Analysis of Statistical Methods Currently used in Toxicology Journals

Jihye Na, Hyeri Yang, SeungJin Bae and Kyung-Min Lim

College of Pharmacy, Ewha Womans University, Seoul, Korea

Statistical methods are frequently used in toxicology, yet it is not clear whether the methods employed by the studies are used consistently and conducted based on sound statistical grounds. The purpose of this paper is to describe statistical methods used in top toxicology journals. More specifically, we sampled 30 papers published in 2014 from *Toxicology and Applied Pharmacology, Archives of Toxicology, and Toxicological Science* and described methodologies used to provide descriptive and inferential statistics. One hundred thirteen endpoints were observed in those 30 papers, and most studies had sample size less than 10, with the median and the mode being 6 and 3 & 6, respectively. Mean (105/113, 93%) was dominantly used to measure central tendency, and standard error of the mean (64/113, 57%) and standard deviation (39/113, 34%) were used to measure dispersion, while few studies provide justifications regarding why the methods being selected. Inferential statistics were frequently conducted (93/113, 82%), with one-way ANOVA being most popular (52/93, 56%), yet few studies conducted either normality or equal variance test. These results suggest that more consistent and appropriate use of statistical method is necessary which may enhance the role of toxicology in public health.

Key words: Toxicology, Biostatistics, Standard deviation, Standard error of mean, Inferential statistics, Descriptive statistics

INTRODUCTION

Inappropriate use of statistical methods are frequently witnessed in clinical and biology journals (1-4). This issue is especially of critical importance in toxicology, since inferences drawn from statistics are frequently referred for enacting regulation, establishing guidance or determination of drug safety. However, in many published papers in toxicology journals, sample size and statistical methods have been frequently selected without proper justification.

In clinical studies, some guidelines regarding the sample size has been referred (either 25 or 30 as a cut-off for being large sample, (5,6)), yet no obvious guidance has been provided in the area of experimental statistics, to our best knowledge. Granting that unlike clinical studies where participants diverge greatly, the baseline characteristics of the experimental animals (inbred or outbred) or (cultured or mono-clonal) cells are tightly controlled and homogeneous; however, small sample size still raise concerns in cases of employing parametric methods for inferential statistics to determine group differences.

Data are usually described and represented by measuring the central tendency and the dispersion of the data; although mean (average of variables), median (value located in middle of the list of variables), and mode (most often occurring value) are all valid measures of central tendency, one becomes more appropriate to use than others, based on the skewedness of the distribution of variables. Mean, the most popular measure to central tendency, is especially vulnerable to outliers, thus median might be a better measure for central tendency.

In the description of data dispersion, standard deviation (SD) and standard error mean (SEM) are often confusedly used (2-4), even though their logical purposes are clearly different. SD is a measure of variability of individual values, whereas SEM represents the precision of sample means. SD should be used to describe the variability of a parameter (3) whereas SEM should be used for hypothesis testing for sample mean. Yet, since $SEM = SD/\sqrt{sample\ size}$,

Correspondence to: Kyung-Min Lim, College of Pharmacy, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 120-808, Korea
E-mail: kmlim@ewha.ac.kr
SeungJin Bae,
E-mail: sjbae@ewha.ac.kr

SEM drops as sample size increases, and accordingly, the size of SEM is usually less than SD making the data appear more reproducible and consistent. Thus, using SEM could underestimate the variability of individual parameter (7).

When inferential statistics are conducted, a null hypothesis is tested and the hypothesis is either rejected or failed to be rejected based on test statistics. At this point, if assumptions of normality are not met (for example, data are severely skewed or sample sizes are small) then the mean and SD of the data do not properly represent either central tendency or the dispersion of the data (6). Hence, non-parametric approach, which makes fewer assumptions, although less efficient, should be considered. However, normality assumptions are often not conducted and parametric approach is inattentively used.

Here, we sampled 30 papers published in major 3 journals in toxicology field in 2014 which have assessed group differences with statistics. The characteristics of data presented were analyzed, categorized and examined. There are more than 110 endpoints or outcome measures in those 30 papers and we reviewed sample size, presentation of the data (SD or SEM) and the employed inferential statistical methods. We also discussed about the logics for selection and appropriateness of statistical methods to give an insight into the proper selection of statistical methods.

