Fundamental Toxicological Sciences (Fundam. Toxicol. Sci.)
Vol.11, No.1, 1-10, 2024

The Japanese Society of Toxicology

Fundamental Toxicological Sciences
URL : http://www.fundtoxicolsci.org/index_e.html

Review

Repeated dose administration toxicity studies - Use of *t*-tests in multiplicity data analysis

Katsumi Kobayashi[1] and Kalathil Sadasivan Pillai[2]

[1]Ex-Cabinet Office, Food Safety Commission of Japan, Akasaka Park Bld, Tokyo, Japan
[2]PNB Vesper Life Science. Kochi, Kerala, India

(Received November 23, 2023; Accepted December 5, 2023)

ABSTRACT — In conducting repeated dose administration toxicity (RDAT) studies with rats and mice, a minimum of three dose groups and one control group are normally set for determing NOEL/NOAEL (no-observed effect level/no-observed adverse effect level) of the test item. For comparison of data among the groups, initially, the data are analysed by analysis of variance (ANOVA). If ANOVA shows a significant difference, then groups means are compared by a multiple comparison range test (MCRT). However, in RDAT studies, at the end of the long duration of the test substance administration, the distribution of the data obtained varies considerably among the groups and the number of animals decreases due to mortality/morbidity, especially in the high-dose groups. Increased variance in the distribution of the data and decreased animals in one or more groups may result in an insignificant ANOVA, though the low-dose group may show a marked difference compared to the control. Dunnett's multiple comparison test is commonly used to compare each treatment group with the control group. However, Dunnett's test has a lower ability to detect significant differences than the *t*-test, and its detection power decreases with the increase in the number of groups. Therefore, we recommend the *t*-test, by-passing ANOVA, which has a high detectable significant difference in the two-group test. In addition, the application of the *t*-test eliminates the need to select an MCRT. However, the final judgment of the adverse effects may be made based on the toxicological relevance in consideration of the statistical analysis results.

Key words: Repeated dose administration toxicity studies, Dunnett's multiple comparison test, *t*-test, Power of statistical analyses

INTRODUCTION

Four dose groups, including a control group, are normally used in repeated dose administration toxicity (RDAT) studies designed to determine no observed adverse effect level/no observed effect level (NOAEL/NOEL) and least toxic level (LOAEL) for pharmaceuticals, pesticides, and general chemicals. The NOAEL/NOEL is established based on statistically and toxicologically significant quantitative and qualitative parameters determined in the study. Until 1982, the *t*-tests were used to analyze the differences between the control group and each dose group for various quantitative values obtained from RDAT studies. However, since then to date, It is a common practice to analyse the data by ANOVA, if the number of groups is more than three (Gad, 1982). If ANOVA shows a significant difference, the groups are compared using multiple comparison range tests like Dunnett, Tukey, Williams, and Scheffé's. ANOVA may show insignificance due to large variance, especially in the high-dose groups (a decrease in the number of observations due to mortality is normally seen in

Correspondence: Katsumi Kobayashi (E-mail: student-kobayashi@beige.plala.or.jp)

the high-dose groups), though the low-dose group may show a significant difference compared to the control. The FDA (2005) and OECD (2010) test guidelines for RDAT studies define NOAEL as the highest dose level of the test item that does not produce a significant increase in adverse effects in comparison to the control group. Dunnett's test has been the preferred choice of statistical tool to compare each treatment group with the control in RDAT studies because it considers multiplicity. In this paper, we tried to explain why the pre-analysis of ANOVA should be avoided for comparing each dose group with the control group. We also attempted to discuss whether multiplicity is required to be considered for the analysis of quantitative data obtained from RDAT studies, based on actual data and several published papers. We propose the *t*-tests, which are more effective in analyzing the data obtained from RDAT studies than Dunnett's multiple comparison test.

MATERIALS AND METHODS

In this paper, we used data from short-term and long-term RDAT studies with rodents conducted at Anpyo Center Inc., Shizuoka, Japan, NIHS (National Institute of Health Sciences, Japan) database, and several publications. For the statistical analysis, SAS JMP software version 5.1 and Excel 2008 were used.

RESULTS

Purpose of repeated dose administration toxicity studies

The purpose of RDAT studies is to determine the dose that induces clear toxic changes, provide details of such changes, and also to determine the dose at which no toxic changes are observed when the test substance is repeatedly administered to animal models. In Japan, NOEL and NOAEL are often used interchangeably in RDAT studies. Moreover, the Chemical Substances Control Law (CSCL) of Japan does not provide a clear expression of NOEL and NOAEL. Generally, NOEL refers to the dose at which no statistically significant difference compared to the control is observed for each analyzed parameter (hematological, clinical chemistry, pathological, *etc*.). On the contrary, NOAEL is the dose level toxicologically determined based on extensive knowledge, experience, and background data of the particular substance, with reference to the results of statistical analysis.

