two generation study in mice, (5) Dunn and Shirley, Jonckheere's, Levene's ANOVA, *t*-, Bonferroni method, Kruskal-Wallis, Mann Whitney's *U*, Turkey's test, and covariance *F*-tests
(1) Germany, 2003, (2) 3323-53-3, (3) Adipic acid, compound with hexane-1,6-diamine (1:1), (4) Subchronic for 4 weeks, (5) Dunnett's test, Bartlett's test, Mann-Whitney test, and Bonferroni test
(1) Germany, 2004, (2) 2855-13-2, (3) 3-Aminomethyl-3,5,5-trimethylcyclo hexylamine, (4) Subchronic for 13 weeks, and inhalation for 14 days, (5) Dunnett-test, Steel-test, ANOVA, and Bartlett's test
(1) Japan, 2001, (2) 5392-40-5, (3) Citral, (4) 14 days, TG 421, (5) William's, Dunnett's test, and Mann-Whitney's *U* tests
(1) Japan, 2002, (2) 126-98-7, (3) Methyl acrylonitrile, (4) Inhalation/days 6 to 20 of gestation, (5) ANOVA, Dunnett's test, and Wilcoxon test
(1) Japan, 2002, (2) 16219-75-3, (3) 5-Ethylidene-2-norbornene, (4) TG 421, (5) Bartlett's, ANOVA, Kruskal-Wallis, nonparametric Dunnett's or parametric Dunnett's test
(1) Japan, 2002, (2) 25321-09-9, (3) Diisopropylbenzene, (4) TG 407, (5) Not mentioned.
(1) Japan, 2004, (2) 56539-66-3, (3) 3-Methoxy-3-methyl-1-butanol, (4) TG 407, (5) Dunnett's or Scheffé's test
(1) Japan, 2004, (2) 793-24-8, (3) N-(1,3-Dimethylbutyl)-N′-phenyl-1,4-phenylenediamine, (4) Subchronic, 13 weeks, (5) Dunnett's test and Mann-Whitney test
(1) Korea, 2002, (2) 94-36-0, (3) Benzoyl peroxide, (4) TG 422, (5) Dunnett's multiple comparison test
(1) Switzerland, 2001, (2) 6386-38-5, (3) Metilox, (4) Reproduction, (5) ANOVA, Dunnett's *t*-test, Kruskal-Wallis, and Mann-Whitney's *U*-test
(1) Switzerland, 2002, (2) 115-95-7, (3) Linalyl acetate, (4) 28 days, (5) ANOVA
(1) USA, 1996, (2) 112-35-6, (3) 2-(2-(2-Methoxyethoxy)ethoxy)-ethanol, (4) TG 408, (5) Levene's test, ANOVA, *t*-tests
(1) USA/IT, 2001, (2) 120-61-6, (3) Dimethyl terephthalate, (4) Reproductive toxicity study, (5) ANOVA and Dunnett's-*t*-test
(1) USA, 2001, (2) 126-73-8, (3) Tributyl phosphate, (4) Subchronic, 13 weeks, (5) ANOVA, Bartlett's test, and Dunnett's test
(1) USA, 2003, (2) 919-30-2, (3) 3-aminopropyl triethoxysilane, (4) TG 408, 91 or 92 days (5) ANOVA, Dunnett's test, Kruskal-Wallis test, and Mann-Whitney's *U*-test

*Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit.

2.2.3. NTP, U.S.A.

The methods used to analyse the data obtained from 84 short-term toxicity studies and 588 long-term carcinogenicity/toxicity studies conducted on chemical substances published in NTP technical reports [23] in 2014 were examined. The NTP technical report series are using the same statistical analysis method. The findings are given below:

- Dunnett [24] and Williams [25],[26] parametric multiple comparison tests for organ and body weights data
- Shirley [27] and Dunn [28] nonparametric multiple comparison test for hematology, clinical chemistry, spermatid, and epididymal spermatozoa/typically skewed distributions
- Jonckheere's test [29] and Williams' or Shirley's test for dose-related trends
- Mann-Whitney's *U* test [30]
- Bartlett's test, ANOVA, Dunnett's test, Kruskal-Wallis test, and Dunn's test for dam and pup data from the *in utero* phases of rats