MATERIALS AND METHODS

Selection of papers. Thirty articles with the purpose of examining group differences from toxicant exposure have been selected, that were published in *Toxicology and Applied Pharmacology* (11 papers (8-17)), *Archives of Toxicology* (8 papers (18-25)), and *Toxicological Science* (11 papers, (26-36)) in 2014. These three journals were high-ranked in the category of Toxicology of Thompson-Reuter, JCR 2012. In the selection of 30 articles, diversity in statistical methods were considered and papers with unclear explanation for statistics were excluded. From these 30 papers, outcome measures (endpoints) were extracted and reviewed for data type, sample size and method of descriptive and inferential statistics.

Data classification. The data has been classified as numeric vs. categorical, based on the outcome measures shown in the paper. Numeric data has been categorized as continuous vs, discrete, and categorical variable was classified as nominal vs. ordinal. Endpoints were classified as "descriptive statistics" if only descriptive statistics (and not inferential statistics) were presented, whereas classified as inferential statistics if p value has been presented and conclusion was drawn based on the p-value. If the method used to estimate p-value could not be identified, then it was classified as "could not be determined". Parametric methods include those tests which assumed that the data are nor-

mally distributed; normality test includes Kolmogorov-Smirnov, D'Agostino and Pearson omnibus and Shapiro-Wilk normality tests; equal variance test, which is required for ANOVA (Analysis of Variance) includes Levene's test and Bartlett's test.

RESULTS

Of 30 papers being selected, 113 outcome measures (or endpoints) were observed (Table 1), and most of them were numeric variables (112/113, 99%) except for one outcome used "pain scale" measure, which falls into ordinal (categorical) variable. Of 112 numeric variables, 109 variables were continuous while 3 outcomes counted the number of neutrophils, immune-stained cells, and number of visits, which could be categorized as discrete variables. 109 continuous variables could be further classified into relative values (68/109, 63% mostly relative to comparison group) and absolute values (41/109, 37%) but few studies have provided justification regarding the why the value has been described as relative or absolute value.

As shown in Fig. 1, the sample size distribution is heavily right-skewed, with most values clustered below 11 and the median, mean, and mode being 6, 16.4, and 3 & 6, respectively. There are some outliers which strongly influenced

Table 1. Data characteristics of endpoints

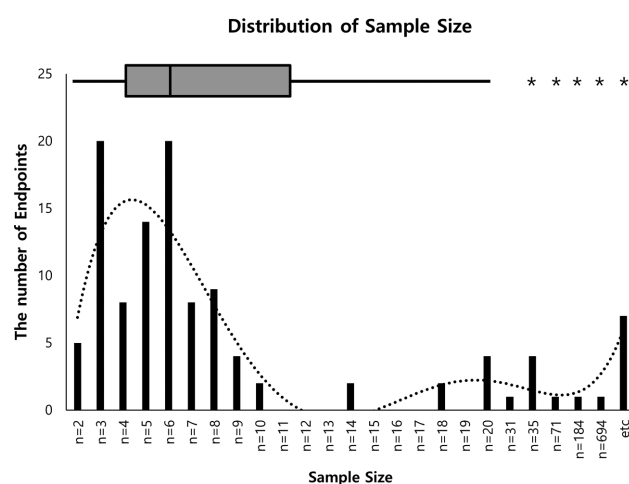| Numerical			Categorical
Continuous (absolute/relative) 109 (41/68)	Discrete 3		Ordinal 1
112			1