We conducted a literature survey in the Japan Existing Chemical Database (JECDB) to understand whether NOEL or NOAEL was used as the criterion for judging the toxic dose in the 28-day RDAT studies and the combined repeated dose toxicity/reproductive and developmental toxicity studies. We observed that during the period, 1991–2001, the toxicity of a chemical was judged in most of the cases by NOEL, and from 2001 onwards, NOAEL alone or both NOAEL and NOEL were often used together for judging the toxicity. Our literature survey indicated during the period, 1991 to 2007, the number of studies in which NOEL and NOAEL were used was 574 and 5, respectively, 24 studies indicated both NOEL and NOAEL.

Guidelines for toxicity testing

The RDAT studies surveyed were conducted according to the standard guidelines for pharmaceuticals, agricultural chemicals, and general chemical substances (OECD, 2008).

Characteristics of toxicity tests

Number of dosing groups: The literature survey in (JECDB) showed that 120 studies were conducted with 4 groups, 33 studies with 5 groups, 4 studies with 6 groups, and 1 study with 7 groups. It may be mentioned here that the power of multiple comparison/range tests to detect significant differences decreases as the number of groups increases (Kobayashi, 2015a).

Table 1 shows how the power to detect significant differences decreases as the number of groups increases. Assuming that five groups were set, no significant difference was observed between the top-dose group and control group, and the high-dose group and control group, when analyzed by Dunnett's test, whereas, the *t*-test showed a significant difference. In multiple comparisons and range tests, as the number of groups increases, the power to detect a significant difference decreases due to the increase in the number of comparison combinations. Therefore, when RDAT studies are conducted with more than four groups, to find a significant difference between the dose groups and the control group, we recommend using the *t*-test, rather than Dunnett’s multiple comparison test.

Common ratio of doses: We present our observations made on the investigation of the common ratio of doses set in 28-day repeated-dose toxicity tests done with 124 existing compounds using rats under the CSCL (JECDB, 2023). The common ratios used were 2, 2.5, 3, 4, 5, 6, and 7 were for 7, 2, 54, 17, 37, 3, and 4 tests, respectively. The most frequently set common ratio was three. The majority of the studies set a high dose of 1000, a mid of 300, and a low dose of 100 mg/kg. A common ratio of 10 was set for a few pesticide toxicity tests, which is

Table 1. Decrease in the power of Dunnett's test to show a significant difference between the dose groups and control group as the number of group increases. Hemoglobin data (g/dL) of 72-week-old B6C3F1 male mice.

No. of groups	Statistical analysis	Dose level				
		Control	Low dose	Mid dose	High dose	Top dose
5	Number of animals	10	10	10	10	10
	Mean ± S.D.	13.9 ± 0.254	13.9 ± 0.503	13.9 ± 0.267	14.2 ± 0.179	14.2 ± 0.279
4	Levene's test[#]			$p = 0.1609$		
	Dunnett's test[##]	-	$p = 0.8185$	$p = 0.6412$	$p = 0.0497$*	
5	Levene's test[#]			$p = 0.2307$		
	Dunnett's test[##]	-	$p = 0.8621$	$p = 0.6955$	$p = 0.0532$	$p = 0.0616$
	Student's *t*-test[##]	-	$p = 0.4341$	$p = 0.3678$	$p = 0.0027$**	$p = 0.0108$*

One-sided test for Dunnett's and Student's *t*-tests.
[#]: Homogeneity variance test. [##]: Comparison with control group.
* $p < 0.05$ and ** $p < 0.01$ from control group.

Table 2. Changes in the number of surviving F344 male rats treated with drugs (dietary administration) in 10 carcinogenicity studies.

Dose level	Number of surviving F344 male rats during the dosing period (weeks)									
	0	52	58	65	71	78	84	91	97	104
Control	50	50 ± 1	49 ± 1	49 ± 1	49 ± 1	48 ± 1	47 ± 1	45 ± 2	44 ± 2	39 ± 5
Low	50	50 ± 0	50 ± 0	49 ± 1	49 ± 1	48 ± 1	47 ± 2	46 ± 3	44 ± 4	41 ± 3
Mid	50	50 ± 0	50 ± 0	50 ± 1	49 ± 1	49 ± 1	48 ± 1	47 ± 2	44 ± 2	40 ± 3
High	50	50 ± 0	49 ± 1	49 ± 1	49 ± 1	46 ± 4	43 ± 8	39 ± 11	35 ± 14	26 ± 16

Values are expressed as mean ± S.D.

too high compared to the common ratio set for pharmaceutical tests and as per the CSCL. We also observed that some tests did not use constant common ratios.

Changes in statistical significance with the reduction in the number of animals and changes in variance in the highest dose group: In the RDAT studies, especially those of >90 days duration, the occurrence of mortality is a normal phenomenon in the higher dose groups, which distorts the mean values, and variance of the measured parameter. When the change with regard to the control group is about ± 20% in the top/high dose group, the variance ratio between the control group and top/high dose group would be ± 3.5 times in body weight, food consumption, hematology parameters, and organ weights.