2.3. A comparison of statistical tools for analyzing the data obtained from repeated dose toxicity studies with rodents in Japan with that of used in other countries

Statistical tools used for analyzing the data obtained from 127 repeated dose toxicity studies with rodents from 45 countries were compared with that of Japan [31]. Scheffé's multiple range parametric and non-parametric tests and Dunnett's type (joint type Dunnett) were commonly used in Japan, but in other countries use of these statistical tools is not so common. However, statistical techniques used for testing the data for homogeneity and inter-group comparisons did not differ much between Japan and other countries. In Japan, the data were not tested for normality and the same was true with the most of the countries investigated. In fact, out of 127 studies, the data obtained from only 6 studies were examined for both homogeneity and normality.

The classification of statistical analyses methods by cluster analysis is given in Fig. 9 and Table 3. As per the analysis, 11 studies fall in cluster 1, two in cluster 2, 109 in cluster 3, and six studies fall in cluster 4. The power for significant difference among the groups using the statistical tools of cluster 1 is extremely low. If the variance of the groups is unequal, using the statistical tools of this cluster may not show a significant difference in the low dose group. The statistical tools of cluster 2 is close to cluster 1, hence the detection power of this cluster is similar to that of cluster 1. If the number of animals is different in the groups, which is usually seen in repeated dose toxicity studies, the power of

detection of a significant difference of the statistical tools of this cluster is further decreased. The statistical tool of cluster 3, which has high detection power, is commonly used in most of the countries. In cluster 4, statistical tools having high detection power were used to examine both homogeneity and normality.

Table 3: Grouping the Studies in Clusters

Cluster	Statistical tools used
1	The parametric data were analyzed by Dunnett' test and the nonparametric data were by Dunnett type rank sum test or Dunn's multiple comparison tests.
2	The parametric data were analyzed by Dunnett' or Scheffé's' test. The nonparametric data were analyzed by Dunnett type rank sum test.
3	After carrying out ANOVA or the data were directly subjected to Dunnett's, Duncan's, and Student's or Mann-Whitney test.
4	The detection power of the analytical method is high. The homogeneity was examined by Levene's test, which has of low detection power. Data were also examined for normality.

Seven studies from Japan are grouped in cluster 1 of 11 analytical tools, two are grouped in cluster 2 of two analytical tools and six are grouped in cluster 3 of 109 analytical tools. No study from Japan is placed in cluster 4 of five analysis tools (Table 4).

Table 4: Number of Toxicity Studies Conducted in Japan in Each Cluster

Cluster (color in Fig. 9)	Rate of the number of studies performed in Japan	
1 (red)	7/11	9/13 (69%)
2 (orange)	2/2	
3 (green)	6/116	6/121 (4.9% = < 5%)
4 (blue)	0/5	

Bartlett's test was used to examine homogeneity in studies conducted in most of the countries. However, six studies used Levene's test (Levene, 1960) to examine homogeneity, which has less power compared to Bartlett's test. Shapiro-Wilk's *W* and Kolmogorov-Smirnov tests were used in two studies each (Table 5). Interestingly, statistical tool used for post hoc comparison was not mentioned in 14 studies. We propose Levene's test for examining homogeneity, since the sensitivity of Bartlett's test is too high to detect a non-homogeneous distribution. However, we propose examining the data for both homogeneity and normality [17].

Table 5: Number of Studies Subjected to Homogeneity and / Or Normality Tests

Test for homogeneity or normality	No. of studies/127 studies
Levene's homogeneity test and Shapiro-Wilk's *W* test or Kolmogorov-Smirnov's test	4
Levene's homogeneity test	2
Shapiro-Wilk's *W* test	1
Kolmogorov-Smirnov's test	2

2.4 Notes on few statistical techniques commonly used in analysis of the data

2.4.1. Use of ANOVA

It is a common practice to use ANOVA for analysing data obtained from three or more than three groups. However, several authors prefer not to use ANOVA, as it may cause type II error. Dunnett [24] never recommended ANOVA for the analysis of data obtained from toxicity studies. A significant difference can be detected by analyzing the data it directly by Dunnett's test even if a significant difference is not found by the ANOVA. An example is given in Table 6, where a significant difference is not shown by ANOVA, but shown by Dunnett's test. In Japan, data obtained from several toxicity studies were not analysed by using ANOVA [32], [33], [34], [35], and [36].