Fig. 1. Distribution of sample size per study. (A) Box plot showing median (6), 25% and 75% quartile ranges around the median (box width) and, 5% and 95% of sample size per study (whiskers) and outliers (*), (B) Histogram showing distribution of sample size per study.
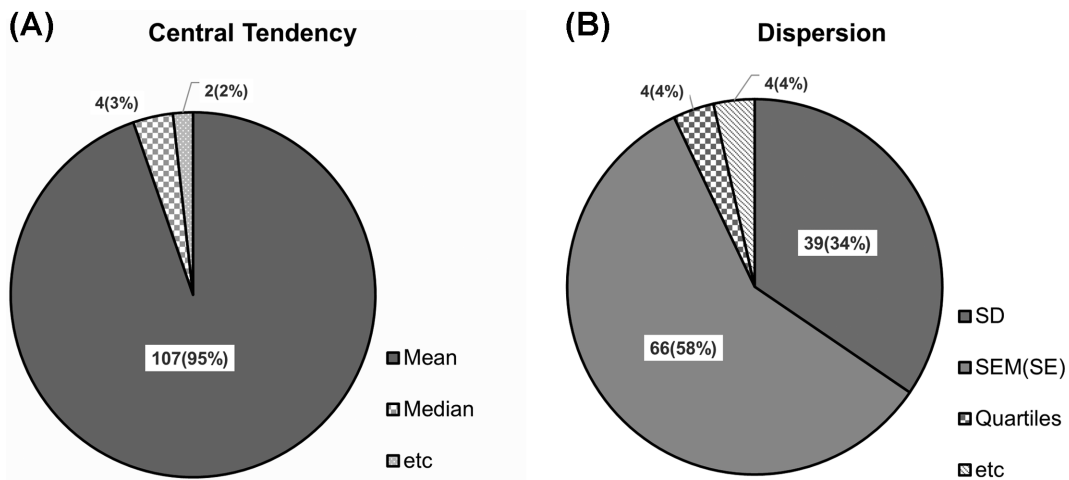
(A)

Central Tendency

4(3%)   2(2%)

107(95%)

□ Mean

□ Median

□ etc

(B)

Dispersion

4(4%)   4(4%)

39(34%)

66(58%)

□ SD

□ SEM(SE)

□ Quartiles

□ etc

Fig. 2. Distribution of measures used to describe summary statistics. (A) shows the distribution of the measures used to describe central tendency of the data (endpoints). (B) shows the distribution of measures used to describe dispersion of the data (endpoints). Legend: SD, standard deviation; SE, standard error; SEM, standard error of the mean.

the mean, with sample size being 71, 184, and 694. The paper which employed large sample sizes of 184 and 694 were about neurotoxicity of heterocyclic amines but the definition of sample number were not conventional (number of neurites for the outcome measurement of neurite length, and mean pixel intensity in the regions of interests) (32) while the one used the average sample size of 71 were about the neurotoxicity of flame retardants and the sample number is the number of wells (of 96 well plate) per group (22). Besides these few outliers, most studies have small sample sizes, raising a concern that the studies may be underpowered to conduct inferential statistics and the dispersion of data might not indicate the true distribution of the

underlying distribution of the outcome variables.

To describe the central tendency of the data, mean was most frequently used (93%, 105 outcomes), and 6 outcomes employed median (Fig. 2A), but none provided explanation regarding the selection of the summary statistics. Median survival time from survival analysis was presented in one study which measured mice's survival (26) and levels in metabolites, mRNA and protein by bisphenol A exposure were presented as median with interquartile range in one paper (35). For the dispersion of the data, SEM/SE was most frequently used (66/113, 58%), followed by SD (39/113, 34%) (Fig. 2B). Interquartile ranges were provided only for 4 outcomes (35,37), all of whom used median and
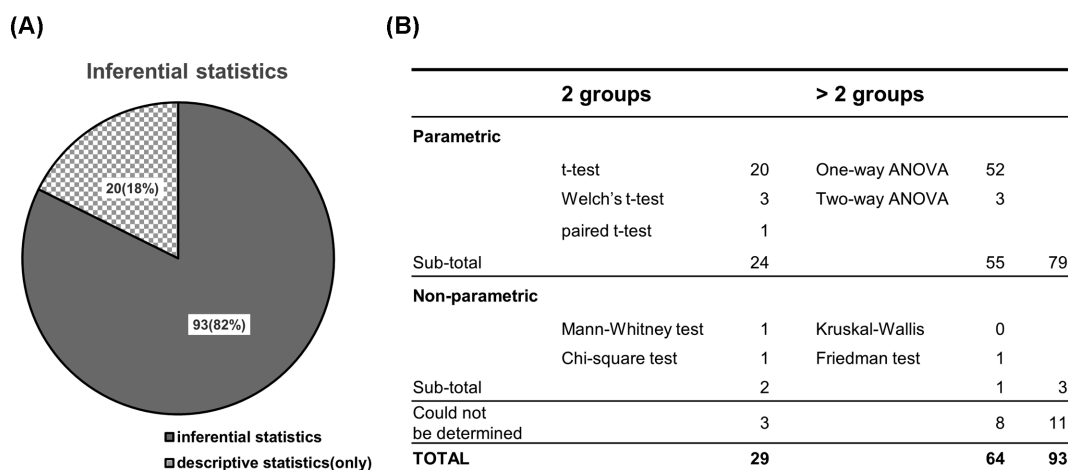
(A)