Changes in the number of samples: The number of surviving F344 male rats in ten 104-week carcinogenicity studies with drugs is given in Table 2. The studies were initiated with 50 rats/group. Until around 78 weeks after the initiation of the studies, the mean number of surviving rats in the control group, low, mid, and high dose groups was similar (46 to 49). From week 78, the mean number of surviving rats declined gradually in all the groups, and the decline was more in the high-dose group. In the high-dose group, at week 104 the mean number of surviving rats was 26 (with a minimum of four in one study), while it was 39 to 41 in the other groups. From 78 weeks onwards variance of the high dose group increased as evidenced by the increased standard deviation. In the other groups, the changes in variance that occurred were minimal.

Changes in absolute organ weight variance, with respect to the control group were determined for several organs at weeks 26, 52, 78, and 104 during the dosing period (Table 3). The variance ratio (*F* values) of absolute weights of adrenals and kidneys inflated at weeks 78 and 104.

Selecting a test of homogeneity of variance when the number of groups is three or more

For testing homogeneity of variance, four types of equal-variance tests are used (Kobayashi, 2015b; Kobayashi and Pillai, 2013). Among these tests, Bartlett's homogeneity test is most commonly used (Kuwagata *et al*., 2023), followed by the Levene, Brown-Foresythe, and O 'Brien's tests. This homogeneity of variance tests typically determine significance at the 5% level. To give formal approval of the difference in the variance of multiple groups (more than three groups), Yoshimura (1987) recommended 10 or more animals in a group. However, in JECDB (TG407, 28-day RDAT studies), most tests

Table 3. Changes in variance ratio (*F* values) of absolute organ weight of F344 male rats treated with drugs (dietary administration) in 10 carcinogenicity studies.

Organs and dose level		On weeks during the dosing period			
		26[+]	52	78	104
Brain	Low	2.5 ± 1.4	3.0 ± 1.6	5.2 ± 10.9	1.7 ± 0.7
	Mid	1.9 ± 0.7	2.1 ± 1.2	1.9 ± 0.5	1.9 ± 1.0
	High	3.5 ± 2.4	2.3 ± 1.1	1.7 ± 0.4	7.4 ± 16.8#
Liver	Low	3.5 ± 3.9	2.1 ± 0.7	9.1± 14.7	2.3 ± 1.2
	Mid	1.8 ± 0.9	1.8 ± 0.8	5.4 ± 6.1	1.9 ± 0.6
	High	1.9 ± 0.5	2.3 ± 0.8	11.8 ± 24.4	1.6 ± 0.5#
Kidneys	Low	2.3 ± 1.0	2.1 ± 1.7	3.0± 4.0	54.6 ± 166
	Mid	2.9 ± 1.3	2.0 ± 0.9	5.7 ± 9.1	13.3 ± 34.4
	High	2.1 ± 0.6	2.6 ± 2.6	5.0 ± 3.7	148 ± 408#
Adrenal glands	Low	8.0 ± 13.0	1.7 ± 1.0	8.7 ± 14.5	330 ± 996
	Mid	2.1 ± 1.2	2.7 ± 1.6	11.8 ± 22.4	210 ± 638
	High	1.9 ± 0.6	2.9 ± 2.0	91.5 ± 251	646 ± 1785#
Testes	Low	1.6 ± 0.6	6.4 ± 9.9	2.6 ± 1.7	2.0 ± 1.1
	Mid	2.9 ± 2.3	3.7 ± 4.3	2.1 ± 1.0	1.4 ± 0.6
	High	3.6 ± 3.1	2.7 ± 1.6	1.7 ± 0.5	1.7 ± 0.4#

Variance ratio was calculated with respect to the respective control group.
Values are expressed as mean ± S.D. [+]: Five studies; #: Eight studies.

Table 4. Number of significant ($p < 0.05$) differences detected when analysed by Dunnett's and *t*-tests in a combined 2-year chronic toxicity and carcinogenicity study with 5 groups including a control group.

Parameter	No. of statistical analyses[#]	Dunnett's test		*t*-test	
		One-sided (α)	Two-sided (2α)	One-sided (α)	Two-sided (2α)
Body weight (b.w.)	528	223	212	246	233
Food consumption	832	235	189	349	279
Hematology	352	123	105	159	126
Blood chemistry	576	215	181	272	235
Urinalysis	64	7	5	11	10
Organ weight	224	47	42	80	61
Organ weight/b.w.	224	82	67	104	89
Total	2800	932 (100)*	801 (100)	1221 (131)	1033 (129)

()*: In % of Dunnett's test. [#] = measurement time points × 2 (sexes) × 4 (number of comparisons). For example, body weight was measured 66 times during the course of the study. Number of statistical analysis done for body weight = 66 × 2 × 4 = 528.

considered five animals/sex per group.

According to Finney (1995) and Kobayashi *et al.* (1999), Bartlett's test is highly sensitive to the homogeneity of variance. In RDAT studies, the non-uniformity of dispersion of several parameters is a normal phenomenon. Recent papers (Moroki *et al*., 2023) use Dunnett's test directly without performing the test of homogeneity of variance. If the Aspin-Welch *t*-test is used instead of the multiple comparison test, there is no need for the test of homogeneity of variance and the analysis of variance (ANOVA). However, for performing an equal variance test, we recommend Leven's homogeneity test (Levene, 1960), the sensitivity of which to the homogeneity of variance is not as high as that of Bartlett's test. Risom *et al*. (2003), Shibui *et al*. (2014), Tanaka *et al.* (2019) and Masubuchi *et al*. (2020) conducted toxicity or pharmacological studies using Levene's test. The authors of the present paper do not suggest Bartlett's test in the case of five animals per group.