Table 6: A Significant Difference Is Not Shown by ANOVA, But Shown by Dennett's Test

Data	Control	Low dose	Middle dose	High dose
B6C3F1 mice, liver weights (g), N = 10	1.08, 1.09, 1.15, 1.09, 1.16, 1.00, 1.12, 1.01, 1.12, 1.02	1.09, 1.12, 1.15, 1.09, 1.04, 0.99, 1.24, 1.15, 0.99, 1.12	1.10, 1.20, 1.09, 1.02, 1.07, 1.12, 1.13, 1.06, 1.11, 1.20	1.16, 1.15, 1.24, 1.16, 1.22, 1.10, 1.18, 1.07, 1.18, 1.09
Mean ± S.D.	1.08 ± 0.06	1.10 ± 0.08	1.11 ± 0.06	1.16 ± 0.05
Bartlett's test, $P = 0.068$, Not significant difference (NS)				
ANOVA, $P = 0.0715$, NS				
Dunnett's test*		$P = 0.9233$	$P = 0.6742$	$P = 0.0399$

S.D; standard deviation. * By two-sided test.

2.4.2. Williams test doesn't give measured raw values without dose-trend

Williams' test [25], [26] is generally carried out to test dose-related trend [37]. The test can be used when the number of animals is equal in each treatment group and the mean values of the treatment effect show a dose related pattern [38, 11, and 39]. This test is not widely used in Japan that has been published.

2.4.3. Power of nonparametric tests

Number of animals required in the low dose group to show a significant difference of this group by rank sum test is given in Table7. Scheffé's test, which is low sensitive, requires 22 and 40 animals, respectively in the four and five group's experimental design, in the low dose group to show a significant difference. Among the rank sum tests in the Table 7, Steel's test is least sensitive. Inaba [40] and Kobayashi et al. [41] explained of the low power of nonparametric type Dunnett's test, for finding a significant difference in the low dose group.

Table 7: Number of Animals in the Lowest Group from Which Low Dosage Group Can Detect Significant Difference by Rank Sum Test

Test	Number of group	
	4	5
Scheffé type	22	40
Hollander-Wolfe or Dunn's test.	19	30
Tukey type	18	32
Dunnett type	15	26
Williams-Wilcoxon	8	12
Steel	4	6
Mann-Whitney's *U* (two groups test)	3	-

The rank sum test widely used in various countries are Williams-Wilcoxon, Hollander-Wolfe [30], and Steel's tests [42].

2.4.4. Joint type and separate type non-parametric Dunnett's test rank sum tests

Two techniques are used in Japan while analyzing the data using nonparametric Dunnett's test (Inaba, 1994) (rank sum test). The technique that uses all groups' orders is called joint type Dunnett's test and the technique that uses the order of the control group and one dosage group is called Dunnett's separate type. Dunnett's separate type is similar to Steel's test. A significant difference is more prone to be detected in separate type if number of samples in each group is four and five. Among the rank sum tests given in Table 8, the power of Dunnett's separate type test (Steel's test) is the highest. The level of the power of the Steel's test is equal to Mann-Whitney's *U* test. The difference between the joint type and separate type non-parametric Dunnett's test rank sum test is described elsewhere in detail [43]. Interested authors can refer to Table 8 for several names of nonparametric Dunnett's test.

Table 8: Source Thesis of Nonparametric Dunnett's Test

Nonparametric Dunnett's test	References
Nonparametric Dunnett	
Dunnett's (mean) rank test	Sakuma [44], Yamazaki et al. [10]
Dunnett-type rank test	
Steel's test	Steel [42], Yoshimura and Oohashi [45], Nagata and Yoshida [46]
The difference of nonparametric Dunnett was described.	Inaba [40], Kobayashi et al.[41], Yoshimura and Oohashi [47]

2.4.5. The rank sum test that is nonparametric procedure is not an analysis of the difference of the mean value

The rank sum test analyzes the difference of average ranks between groups. An example is given in Table 9. Though the mean values of the control group and the high dose group are the same, the rank sum analysis shows a significant difference between the ranks. The reason for the mean value of the high dose group is similar to that of the control is because of the value, 2.96 in the high dose group. This situation warrants examining the data for outliers by Smirnov-Grubbs test for high dose group [48]. The outliers may be excluded from the data before ranking them. It may be possible that after removing the outliers, the data may show a normality and homogeneity of variance. This example is seldom occurring. Takizawa [49] has pointed out a similar case that a statistically significant difference was observed in the same average value.