Inferential statistics

20(18%)

93(82%)

□ inferential statistics

□ descriptive statistics(only)

(B)

	2 groups		> 2 groups		
Parametric					
	t-test	20	One-way ANOVA	52	
	Welch's t-test	3	Two-way ANOVA	3	
	paired t-test	1			
Sub-total		24		55	79
Non-parametric					
	Mann-Whitney test	1	Kruskal-Wallis	0	
	Chi-square test	1	Friedman test	1	
Sub-total		2		1	3
Could not be determined		3		8	11
TOTAL		**29**		**64**	**93**

Fig. 3. Description of studies conducting inferential statistics. (A) Proportion of endpoints conducting inferential statistics; endpoints were classified as "descriptive statistics" if only descriptive statistics (and not inferential statistics) were presented, whereas classified as inferential statistics if inferential statistics were conducted, namely, p value has been presented and conclusion was drawn based on the p-value. (B) Classification of endpoints conducting inferential statistics. If the method used to estimate p-value could not be identified, then it was classified as "could not be determined".

(A) Parametric Method
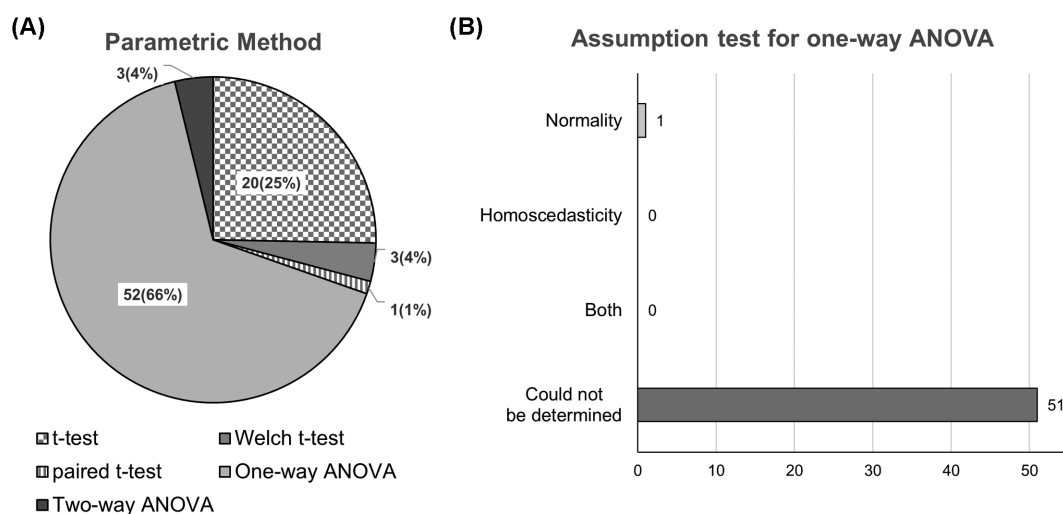
(B) Assumption test for one-way ANOVA

Fig. 4. Detailed description of studies conducting ANOVA (analysis of variance). (A) Analysis of studies conducting parametric method. (B) Analysis of 52 studies conducting one-way ANOVA.

box-whisker plots for data description, and dispersion was not presented in 4 endpoints.

Conspicuously, majority of outcome measures were analyzed by inferential statistics (93/113, 82%) (Fig. 3A), most probably because the original purpose of the studies aimed to examine the difference between the group means or to examine the effects of toxicants. Those that did not conducted inferential statistics (20/113), were mostly describing just observations or baseline characteristics where the determination of group difference was not necessary. For inferential statistics, parametric methods were frequently employed (parametric 77/93, 82.8%), while non-parametric methods (3/93, 3.2%) were rarely employed (Fig. 3B). Of 93 outcomes that underwent inferential statistics, two-group comparison was 29 and others (64/93, 68.8%) were multi-group (more than 2 groups) comparison. Interestingly, of 64 multi-group comparisons, ANOVA was conducted for 55, representing most favored inferential statistical method for multi-group comparisons.