Multiplicity issues

Multiple testing in RDAT studies can have an impact on Type I and Type II error rates (Li *et al.*, 2017). The multiple comparison/range tests that are commonly used in RDAT studies are Dunnett, Tukey, Williams, and Scheffé's. Among these, Dunnett's test has the highest detection power and is one of the preferred statistical tools for examining differences between the con-

Table 5. Comparison of day-27 body weights (g) of CD (SD) IGS female rats exposed to different doses of a test substance with control group (0 mg/kg) using Dunnett's and *t*-tests.

Statistical analysis	Dose groups				
	0 mg/kg	1 mg/kg	3 mg/kg	10 mg/kg	30 mg/kg
Body weight	205, 200	208	224	209	181, 211
	213, 211	199	237	215	175, 201
	248, 221	210	225	229	196, 174
	187, 200	204	224	226	189, 207
	190, 196	207	179	227	201, 153
No. of animals	10	5	5	5	10
Mean ± S.D. (%)	207 ± 18 (100)	206 ± 4 (99.5)	218 ± 22 (105)	221 ± 9 (107)	189 ± 18 (91.3)
Bartlett 's test			$p = 0.2001$		
ANOVA			$p = 0.0063$**		
Dunnett's test	-	$p = 0.9996$	$p = 0.6197$	$p = 0.3766$	$p = 0.0660$
Homogeneity by *F* test	-	$p = 0.0147$	$p = 0.5185$	$p = 0.1844$	$p = 0.9658$
t-test†	-	$p = 0.4026^{\#}$	$p = 0.1650$	$P = 0.0609$	$p = 0.0346$*

†: By Student's *t*-test; #by Aspin-Welch's *t*-test.
* $p < 0.05$ and ** $p < 0.01$ from control group or among the groups.

Table 6. Classification of number of studies conducted during the period 1985–2004 in Japan 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 7. Minimum number of animals required within a group for obtaining a significant difference by rank sum tests.

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

trol group and each dose group in RDAT studies, and is incorporated into the statistical decision tree for toxicology (OECD, 2010). The multiple comparison tests use the variance of the error term of the analysis of variance. This means, for example, that variance in the mid and high-dose groups are also used for the comparison of the low-dose group with the control. Therefore, a large variance and a decrease in the number of observations due to death in the high-dose group affect the statistical analysis results of the other dose groups. Therefore, for comparing the low-dose group with the control, using the variance in other groups (*e.g.*, mid-dose and high-dose groups) in the calculation procedure is contentious.

In a study with five groups, a significant difference was detected in 932 (one-sided) and 801 (two-sided) analyses by Dunnett's test, and in 1221 (one-sided) and 1033 (two-sided) by the *t*-tests, indicating that a significant difference was detected at a 29 (two-sided) to 31 (one-sided) % higher frequency by the *t*-test than by the Dunnett's test (Table 4, Kobayashi and Pillai, 2003). Over 30 years ago, at the 55th annual meeting (1993) of the Japanese Society for Pharmaceutical Statistics, the choice of statistical analysis method for RDAT studies was discussed. Dr. Yoshimura recommended the use of multiple comparison tests from the standpoint of evaluation of toxicity, and Dr. Tsubaki recommended the use of sensitive *t*-tests from the standpoint of consumers.

Differences in power between Dunnett and *t*-tests to detect a significant difference in body weight in a 28-day RDAT study (partially modified) conducted in rats under the Japanese CSCL are given in Table 5. A significant difference was observed in the analysis of variance (ANOVA), but Dunnett's test showed no significant difference between any dose group and control group. However, the *t*-test showed a significant difference in the top dose group (30 mg/kg b.w.), compared with the control group.

History of multiple comparison tests using analysis of variance as the first step of analysis

Several statistical methods have been proposed to analyze the data obtained from toxicity, pharmacology, and efficacy studies with three or more groups, including a control group. Usually, the statistical analysis is done to compare each dose group with the control group or among all groups. Differences among the groups are analyzed by ANOVA and the variance of the error term calculated from the ANOVA is used to compare each dose group with the control group. Fisher first published ANOVA in 1921. Thereafter, Tukey in 1949, Scheffé in 1953, Dunnett, and Duncan in 1955, and Williams in 1971 discussed ANOVA in their publications.

ANOVA (one-way ANOVA): Sir Ronald Aylmer Fisher (1980–1962) was a British statistician, geneticist, and evolutionary biologist who worked at the UK agricultural research institution for 14 years, thereafter worked at several universities and research institutes. He described ANOVA under the title "On the "Probable Error" of a coefficient of correlation deduced from a small sample" (Fisher, 1921). ANOVA provides fundamental numerical values required for multiple comparison and/or range tests involving three or more than three groups. It also provides information on the distribution state of data of all the groups.