Table 9: Notation of Result for Rank Sum Test

Group	Creatinine values (mg/dL) in rats at Week 52 (N = 20)	Variance ratio (*F*)	Mean ± S.D.	Mean rank#
Control	0.70 0.68 0.70 0.74 0.60 0.65 0.65 0.72 0.63 0.78 0.67 0.64 0.63 0.66 0.88 0.73 0.57 0.79 0.78 0.65	52 $P < 0.001$	0.69 ± 0.07N	27.95
High dose	0.51 0.59 0.49 0.60 0.58 0.62 0.51 0.57 0.60 2.96 0.56 0.65 0.71 0.55 0.54 0.41 0.52 0.62 0.59 0.59		0.69 ± 0.54**	13.05** $U = 51$

N, nonparametric rank sum test.
By Mann-Whitney's *U* test.
**Significantly different ($P < 0.01$) from control group.

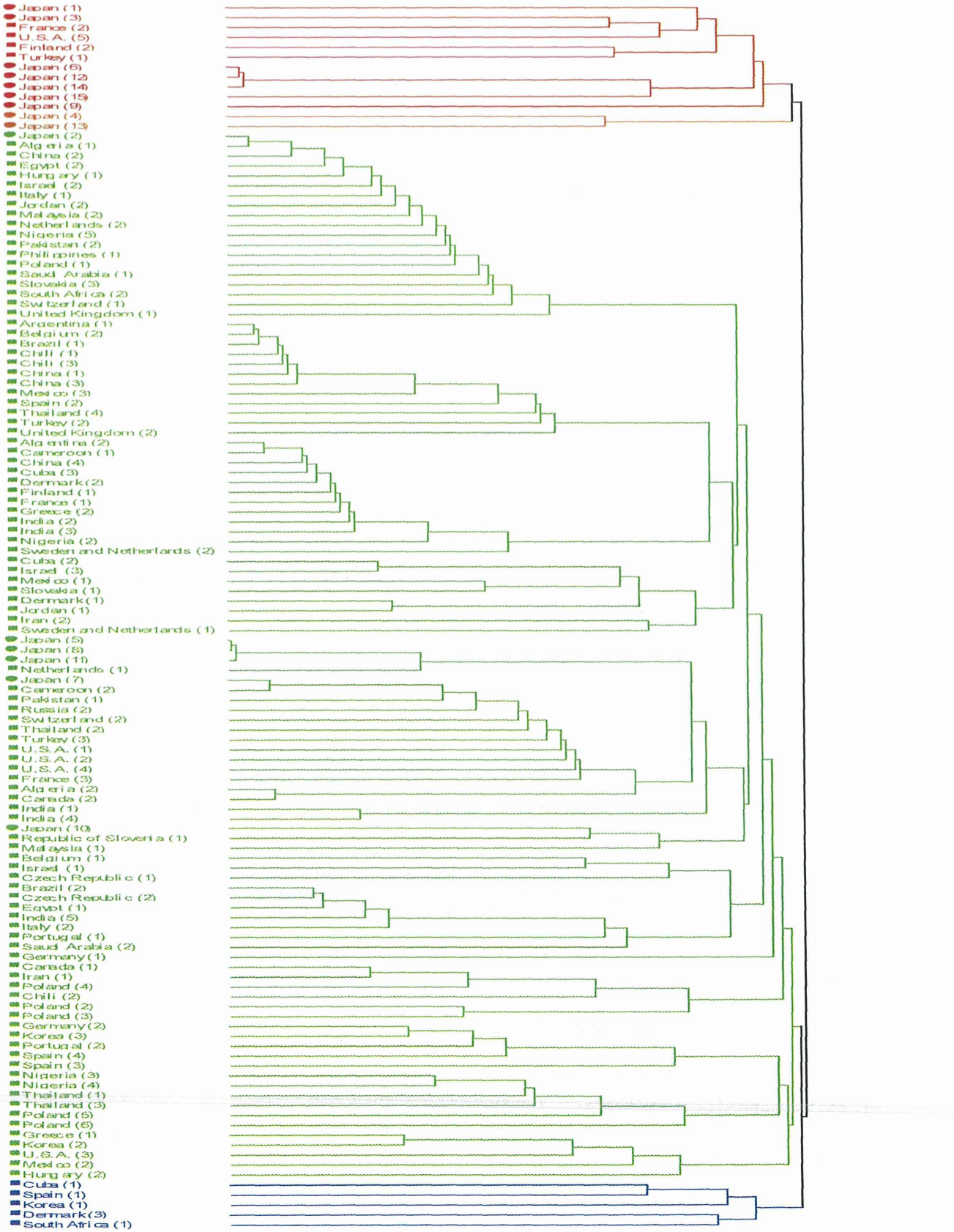

Fig 9: Classification of Statistical Analysis Methods by Cluster Analysis.

2.4.6. Which test to be used one-sided or two-sided?

Kobayashi et al. [21] examined whether a one-sided or two-sided test was used in the analysis of the data obtained from 122 numbers of 28-day repeated dose administration studies in rats. The studies were conducted as per CSCL or OECD test guideline (TG 407) [22]. Out of 122 studies examined, quantitative data of 22 studies were analysed by the one-sided test, 13 studies were analysed by two-sided test, whereas there was no mention about whether the one-sided or two-sided test was used in 87 studies. With regard to qualitative data, in 34 and 22 studies the data were analysed by the one-sided and two-sided tests, respectively, whereas there was no mention about whether the one-sided or two-sided test was used in 70 studies (Table 10).

Table 10: Use of One-Sided or Two-Sided Test for Short-Term Repeated Dose Administration Toxicity Studies with Rats

Data	One-sided	Two-sided	No description	Total
Quantitative	22	13	87	122
Qualitative	34	22	70	126

Kobayashi [50] recommended a one-sided test for the analysis of data obtained from toxicological studies. A significant difference is more apt to be observed in a one-sided test than in a two-sided test. According to a survey, the delectability of a significant difference by the two-sided test was 71–95% of that by a one-sided test in Dunnett's multiple comparison test (Table 11).

Table 11: Difference In Number of Detected Significant Differences ($P < 0.05$) One- and Two-Sided Test by Dunnett's *t*-Test in a Combined Chronic Toxicity/Carcinogenicity Study in Rats