Fifteen out of ninety three endpoints that conducted inferential statistics underwent assumption test which is critical for selecting whether to use parametric or non-parametric methods. If we look further into parametric methods, one-way ANOVA is the most frequently employed (55/79, 69.6%) followed by t-test (20/79, 27%, Fig. 4A). Interestingly, for one-way ANOVA, homoscedasticity or normality assumptions were rarely tested (Fig. 4B, 51/52, 98.1%). We could observe that only one paper described that they conducted normality test before one-way ANOVA (36) while none of the endpoints conducted equal variance test (0/49, 0%).

DISCUSSION

Our study illustrates that in the sampled 30 articles pub-

lished in *Toxicological Sciences*, *Toxicology and Applied Pharmacology* and *Archives of Toxicology*, which are top-ranked toxicology journals, inconsistent statistical methods were frequently used when describing and inferring data.

Diverse methods were often mixed for the description of data; for example, mean and median were used but it is difficult whether they were used properly since we could not identify the distribution of variables in the paper. SD and SEM were used without proper explanation, too. Descriptive statistics sought to explain a given study sample, thus if the purpose of the study is to explain the variability within the sample, then SD should be used; if the purpose is to estimate how the mean of the sample is related to the mean of underlying population, then SEM is used. Since SEM is always smaller than SD, one might speculate that SEM has been frequently used where SM should be used instead to underestimate the variability of the sample, which may lead readers to assume smaller variation of the sample (3). Our study identified that SEM is more frequently used than SD, and both methods are selected without clear justification. Moreover, there is a high possibility of SEM being improperly used, yet further study should be conducted to substantiate the issue of SEM being inappropriately used.

Our study also shows that many studies have small sample size; the mode of sample size being 3 & 6 and the median being 6, thus it is not clear whether the results obtained from the studies can be reproduced when the sample size has been expanded, since data with small sample size are especially vulnerable to outliers. Given that our study includes *in vitro* as well as *in vivo* studies, the median of the sample size could even diminish further if we limit our study to *in vitro* or *in vivo* only which warrants future study.

To make an inference regarding the test result, then para-

metric or non-parametric methods can be used; parametric method, which includes ANOVA and t-test, are conventional and well-known, yet the sample is assumed to be normally distributed or approximately so. If the normality assumption is violated, then non-parametric method, which is sometimes called as distribution-free method, should be considered (6). Since the studies usually have small sample size (as shown in Fig. 1), the data might not conform to normality assumptions; yet our study shows that parametric approach was frequently used without testing relevant assumptions, (Fig. 4B) One-way ANOVA, which is the most frequently used parametric method when comparing multiple groups, requires homoscedasticity assumption. ANOVA, whose full-name is Analysis of Variance, compares within-groups variance with between-group variance, thus homoscedasticity assumption is critical, yet our study shows that the key assumption is frequently ignored.

Statistical significance does not necessarily translate into clinical significance (38), implying that statistical difference does not indicate clinical (and possibly toxicological) differences. Thus, more attention should be paid when interpreting the statistical significance of the study result. Also, transforming absolute into relative values (such as use of change from baseline) or using absolute value should be determined based on the clinical (toxicological) interpretation of the results (5), yet our study shows that relative and absolute values were used together (as shown in Table 1) without proper explanation.

Insufficient explanation of statistical method was prevalent, which prevented us from conducting more detailed explanation. For example, the study methods specify "the homogeneity of variance for all the data sets was firstly examined using Bartlett test, and then the data succeeding or failing to pass the test were analyzed by one-way analysis of variance or Kruskal–Wallis test", respectively but it did not specify whether ANOVA was used (namely, test passed) or not (test failed) in the study (15). These studies were categorized as "inferential statistics conducted, yet parametric/non-parametric unclear (could not be determined)", which might underestimate the overall performance of the study. Also, one should interpret our result with caution, since we sampled only 30 papers from the journals, thus our result may not represent the general trend in toxicology journals.