Tukey's test: John W. Tukey (1915–2000) was an American mathematician and statistician, who published a multiple comparison/range test under the title "Comparing individual means in the analysis of variance" (Tukey, 1949). In the test example, Fisher's paper was cited and the yield of potatoes under six types of fertilizers was analyzed. In this example, arbitrary groups were selected and analyzed. Tukey analyzed the difference between the largest average value and the smallest average value and stated that if the difference is insignificant, testing the difference between other groups was unnecessary. Tukey's test can be applied to all pairwise comparisons. One of the requirements of this test is that the number of animals in each group should be the same.

Sugimoto *et al*. (2020) used Tukey's test for analysing the data from repeated 28-day and 13-week dose toxicity studies of oils prepared from the internal organs of the Japanese giant scallop in rats. The groups consisted of two concentrations each of scallop oil-A, scallop oil-B, tuna oil (existing oil), and a control group, a total of seven groups. The authors compared all pairs using Tukey's test.

Scheffé's test: Henry Scheffé (1907–1977) was a mathematician of the United States (Columbia University). The multiple comparison test published by Scheffé (1953)

performs simultaneous, joint pairwise comparisons for all possible pairwise combinations of each group mean. This test applies to a series of estimates of all possible contrasts between averages. The test is an extension of ANOVA, therefore, it is considered that calculation is possible even if the number of animals in each group is different. Since there are many combinations in the multiple comparison/range test, the power of Scheffé's test is low. A statistically significant difference at $p = 0.05$ may not be observed even if the mean value of a treatment group shows a 30% difference compared to the control group. The test is used relatively widely in the field of agriculture.

The use of Scheffé's test is not advisable in RDAT studies due to its low power. Hirata (2012) investigated the history of multiple comparison tests, where a description of Scheffé's test is given (Yamazaki *et al.*, 1981). For RDAT studies, earlier decision tree for the statistical analysis of the data included a Scheffé's test, and the Japanese registered examiner pointed out that the data should be reanalyzed using a *t*-test also (Hirata, 2012). The authors of this paper are concerned about the use of Scheffé's test in RDAT studies, the power of which is low to detect a significant difference, and also reanalyzing the data using the *t*-test.

Dunnett's test: Canadian statistician, Charles W. Dunnett (1921–2007), a professor of the departments of mathematics, statistics, clinical epidemiology, and biostatistics of the McMaster University, published this test (Dunnett, 1955). The Dunnett test is used in testing two or more experimental groups against a single control (Lee and Lee, 2018).

Dunnett's original paper (Dunnett, 1955) is presented below: Case 1 study consisted of five groups, including a control group, with three measurements in each group. In this study, the breaking strength is analysed when three different chemical substances are treated in cloth. In the case 2 study, the influence of species difference on erythrocytes counts is compared with the control group by administration of two drugs. The number of individuals in Drug A and B groups varies according to the missing values (accidental losses). The number of animals in each group ranged from six to four in the erythrocyte count. Yamazaki *et al*. (1981) was the first Japanese author to adopt Dunnett's test in the decision tree. Kobayashi (1983) published Dunnett's calculation example in Japanese using actual data. The results are similar to those of the *t*-test when the two-group setting is calculated with Dunnett's test. The Dunnett's test is an extended version of the *t*-test. ANOVA F-test is not recommended before performing Dunnett's comparison test against control (Hothorn, 2016).

Duncan's test: David B. Duncan, a member of Virginia Polytechnic Institute and State University used this test for analyzing the data gathered from a barley harvest (barley grain yield per acre). For setting seven groups (A–G), 21 calculation steps were required (Duncan, 1955).

Williams's test: Williams (1971) published a test for comparing treatment means with a control mean. The test is generally used to compare multiple dose groups with a control group assuming a monotonic dose-response relationship. In this test, if there is no significant difference between the mean value of the high/top dose group and the control mean, it is considered difference between the mean values of other treatment groups and the control mean is insignificant, even if a significant difference is observed at the lowest dose. The use of William's test is not recommended when the number of animals in the group is different (Williams, 1972) and extremely less (Williams, 1971 and 1972). However, Sakaki *et al*. (2000) stated that Williams's test can be used even if number of the animals in a group differs about two times compared to other group/s.

Among the multiple comparison tests, the Dunnett test has more power followed by Williams, Duncan, and Tukey's test rank in order of power. The Dunnett and Williams's tests are for comparing each treatment group with the control group, whereas the Duncan and Tukey tests are an all-pairs comparison test. Scheffé's test has the lowest power due to the arbitrary number of possible combinations. It may be noted that the examples given by the authors of the above tests in their original paper are not data related to RDAT studies.

t-test used to test the difference between two groups

Three types of *t*-tests are commonly used, depending on the size of the variance ratio and the number of samples in each group. For equal variances, Student, for unequal variances, Aspin-Welch, and for unequal variances with different sample size, Cochran-Cox *t*-tests are used. Dunnett's test is also called Dunnett's *t*-test. Since the number of groups in RDAT studies is usually more than three with concurrent control group, the use of the *t*-test for comparing two groups has become less common. In addition, in the case of a two-group setting, the test for unequal variance with different sample sizes is often ignored and the Aspin-Welch *t*-test is used. Student's *t*-test (Student, 1908) was published in 1908 by Gosset, under the pseudonym, Student.