Measurement item	No. of statistical analyses	Dunnett's *t*-test	
		One-sided	Two-sided
Body weight	528	223 (100)	212 (95)
Feed consumption	832	235 (100)	189 (80)
Hematology	352	123 (100)	105 (85)
Blood chemistry	576	215 (100)	181 (84)
Urinalysis	64	7 (100)	5 (71)
Organ weight	224	47 (100)	42 (89)
Organ weight/BW	224	82 (100)	67 (81)
Total	2800	932 (100)	801 (86)

( ): In % of one-sided value in each item.

In the references, out of 700 items for all tests, 578 showed significant differences in unidirectional changes in relation to the control group, and 39 items indicated significant differences with bidirectional changes in values higher and lower than those in the control group. The ratio of the bidirectional pattern (39) to the unidirectional pattern (578) was 1 to 15. Consequently, the one-sided test may be recommended for statistical analyses of toxicological bioassay data that control values, since more rigid evaluation of the data of the chemical effects on the living body and the environmental is necessary [51]. Properties of quantitative data from a combined chronic toxicity/carcinogenicity study are described. Trend of significant differences between each treated group and the control group obtained from actual two years long-term studies are shown in Table 12.

Table 12: Trend of Significant Differences Compared with Control Group

Measurement item	No. of statistical analyses	Changes	
		Unidirectional	Bidirectional
Body weight	132	130	0
Feed consumption	208	156	22
Hematology	88	72	7
Blood chemistry	144	125	8
Urinalysis	16	9	0
Organ weight	56	40	2
Organ weight/BW	56	46	0
Total	700	578	39

Scientists have different views on the use of one- or two-sided test. Shirley [27] used the two-sided test for Student's *t*-test and Cochran's *t*-test, and if significant differences were observed in the ANOVA, they used the one-sided test in Dunnett's test. Dunnett [24] recommended use of the two-sided test to determine simultaneous upper and lower limits for the difference between the control group and each treated group; he used the one-sided test to determine either the upper or lower limit on the difference between the control group and each treated group.