However, our study attempted to illustrate that statistical part of the toxicological experimental studies are inconsistently used with insufficient explanation regarding the method selected for the analysis. It is deplorable to see the absence of relevant explanation on statistics considering the prestige of the journals in toxicology. However, considering that conclusions drawn from experimental studies provide critical evidence or ground in regulatory affairs, more consistent and transparent approach should be taken when analyzing the study results. It would be important to review more

papers and to invite biostatisticians to discuss the appropriateness of the statistical methods employed in toxicology. Such a study will provide a good reference for establishing appropriate decision tree for selection of statistical methods, which is invaluable for improvement of the reliability of the results and advancement of toxicological research.

ACKNOWLEDGEMENTS

REFERENCES

1. Schor, S. and Karten, I. (1966) Statistical evaluation of medical journal manuscripts. *JAMA*, **195**, 1123-1128.
2. Ko, W.R., Hung, W.T., Chang, H.C. and Lin, L.Y. (2014) Inappropriate use of standard error of the mean when reporting variability of study samples: A critical evaluation of four selected journals of obstetrics and gynecology. *Taiwan J. Obstet. Gynecol.*, **53**, 26-29.
3. Nagele, P. (2003) Misuse of standard error of the mean (sem) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *Br. J. Anaesth.*, **90**, 514-516.
4. Altman, D.G. and Bland, J.M. (2005) Standard deviations and standard errors. *BMJ*, **331**, 903.
5. U.S. Department of Health and Human Services FaDA, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). (1998) Guidance for Industry:E9 Statistical Principles for Clinical Trials. pp. 1-22.
6. Pagano, M. and Gauvreau, K. (2000) Survival analysis-thelife table method (2nd edition), Principles of Biostatistics, Pacific Grove, Duxbury, pp. 488-494.
7. Cumming, G., Fidler, F. and Vaux, D.L. (2007) Error bars in experimental biology. *J. Cell Biol.*, **177**, 7-11.
8. Ahad, A., Ganai, A.A., Mujeeb, M. and Siddiqui, W.A. (2014) Chrysin, an anti-inflammatory molecule, abrogates renal dysfunction in type 2 diabetic rats. *Toxicol. Appl. Pharmacol.*, **279**, 1-7.
9. Baumann, L., Knörr, S., Keiter, S., Rehberger, K., Volz, S., Schiller, V., Fenske, M., Holbech, H., Segner, H. and Braunbeck, T. (2014) Reversibility of endocrine disruption in zebrafish (Danio rerio) after discontinued exposure to the estrogen 17alpha-ethinylestradiol. *Toxicol. Appl. Pharmacol.*, **278**, 230-237.
10. Chen, J., Shetty, S., Zhang, P., Gao, R., Hu, Y., Wang, S., Li, Z. and Fu, J. (2014) Aspirin-triggered resolvin D1 down-regulates inflammatory responses and protects against endotoxin-induced acute kidney injury. *Toxicol. Appl. Pharmacol.*, **277**, 118-123.
11. Galvis-Pareja, D., Zapata-Torres, G., Hidalgo, J., Ayala, P., Pedrozo, Z., Ibarra, C., Diaz-Araya, G., Hall, A.R., Vicencio, J.M., Nuñez-Vergara, L. and Lavandero, S. (2014) A novel dihydropyridine with 3-aryl meta-hydroxyl substitution blocks L-type calcium channels in rat cardiomyocytes. *Toxicol. Appl. Pharmacol.*, **279**, 53-62.
12. Khanal, T., Kim, H.G., Do, M.T., Choi, J.H., Won, S.S., Kang,

W., Chung, Y.C., Jeong, T.C. and Jeong, H.G. (2014) Leptin induces CYP1B1 expression in MCF-7 cells through ligand-independent activation of the ERalpha pathway. *Toxicol. Appl. Pharmacol.*, **277**, 39-48.

13. Koller, V.J., Auwärter, V., Grummt, T., Moosmann, B., Mišik, M. and Knasmüller, S. (2014) Investigation of the *in vitro* toxicological properties of the synthetic cannabimimetic drug CP-47,497-C8. *Toxicol. Appl. Pharmacol.*, **277**, 164-171.

14. Li, H.B., Qin, D.N., Ma, L., Miao, Y.W., Zhang, D.M., Lu, Y., Song, X.A., Zhu, G.Q. and Kang, Y.M. (2014) Chronic infusion of lisinopril into hypothalamic paraventricular nucleus modulates cytokines and attenuates oxidative stress in rostral ventrolateral medulla in hypertension. *Toxicol. Appl. Pharmacol.*, **279**, 141-149.