History of statistical analysis methods using organisms with three or more groups in Japan

From 1960 to 1975, most of the scientific publications on toxicology did not clearly specify the statistical analysis method or conduct any statistical analysis. However, a few papers stated the statistical results in a table, but the statistical analysis method used was unclear. In Japan, multi-group data were first analyzed in 1976 by Maita *et al*. under the title "Long-term feeding test of Sanpoly-305 in rats".

History of statistical analysis methods using organisms with three or more groups in other countries

Petering *et al*. (1967) analyzed body weight and tumor diameter data of six groups of rats (10 rats/group) using Student's *t*-test. In the same year, Wexler *et al*. (1967) analyzed body weight, serum biochemical values, and organ weights of seven groups (9 to 33 rats/group) using ANOVA.

Analysis of continuous variables of NTP (National Toxicology Program, USA) technical report on the toxicology and carcinogenesis studies

The methods used for analyzing data obtained from 106 short-term toxicity studies and 602 long-term carcinogenicity/toxicity studies conducted on chemical substances published in NTP technical reports in 2023 were examined. For analyzing the organ and body weight data, the NTP technical report series (NTP, 2023) used Dunnett (1955) and Williams (1971 and 1972) parametric multiple comparison tests. For nonparametric multiple comparisons of hematology, clinical chemistry, spermatid, and epididymal spermatozoa Shirley's (1977) and Dunn's (1964) tests were used.

Statistical data analysis of RDAT studies according to OECD guidelines

The OECD guidelines TG407 for repeated dose 28-day oral toxicity study in rodents (OECD, 2008) and combined repeated dose toxicity study with the reproduction/developmental toxicity screening test, TG422 (OECD, 2015) do not recommend comparisons of the effect along a dose range using multiple *t*-tests. The guidelines for repeated dose 90-day oral toxicity study in rodents (TG408, OECD, 2018a) and chronic toxicity studies (TG452, OECD, 2018b) recommend that the data should be evaluated by an appropriate and generally acceptable statistical method, and the statistical methods should be selected during the design of the study.

The data of the recovery groups (top dose and control) in the RDAT studies are analyzed by either the *t*-test or Dunnett's test. In analyzing two groups, the results obtained by Dunnett's test are identical to those obtained by the *t*-test. This is because Dunnett's test is an extension of the *t*-test.

Difference in the use of statistical tools for analyzing data obtained from 28-day repeated dose toxicity studies in various test facilities 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 accessed from JECDB (2023) were examined. The studies were conducted following the guidelines of the CSCL. Most of the studies used Bartlett's homogeneity test, ANOVA, Dunnett's test, and Scheffé's test. Student's or Aspin-Welch's *t*-test and Mann-Whitney's *U* test were used for analyzing the data of recovery groups (Table 6).

Changes in the *t*-test and Dunnett's test by the Anpyo Center, Japan (a contract testing facility)

During 1980–1983, among the statistical tools used for the analysis of data obtained from RDAT studies with pesticides, pharmaceuticals, and new chemical substances, more than 80% was *t*-test. But in 1992, the usage rate of the *t*-test decreased to 30%, and the usage of Dunnett's test of the analysis of variance system increased to s 70%. Since 2000, Dunnett's test is more commonly used.

Nonparametric multiple comparison tests and two-group tests

Currently, in RDAT studies, where three or more groups are used, if Bartlett's test shows unequal variances, Steel's test, which considers multiplicity, is commonly used. However, the Mann-Whitney's *U* test (comparing two groups) is used in about 10% of toxicity studies. When using these tests, consideration should be given to the number of animals in groups. Table 7 shows a comparison of the power of nonparametric tests in multiplicity analysis (except for Mann-Whitney's *U* test, which is shown for reference, the remaining are multiple comparison tests). These nonparametric tests convert all individual values into ranks and analyze the ranks of their average values. Steel's test and Mann-Whitney's *U* have about the same power.

DISCUSSION

RDAT studies are conducted to determine the adverse

effects of the test item on the test organism. Normally these studies are conducted with a minimum 3 treatment groups and a control group. The dose levels of the test item for the treatment groups are selected in such a way as to estimate NOEL/NOAEL of the test item for that particular test organism. FDA (2005) defines NOAEL as the highest dose level of the test item that does not produce a significant increase in adverse effects in comparison to the control group. The definition further states that any biologically significant effect is considered an adverse effect, which may or may not be statistically significant. OECD guidelines 407 (OECD, 2008) and 422 (OECD, 2015) define NOAEL as the highest dose level where no adverse treatment-related findings are observed. But the OECD Guidance document (OECD, 2010) defines NOEL as the highest dose level where there is no significant increase in treatment-related effects compared with the negative/vehicle control. The document explains that the terminology NOAEL is used to distinguish between changes that are adverse rather than any treatment-related effect which may in some cases not be adverse. Further, OECD test guideline 407 for 28-day RDAT studies recommends not to use multiple *t*-tests for *c*omparisons of the effect along a dose range (OECD, 2008), whereas NTP (2023) recommends multiple comparison procedures of Dunnett (1955) and Williams (1971 and 1972).