Gad and Weil (1986) explained the significant difference between control and treated groups in body weight by using the two-sided test. Yoshimura and Ohashi [52] recommended using the one-sided test in the analysis of toxicological study data, because toxicity is the absence of an increase in the mean values in most of the parameters. Additionally, quantitative data obtained in the toxicological test should be analysed using the one-sided test when a difference compared with the control group is anticipated unidirectional (either increase or decrease) before the experiment. Two-sided test should be employed when the difference cannot be anticipated unidirectional before the experiment. The

change in the hypothesis can cause type I error. It is very important to increase the power of detection of the differences among the groups.

Generally in toxicity studies, the study director can make a decision on it based on the preliminary studies. Sakuma [44] suggested selecting a one- or two-sided test based on the reports on similar studies conducted. In screening tests for new drugs, the two-sided test is recommended. The study director should not change the hypothesis from a two-sided to a one-sided test after the test has been finished. Nakamura [51] stated selection of the tests depends on the purpose and content of the study, and the statistical significance of the data should not be foreseen. Ishii [53] stated that it is necessary to select properly according to the situation in which the difference between two cases has to be considered to be either plus or minus alone and in which the difference has to be considered to be both plus and minus.

2.4.7. What is the multiple of the statistical analysis and toxicity studies?

Difference in detection of significant differences between Student's *t*-test and Dunnett's test is showed in Table 13 [48]. In analyses with the *t*-test, the number of significant differences detected was more than that detected by Dunnett's test.

Table 13: Difference In Number of Detected Significant Differences ($P < 0.05$) Between Student's *t*-Test and Dunnett's *t*-Test by One-Sided Test in a Combined Chronic Toxicity/Carcinogenicity Study in Rats

Measurement items	No. of statistical analyses	One-sided *t*-test	
		Student's	Dunnett's
Body weight (BW)	528	246 (100)	223 (91)
Feed consumption	832	349 (100)	235 (67)
Hematology	352	159 (100)	123 (77)
Blood chemistry	576	272 (100)	215 (79)
Urinalysis	64	11 (100)	7 (64)
Organ weight	224	80 (100)	47 (59)
Organ weight/BW	224	104 (100)	82 (79)
Total	2800	1221 (100)	932 (76)

( ): In % of Student's *t*-test value in each item.

Repeated dose administration studies are usually conducted with four or more than four groups. The mean values of the findings of are compared among the treated groups and between a treated group and control. In this situation multivariate analytical tool is the ideal one [11]. Use of *t*-test for analyzing data derived from more than 2 groups may cause the type I error. The significance level value P becomes $1 - (1 - 0.05)^3 = 0.142$, if the data of a four-group setting is analysed three times by *t*-test.

2.4.8. Which technique does the homogeneity tests of variance?

Finney [18] did not recommend use of Bartlett's homogeneity test, because of its strong power to detect a non-homogeneity distribution. Power of various homogeneity tests is given in Table 14. Power to detect a significant difference is highest in Bartlett's test, followed by Levene's, Brown-Foresythe's, and O'Brien's tests. When the groups contain more number of animals, it is more likely that Bartlett's test show a significant difference [54]. All the tests mentioned above will have a similar power to detect a significant difference, when all the groups show a similar distribution. For testing homogeneity, we recommend Levene's test [55]. The OECD recommends this method [56].

Table 14: Water Consumption (G/Week) In B6C3F1 Mice at Week 13- Power of Various Homogeneity Tests

Group	No. of animals	Mean ± S.D.	*P* value by each homogeneity test			
			O'Brien	Brown-Foresythe	Levene	Bartlett
1	10	43.8 ± 9.0	0.0459	0.0340	0.0014	< 0.0001
2	10	35.4 ± 3.4				
3	10	31.9 ± 1.5				
4	10	30.7 ± 2.1				

3. Conclusion

Detection of a significant difference using statistical analysis in repeated dose administration studies is influenced by the magnitude of difference between the means and the variance, and number of animals of the groups. For the analysis of data obtained from repeated dose administration studies, we may suggest to use a the decision tree with a simple route to select an appropriate statistical tool, examine the data for both homogeneity and normality and use the one-sided test with high power for detecting a significant difference. As far as possible avoid carrying out statistical analysis on the transformed of data, as interpretation of such statistical analysis is difficult. For examining homogeneity of variance, we may propose Levene's test. And finally, when the statistical analysis is interpreted, more important thing should be given to biological relevance than to statistical relevance.