15. Takeda, T., Taura, J., Hattori, Y., Ishii, Y. and Yamada, H. (2014) Dioxin-induced retardation of development through a reduction in the expression of pituitary hormones and possible involvement of an aryl hydrocarbon receptor in this defect: a comparative study using two strains of mice with different sensitivities to dioxin. *Toxicol. Appl. Pharmacol.*, **278**, 220-229.

16. Wu, W., He, K., Zhou, H.R., Berthiller, F., Adam, G., Sugita-Konishi, Y., Watanabe, M., Krantis, A., Durst, T., Zhang, H. and Pestka, J.J. (2014) Effects of oral exposure to naturally-occurring and synthetic deoxynivalenol congeners on proinflammatory cytokine and chemokine mRNA expression in the mouse. *Toxicol. Appl. Pharmacol.*, **278**, 107-115.

17. Zhou, C.F., Zhou, D.C., Zhang, J.X., Wang, F., Cha, W.S., Wu, C.H. and Zhu, Q.X. (2014) Bleomycin-induced epithelial-mesenchymal transition in sclerotic skin of mice: possible role of oxidative stress in the pathogenesis. *Toxicol. Appl. Pharmacol.*, **277**, 250-258.

18. Stern, M., Gierse, A., Tan, S. and Bicker, G. (2014) Human Ntera2 cells as a predictive in vitro test system for developmental neurotoxicity. *Arch. Toxicol.*, **88**, 127-136.

19. Ramaiahgari, S.C., den Braver, M.W., Herpers, B., Terpstra, V., Commandeur, J.N., van de Water, B. and Price, L.S. (2014) A 3D *in vitro* model of differentiated HepG2 cell spheroids with improved liver-like properties for repeated dose high-throughput toxicity studies. *Arch. Toxicol.*, **88**, 1083-1095.

20. Onami, S., Cho, Y.M., Toyoda, T., Mizuta, Y., Yoshida, M., Nishikawa, A. and Ogawa, K. (2014) A 13-week repeated dose study of three 3-monochloropropane-1,2-diol fatty acid esters in F344 rats. *Arch. Toxicol.*, **88**, 871-880.

21. Nakagawa, Y., Suzuki, T., Nakajima, K., Inomata, A., Ogata, A. and Nakae, D. (2014) Effects of N-acetyl-L-cysteine on target sites of hydroxylated fullerene-induced cytotoxicity in isolated rat hepatocytes. *Arch. Toxicol.*, **88**, 115-126.

22. Hendriks, H.S., Meijer, M., Muilwijk, M., van den Berg, M. and Westerink, R.H. (2014) A comparison of the *in vitro* cyto- and neurotoxicity of brominated and halogen-free flame retardants: prioritization in search for safe(r) alternatives. *Arch. Toxicol.*, **88**, 857-869.

23. Follmann, W., Behm, C. and Degen, G.H. (2014) Toxicity of the mycotoxin citrinin and its metabolite dihydrocitrinone and of mixtures of citrinin and ochratoxin A in vitro. *Arch. Toxicol.*, **88**, 1097-1107.

24. Dahdouh, F., Raane, M., Thevenod, F. and Lee, W.K. (2014) Nickel-induced cell death and survival pathways in cultured renal proximal tubule cells: roles of reactive oxygen species, ceramide and ABCB1. *Arch. Toxicol.*, **88**, 881-892.

25. Barbosa, D.J., Serrat, R., Mirra, S., Quevedo, M., Gómez de Barreda, E., Avila, J., Fernandes, E., Bastos Mde, L., Capela, J.P., Carvalho, F. and Soriano, E. (2014) MDMA impairs mitochondrial neuronal trafficking in a Tau- and Mitofusin2/Drp1-dependent manner. *Arch. Toxicol.*, **88**, 1561-1572.

26. Shmarakov, I.O., Borschovetska, V.L., Marchenko, M.M. and Blaner, W.S. (2014) Retinoids modulate thioacetamide-induced acute hepatotoxicity. *Toxicol. Sci.*, **139**, 284-292.