Classical textbooks and most of the decision trees prescribe conducting ANOVA before multiple comparisons. Comparisons among the groups are made only when ANOVA shows a significant difference A judgment on a significant difference becomes conservative if the results are judged based on both ANOVA and multiple comparisons (Hamada, 2018). In RDAT studies, at the end of the long duration of the test substance administration, the distribution of the data obtained varies considerably among the groups and the number of animals decreases, especially in the high-dose groups. Increased variance in the distribution of the data and decreased animals in the high-dose groups may result in an insignificant ANOVA, though the low-dose group may show a marked difference compared to the control. Therefore, we recommend the *t*-test, by-passing ANOVA, which has a high detectable significant difference in the two-group test. In addition, the application of the *t*-test eliminates the need to select methods for the multiple comparison test (post-hoc analysis). However, the final judgment of the adverse effect may be made based on the toxicological relevance in consideration of the statistical analysis results.

Conflict of interest---- The authors declare that there is no conflict of interest.

REFERENCES

Duncan, D.B. (1955): Multiple Range and Multiple F-Test. Biometrics, **11**, 1-5.

Dunn, O.J. (1964): Multiple comparisons using rank sums. Technometrics, **6**, 106-107.

Dunnett, C.W. (1955): A multiple comparison procedure for comparing several treatments with a control. Am. Stat. Assoc., **50**, 1096-1211.

FDA. (2005): Guidance for industry estimating the maximum safe starting dose in initial clinical trials for therapeutics in adult healthy volunteers, U.S. department of health and human services food and drug administration center for drug evaluation and research (CDER), Rockville, USA.

Finney, D.J. (1995): Thoughts suggested by a recent paper: questions on non-parametric analysis of quantitative data (letter to the editor). J. Toxicol. Sci., **20**, 165-170.

Fisher, R.A. (1921): On the "probable error" of a coefficient of correlation deduced from a small sample. Metron, **1**, 3-32.

Gad, S.C. (1982): Statistic for toxicologist. principles and methods of toxicology (Hayes A. H., ed.), pp. 276, Raven Press, New York.

Hamada, C. (2018): Statistical analysis for toxicity studies. J. Toxicol. Pathol., **31**, 15-22.

Hirata, A. (2012): Multiple comparison test –looking back on the past–. Japanese Society for Biopharmaceutical Statistics, 10th regular meeting, June 2, 2012. (in Japanese)

Hothorn, L.A. (2016): The two-step approach–a significant ANOVA F-test before Dunnett's comparisons against a control– is not recommended. Commun. Stat. Theory Methods, **45**, 3332-3343.

JECDB (Japan Existing Chemical Database): National Institute of Health Sciences (nihs.go.jp) Accessed on August 27, 2023.

Kobayashi, K. (1983): Dunnett's multiple comparison test. Japanese society for biopharmaceutical statistics, No. 10, 11-15. (in Japanese)

Kobayashi, K. (2015a): Statistical analysis methods for toxicological studies 2015, pp. 93–94, Yakuji Nippo, Limited, Tokyo.

Kobayashi, K. (2015b): Statistical analysis methods for toxicological studies 2015, pp. 57, Yakuji Nippo, Limited, Tokyo.

Kobayashi, K. and Pillai, K.S. (2003): Applied statistics in toxicology and pharmacology. pp. 29, Science Publishers, Inc., New Hampshire.

Kobayashi, K. and Pillai, K.S. (2013): A handbook of applied Statistics in pharmacology. pp. 34-34, CRS Press, New York.

Kobayashi, K., Kitajima, S., Miura, D., Inoue, H., Ohori, K., Takeuchi, H. and Takasaki, K. (1999): Characteristics of quantitative data obtained in toxicity rodents –The necessity of Bartlett's test for homogeneity of variance to introduce a rank test–. J. Environ. Biol., **20**, 37-48.

Kuwagata, M., Tsuboi, M., Igarashi, T., Tsurumoto, M., Nishimura, T., Taquahashi, Y. and Kitajima, S. (2023): A 90-day dose toxicity study of 2-(2H-benzotriazol-2-yl)-6-dodecyl-4-methylphenol in rats. Fundam. Toxicol. Sci., **10**, 59-68.

Lee, S. and Lee, D.K. (2018): What is the proper way to apply the multiple comparison test? Korean J Anesthesiol., **71**, 353-360. Erratum in (2020). Korean J. Anesthesiol., **73**, 572.