Acknowledgments

We would like to acknowledge the continuing guidance and encouragement of Dr. Akihiko Hirose, Division of Risk Assessment, Biological Safety Research Center, National Institute of Health Sciences.

References

[1] Mitsumori K, Usui T, Takahashi K, Shirasu Y. Twenty-four month chronic toxicity studies of dichlorodisopropyl ether in mice. J Pesticide Sci. 1979; 4(3): 323–335. http://dx.doi.org/10.1584/jpestics.4.323.

[2] Maita K, Hirano, M, Mitsumori K, Takahashi K, Shirasu T. Subchronic toxicity studies with zinc sulfate in mice and rats. J Pesticide Sci. 1981; 6(3): 327–336. http://dx.doi.org/10.1584/jpestics.6.327.

[3] Hashimoto K, Imai K, Yoshimura S, Ohtaki T. Toxicity evaluation of a potential inhibitor of angiotension converting enzyme, 3. Twelve months studies on the chronic toxicity of captopril in rats. J Toxicol Sci. 1981; 6 (Supple II): 215–246. http://dx.doi.org/10.2131/jts.6.SupplementII_215.

[4] Shinpo K, Yokoi Y, Fujiwara S. General toxicity of α, β-adrenoceptor- blocking agent labetalol hydrochloride, 3th. 90-day oral administration toxicity and recovery studies in rats. J Toxicol Sci. 1981; 6(1): 37–59. http://dx.doi.org/10.2131/jts.6.37.

[5] Hirayama H, Wada S, Shikuma H, et al. Acute toxicity and subacute oral toxicity tests of thymoxamine hydrochloride (M-101) in rats. Kisotorinsho. 1982; 16: 1147–1173.

[6] Shimazu H, Takeda K, Onodera C. et al. Intravenous chronic toxicity of lentinan in rats: 6-month treated and 3-month recovery. J Toxicol Sci. 1980; 5(Supple): 33–57. http://dx.doi.org/10.2131/jts.5.Supplement_33.

[7] Takeuchi M, Iwata M, Kiguchi M, Kaga M, Shimpo K. Chronic toxicity study of AC-1370 sodium, an antibiotics, intravenously administration in rats. J Toxicol Sci. 1984; 9(4): 363–388. http://dx.doi.org/10.2131/jts.9.363.

[8] Imoto S, Yahata A, Kosaka M, et al. Peri- and postnatal study on halopredone acetate in rats. J Toxicol Sci. 1985; 10(Supple II): 105–122. http://dx.doi.org/10.2131/jts.10.SupplementI_105.

[9] Kobayashi, K. Dunnett's multiple comparison tests. Bull. Jap Soc Biopharm Stat. 1983; No. 10: 11–15.

[10] Yamazaki M, Noguchi Y, Tanda M, Shintani S. Statistical method appropriates for general toxicological studies in rats. J Takeda Res Lab. 1981; 40 (3/4): 163–187.

[11] Gad CS, Weil SW. Statistics and experimental design for toxicologists. The Telford Press Inc., NJ U.S.A., 1986; pp 18.

[12] Sano M, Okayama, K. *P* value programs and highly accurate percent point table of one and two sided tests for Dunnett multiple comparison test. Jap Soc Biopharm Stat. 1990; No. 32: 21–44.

[13] Hamada C, Yoshino K, Matsumoto K, Nomura M, Yoshimura I. Three-type algorithm for statistical analysis in chronic toxicity studies. J Toxicol Sci. 1998; 23(3): 173–181. http://dx.doi.org/10.2131/jts.23.3_173.

[14] Kobayashi K, Kanamori M, Ohhori K, Takeuchi H. A new decision tree method for statistical analysis of quantitative data obtained in toxicity studies on rodents. San Ei Shi. 2000; 42(4): 125–129.

[15] Sakaki H, Igarashi S, Ikeda T, et al. Statistical method appropriate for general toxicological studies in rats. J Toxicol Sci. 2000; 25(4 app): 71–81.