27. Otsuka, S., Ishihara, A. and Yamauchi, K. (2014) Ioxynil and tetrabromobisphenol A suppress thyroid-hormone-induced activation of transcriptional elongation mediated by histone modifications and RNA polymerase II phosphorylation. *Toxicol. Sci.*, **138**, 290-299.

28. Ngalame, N.N., Tokar, E.J., Person, R.J., Xu, Y. and Waalkes, M.P. (2014) Aberrant microRNA expression likely controls RAS oncogene activation during malignant transformation of human prostate epithelial and stem cells by arsenic. *Toxicol. Sci.*, **138**, 268-277.

29. Lu, Y., Wang, Y.Y., Yang, N., Zhang, D., Zhang, F.Y., Gao, H.T., Rong, W.T., Yu, S.Q. and Xu, Q. (2014) Food emulsifier polysorbate 80 increases intestinal absorption of di-(2-ethylhexyl) phthalate in rats. *Toxicol. Sci.*, **139**, 317-327.

30. Li, X., Qu, F., Xie, W., Wang, F., Liu, H., Song, S., Chen, T., Zhang, Y., Zhu, S., Wang, Y., Guo, C. and Tang, T.S. (2014) Transcriptomic analyses of neurotoxic effects in mouse brain after intermittent neonatal administration of thimerosal. *Toxicol. Sci.*, **139**, 452-465.

31. Lefebvre, D.E., Pearce, B., Fine, J.H., Chomyshyn, E., Ross, N., Halappanavar, S., Tayabali, A.F., Curran, I. and Bondy, G.S. (2014) In vitro enhancement of mouse T helper 2 cell sensitization to ovalbumin allergen by carbon black nanoparticles. *Toxicol. Sci.*, **138**, 322-332.

32. Griggs, A.M., Agim, Z.S., Mishra, V.R., Tambe, M.A., Director-Myska, A.E., Turteltaub, K.W., McCabe, G.P., Rochet, J.C. and Cannon, J.R. (2014) 2-Amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP) is selectively toxic to primary dopaminergic neurons in vitro. *Toxicol. Sci.*, **140**, 179-189.

33. Conti, A., Kobets, T., Escudero-Lourdes, C., Montgomery, B., Tryndyak, V., Beland, F.A., Doerge, D.R. and Pogribny, I.P. (2014) Dose- and time-dependent epigenetic changes in the livers of Fisher 344 rats exposed to furan. *Toxicol. Sci.*, **139**, 371-380.

34. Churchwell, M.I., Camacho, L., Vanlandingham, M.M., Twaddle, N.C., Sepehr, E., Delclos, K.B., Fisher, J.W. and Doerge, D.R. (2014) Comparison of life-stage-dependent internal dosimetry for bisphenol A, ethinyl estradiol, a reference estrogen, and endogenous estradiol to test an estrogenic mode of action in Sprague Dawley rats. *Toxicol. Sci.*, **139**, 4-20.

35. Chen, M., Zhou, K., Chen, X., Qiao, S., Hu, Y., Xu, B., Xu, B., Han, X., Tang, R., Mao, Z., Dong, C., Wu, D., Wang, Y., Wang, S., Zhou, Z., Xia, Y. and Wang, X. (2014) Metabolomic analysis reveals metabolic changes caused by bisphenol A in rats. *Toxicol. Sci.*, **138**, 256-267.

36. Barbosa, D.J., Serrat, R., Mirra, S., Quevedo, M., de Barreda, E.G., Àvila, J., Ferreira, L.M., Branco, P.S., Fernandes, E., Lourdes Bastos, M., Capela, J.P., Soriano, E. and Carvalho, F. (2014) The mixture of "ecstasy" and its metabolites impairs

mitochondrial fusion/fission equilibrium and trafficking in hippocampal neurons, at in vivo relevant concentrations. *Toxicol. Sci*., **139**, 407-420.

37. Schellenberger, M.T., Grova, N., Farinelle, S., Willième, S., Schroeder, H. and Muller, C.P. (2013) Modulation of benzo[a]pyrene induced neurotoxicity in female mice actively immunized with a B[a]P-diphtheria toxoid conjugate. *Toxicol. Appl. Pharmacol*., **271**, 175-183.

38. Brahman, L.E. (1991) Confidence Intervals Assess Both Clinical Significance and Statistical Significance. *Ann. Intern. Med*., **114**, 515-517.