Levene, H. (1960): Robust tests for equality of variances. In Olkin, I., Ghurye, G., Hoeffding, W., Madow, W.G. and Mann, H.B. (eds.), Contributions to probability and statistics: Stanford

University Press, Stanford, California, 278-292.
Li, G., Taljaard, M., Van den Heuvel, E.R., Levine, M.A., Cook, D.J., Wells, G.A., Devereaux, P.J. and Thabane, L. (2017): An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int. J. Epidemiol., **46**, 746-755.
Maita, K., Masuda, H. and Suzuki, Y. (1976): Long-term feeding test of Sanpoly-305 in rats. J. Toxicol. Sci., **1**, 39-49. (in Japanese)
Masubuchi, Y., Kikuchi, S., Okano, H., Takahashi, Y., Takashima, K., Ojiro, R., Tang, Q., Yoshida, T., Koyanagi, M., Maronpot, R.R., Hayashi, S. and Shibutani, M. (2020): Lack of combined effect of continuous exposure to α-glycosyl isoquercitrin from fetal stages to adulthood and voluntary exercise or environmental enrichment on learning and behaviors in rats. Fundam. Toxicol. Sci., **7**, 241-248.
Moroki, T., Akizawa, F., Kondo, S., Fujiwara, S., Yoshikawa, T. and Inoue, Y. (2023): Toxicological effects of repeated subcutaneous administration of corn oil for 4 weeks in rats. Fundam. Toxicol. Sci., **10**, 168-178.
NTP (National Toxicology Program) technical report series (2023): TR-602: Isomeric mixture of tris (chloropropyl) phosphate administered in feed to Sprague Dawley (Hsd: Sprague Dawley SD) rats and B6C3F1/N Mice (nih.gov).
OECD. (2008): Test No. 407: Repeated dose 28-day oral toxicity study in rodents, OECD guidelines for the testing of chemicals, Section 4, OECD publishing, Paris.
OECD (2010): OECD guidance document for the design and conduct of chronic toxicity and carcinogenicity studies, Supporting TG451, 452 and 453. Section 4: Statistical and dose response analysis, Including benchmark dose and linear extrapolation, NOAELS and NOELS, LOAELS and LOELS.
OECD. (2015): Test No. 422: Combined repeated dose toxicity study with the reproduction/developmental toxicity screening test, OECD guidelines for the testing of chemicals, Section 4, OECD publishing, Paris.
OECD. (2018a): Test No. 408: Repeated dose 90-day oral toxicity study in rodents, OECD guidelines for the testing of chemicals, Section 4, OECD publishing, Paris.
OECD. (2018b): Test No. 452: Chronic toxicity studies, OECD guidelines for the testing of chemicals, Section 4, OECD publishing, Paris.
Petering, H.G., Buskirk, H.H. and Crim, J.A. (1967): The effect of dietary mineral supplements of the rat on the antitumor activity of 3-ethoxy-2-oxobutyraldehyde bis(thiosemicarbazone). Cancer Res., **27**, 1115-1121.
Risom, L., Dybdahl, M., Bornholdt, J., Vogel, U., Wallin, H., Møller, P. and Loft, S. (2003): Oxidative DNA damage and defence gene expression in the mouse lung after short-term exposure to diesel exhaust particles by inhalation. Carcinogenesis, **24**, 1847-1852.
Sakaki, H., Igarashi, S., Ikeda, T., Imamizo, K., Omichi, T., Kadota, M., Kawaguchi, T., Takizawa, T., Tsukamoto, O., Terai, K., Tozuka, K., Hirata, J., Handa, J., Mizuma, H., Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., **25**, 71-98.
Scheffé, H. (1953): A method for judging all contrasts in the analysis of variance. Biometrika, **40**, 87-104.
Shirley, E. (1977): A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. Biometrics, **33**, 386-389.
Shibui, Y., Miwa, T., Kodama, T. and Gonsho, A. (2014): 28-day dietary toxicity study of L-phenylalanine in rats. Fundam. Toxicol. Sci., **1**, 29-38.
Student (1908): The probable error of a mean. Biometrika, **6**, 1-25.
Sugimoto, K., Shimizu, E., Hagihara, N., Hosomi, R., Fukunaga, K., Yoshida, M., Yoshioka, T. and Takahashi, K. (2020): Repeated 28-day and 13-week dose toxicity studies of oils prepared from the internal organs of the Japanese giant scallop (*Patinopecten yessoensis*) in rats. Fundam. Toxicol. Sci., **7**, 177-188.
Tanaka, T., Masubuchi, Y., Okada, R., Nakajima, K., Nakamura, K., Masuda, S., Nakahara, J., Maronpot, R.R., Yoshida, T., Koyanagi, M., Hayashi, S.M. and Shibutani, M. (2019): Ameliorating effect of postweaning exposure to antioxidant on disruption of hippocampal neurogenesis induced by developmental hypothyroidism in rats. J. Toxicol. Sci., **44**, 357-372.
Tukey, J.W. (1949): Comparing individual means in the analysis of variance. Biometrics, **5**, 99-114.
Wexler, B.C., Kittinger, G.W. and Judd, J.T. (1967): Responses to drug-induced myocardial necrosis in rats with various degrees of arteriosclerosis. Circ. Res., **20**, 78-87.
Williams, D.A. (1971): A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics, **27**, 103-117.
Williams, D.A. (1972): The comparison of several dose levels with a zero dose control. Biometrics, **28**, 519-531.
Yamazaki, M., Noguchi, Y., Tanda, M. and Shintani, S. (1981): Statistical method appropriates for general toxicological studies in rats. J. Takeda Res. Lab., **40**, 163-187. (in Japanese)
Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Press, pp. 45-46, Tokyo. (in Japanese)