**Analysis of historical negative control group data from the rat *in vivo* micronucleus assay**

Lovell, D. P1., Fellows, M2., Saul, J3., Whitwell, J3., Custer, L4., Dertinger, S5., Escobar, P6., Fiedler, R7., Hemmann,U8., Kenny, J9., Smith, R3., van der Leede, B.M.j10 & Zeller, A11.

1 St George’s Medical School, University of London, Cranmer Terrace, London, SW17 0RE, UK

2 Astra Zeneca, Drug Safety and Metabolism, Cambridge, CB4 0WG, UK

3 Covance Laboratories, Harrogate, North Yorkshire, HG3 1PY, UK

 4 Bristol-Myers Squibb, New Brunswick, NJ, USA

5 Litron Laboratories, Rochester, NY, USA

6 Merck Co. & Inc., West Point, PA, USA

7 Covance Laboratories, Greenfield, IN, USA

8 Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany

9 GlaxoSmithKline, Ware, Hertfordshire, UK

10 Janssen R&D, a Division of Janssen Pharmaceutica N.V., Beerse, Belgium

11 Pharmaceutical Sciences, pRED Innovation Center Basel, F. Hoffmann-La Roche Ltd, 4070 Basel, Switzerland

\* Corresponding author: St George’s Medical School, University of London, London, SW17 0RE, UK; email: dlovell@sgul.ac.uk

**Key words:** *In vivo* micronucleus assay, Historical negative control data, Quality control statistics, Wistar and Sprague Dawley rats

**Headline points:**

* Historical negative control data are important for assessing data quality and for interpretation of the *in vivo* micronucleus assay
* 10 laboratories’ data from rat *in vivo* micronucleus experiments were analyzed statistically.
* Mean incidence ranged from 0.44 to 2.22 micronucleated cells/1000 cells

**Abstract**

A database of micronuclei counts for historical negative control data from rat *in vivo* micronuclei tests performed in 10 different laboratories was established. Data were available from over 4000 negative control rats from 10 laboratories. The mean frequency of micronucleated cells (MN)/1000 cells ranged from 0.44 to 2.22, a 5-fold range. Overall there were no major sex or strain differences in frequency, although there were some small but statistically significant differences within laboratories. There was appreciable variability between experiments compared with variability within experiments in some laboratories. No specific factor was identified which could explain this variability although it was noted that many different vehicles were used in the experiments. It is hoped that these data will help laboratories beginning studies with the rat micronucleus assay and those involved in the assessment of micronucleus assay results.

**Introduction**

The mammalian (in vivo) micronucleus test has been developed and used for over 40 years (Heddle, 1973, Schmid, 1975 and Hayashi, 2016). It provides a method for assessing the genotoxicity of chemical or physical agents in a mammalian species, providing a system for detecting cytogenetic damage resulting from clastogenic or aneugenic activity. It has an important role in a number of the testing strategies for assessing the genotoxicity of chemicals (COM, 2011). In the case of the International Conference on Harmonization ICH S2(R1)) guidance document (ICH 2011) it is one of the recommendations as part of the required test battery for genotoxicity to be performed.

Methods for carrying out the assay have been published in several papers. The first Organisation for Economic Cooperation and Development (OECD) Guideline (TG 474) for the Mammalian erythrocyte micronucleus test was in 1983, revised in 1997 and extensively revised in 2014.The latest revision was in 2016 (OECD, 2016a) and provides recommendations for the conduct of the mammalian (in vivo) micronucleus test.

The revised genotoxicity guidelines (OECD, 2016b) emphasised a more important role for historical negative control data in the analysis and interpretation of *in vivo* micronucleus assays. The guidance has indicated that comparison of individual study results with a laboratory’s historical negative control data should form part of result evaluation and laboratories are encouraged to build up a dataset of negative control data to provide evidence of their competence with the test.

The ILSI/HESI Genetic Toxicology Technical Committee (GTTC) Data Interpretation Workgroup began a project to collect and collate sets of data from experienced laboratories into a standardized database intended to provide high quality negative control data from genotoxicity tests carried out under OECD guidelines. This group has previously reported on the negative control data from a set of *in vitro* micronuclei studies from a series of experienced laboratories (Lovell *et al*, 2018a and Lovell *et al*, 2018b). The GTTC decided to investigate the degree of variability in the rat *in vivo* micronucleus assay. Tests using the rat were chosen because of the increasing interest in including the micronucleus as part of other standard toxicology tests. Data were not requested from other species (such as the mouse) on which the test is also carried out. This choice was made because a large and relevant historical database was likely to be available from studies in rats. Also, pre-existing toxicity and ADME data in the Pharmaceutical Industry are available from studies carried out in the rat, facilitating follow-up genotoxicity testing in the same species.