[16] Takizawa T, Igarashi T, Imamizo, H, et al. A study on the consistency between flagging by statistical tests and biological evaluation. Drug Information Journal. 2000; 34(2): 501–509.

[17] Kobayashi K, Pillai KS, Suzuki M, Wang J. Do we need to examine the quantitative data obtained from toxicity studies for both normality and homogeneity of variance? J Environ Biol. 2008; 29(1): 47–52.

[18] Finney DJ. Thoughts suggested by a recent paper: Questions on non-parametric analysis of quantitative data (Letter to editor). J Toxicol Sci. 1955; 20(2): 165–170.

[19] Chemical Substances Control Law (1986): http://www.safe.nite.go.jp/kasinn/genkou/kasinhou04.html.

[20] MHLW (2014): http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPage.jsp. Accessed July 31, 2014.

[21] Kobayashi K, Pillai KS, Sakuratani Y, Abe T, Kamata E, Hayashi M. Evaluation of statistical tools used in short-term repeated dose administration toxicity studies with rodents. J Toxicol Sci. 2008; 33(1): 97–104. http://dx.doi.org/10.2131/jts.33.97.

[22] OECD TG 407 OECD GUIDELINES FOR THE TESTING OF CHEMICALS, Repeated Dose 28-Day Oral Toxicity Study in Rodents. Adopted: 3 October 2008, http://ntp.niehs.nih.gov/iccvam/suppdocs/feddocs/oecd/oecdtg407-2008.pdf. Accessed July 31, 2014.

[23] NTP (2014) http://ntp.niehs.nih.gov/?objectid=D1512B41-F1F6-975E-7FBA3 D4A2132F1C1 and http://ntp.niehs.nih.gov/?objectid=D16D6C59-F1F6-975E-7D23D 1519B8CD7A5. Accessed July 31, 2014.

[24] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Am Stat Assoc. 1955; 50: 1096–1211. http://dx.doi.org/10.1080/01621459.1955.10501294.

[25] Williams DA. A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics. 1971; 27: 103–117. http://dx.doi.org/10.2307/2528930.

[26] Williams DA. The comparison of several dose levels with zero dose control. Biometrics. 1972; 28: 519–531. http://dx.doi.org/10.2307/2556164.

[27] Shirley EA. Non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. Biometrics 1977; 33: 386–389. http://dx.doi.org/10.2307/2529789.

[28] Dunn OJ. Multiple comparisons using rank sums. Technometrics 1964; 6: 106–107. http://dx.doi.org/10.1080/00401706.1964.10490181.

[29] Jonckheere A. A distribution-free k-sample test against ordered alternatives. Biometrika 1954; 41: 133–145. http://dx.doi.org/10.1093/biomet/41.1-2.133.

[30] Hollander M, Wolf DA. Nonparametric statistical methods. John Wiley and Sons, NY U.S.A., 1973; pp 120–123.

[31] Kobayashi K, Pillai KS, Guhatakurta S, Cherian KM, Ohnishi M. Statistical tools for analysing the data obtained from repeated dose toxicity studies with rodents. A comparison of the statistical tools used in Japan with that of used in other countries. J Environ Biol. 2011; 32(1): 11–16.

[32] Hagiwara A, Imai N, Numano T, et al. A twenty eight-day repeated dose toxicity study of black soybean extract in Sprague-Dawley rats. J Toxicol Sci. 2010; 35(1): 87–96. http://dx.doi.org/10.2131/jts.35.87.

[33] Kojima S, Sasaki J, Tomita M, et al. Multiple organ toxicity, including hypochromic anemia, following repeated dose oral administration of phenobarbital (PB) in rats. J Toxicol Sci. 2009; 34(5): 527–539. http://dx.doi.org/10.2131/jts.34.527.

[34] Honda K, Enoshima T, Oshikata T, et al. Toxicity studies of Asahi Kasei PI, purified phosphatidylinositol from soy lecithin. J Toxicol Sci. 2009; 34(3): 265–280. http://dx.doi.org/10.2131/jts.34.265.

[35] Tamano S, Yoshino H, Ichihara T, et al., 28-day oral toxicity of macrophomopsis gum in F344/DuCrj rats. Jpn J Food Chem 2005; 12(3): 128–134.