**Methods**

The GTTC canvased a number of laboratories which would have relevant data for their interest in the study. Data were collected from those laboratories agreeing to take part in the project. In total, ten laboratories (5 from the USA and 5 from Europe) participated. Two of these laboratories were from a single company with one laboratory in Europe and another in the USA. One laboratory reported separate datasets for two different rat strains. The ten laboratories were assigned letters A to J to provide a degree of anonymization*.*

It was found that scoring methods differed among the laboratories. Some used manual scoring methods, while others used flow cytometry. In addition, one laboratory (Lab G) collected data from a separate group of animals and used automatic slide evaluation with Metafer 4 (MetaSystems, Germany).

8 laboratories provided data on samples derived from bone marrow (BM) and 5 provided data on samples derived from peripheral blood (PB), with three laboratories (C, H and I) providing separate datasets on BM and PB samples. Four laboratories, (D, E, H and I) provided separate sets of data from Wistar and Sprague Dawley rat strains. One laboratory (G) provided separate datasets of samples receiving either a single or repeat dosing. In all, fifteen separate datasets were created (S1 to S15).

The number of experiments (over 600), the number of rats (males and females) and other relevant information are shown in Table 2. The number of rats per laboratory providing data ranged from 148 to 1358. In all, data from 4356 rats (3549 males, 787 females and 20 unknowns (i.e. where the animal's sex was unavailable) were provided.

Most experiments used just one sex (usually males) but females were included in an appreciable number of the experiments. Where both sexes were used the experiments often consisted of 5 males and 5 females, but some experiments used appreciably more (n=24). Direct comparisons between the sexes from individual laboratories were possible in about 12% of the experiments.

*Data collection and management*

Two Excel spreadsheets (See Supplementary Material Figs 1 and 2) were developed and trialled by Covance - one for data obtained from bone marrow (BM) and the other for data from peripheral blood (PB) samples. The spreadsheets required participants to complete a short questionnaire on the laboratory and husbandry conditions (Table 1) based on the ARRIVE guidelines (Kilkenny *et al*, 2010). Most laboratories completed this satisfactorily.

A considerable number of vehicles were used including (amongst many others): deionized water, corn oil, PBS, PE and HPMC. (A complete listing is provided in the Supplementary Material). Routes of administration were: intravenous (infusion and slow bolus), oral gavage (p.o.), subcutaneous (s.c.) and intra-peritoneal (i.p.).

No restriction was placed upon the amount of data that could be submitted other than that laboratories were expected to provide data from a minimum of 20 experiments. In some laboratories, multiple vehicles were used which meant that there were often less than 20 experiments available for comparisons of specific vehicles. In these circumstances data from experiments using different vehicles were combined.

Samples were scored either manually when samples were derived from the bone marrow (BM) cells or using automated flow cytometry when the samples were obtained from the peripheral blood (PB) samples. Laboratories varied in the number of cells scored but, in the case of bone marrow samples, approximately 500 to 2000 cells were scored for toxicity and an estimation of %PCE values with approximately 1000 to 4000 PCE’s scored for micronuclei. In the case of peripheral blood samples large numbers of cells were sampled (ranging from about 160,000 to 6 million) to arrive at approximately 20,000 immature reticulocytes which were scored for the presence of micronuclei. The remaining mature normochromatic erythrocytes (NCE) were used in the assessment of toxicity by calculating the percentage immature reticulocytes (%RET) in the total sample of cells scored.

Six items of data were included on the 'Bone Marrow' Excel spreadsheet (3 for toxicity and 3 for the micronucleus endpoint). For toxicity: total cells scored, number of immature cells (PCE) and number of mature cell (NCE). For the micronucleus endpoint: total cells scored for MN, number of MN PCE and number of MN NCE. Five items of data were included on the 'Peripheral Blood' Excel spreadsheet: total cells scored, number of mature erythrocytes, number of MN mature erythrocytes, number of immature RET and number of MN RET.

Most of the data were provided in the form of the number of micronucleated cells among the number of cells scored for each animal. In one case (Lab B) data were reported solely as frequencies. Data were converted from manual counting of bone marrow cells into MN/1000 PCEs (datasets S1, S4, S7, S8, S9, S10, S11, S13, S15) with these datasets also reporting the PCE:NCE ratio expressed as %PCEs. Data were converted from flow cytometry counting peripheral blood cells into MN/1000 immature reticulocytes (RET) (datasets S2, S3, S5, S6, S12, S14). %RETS were also reported for these datasets.

Some laboratories (datasets Lab C (S4) and Lab F (S8)) also scored bone marrow normochromatic erythrocytes (NCE) for micronuclei (analysed as MN/1000 normochromatic erythrocytes) while others (Lab C (S3), Lab D (S5& S6), Lab H (S12) and Lab I (S14)) using peripheral blood samples, scored the number of 'mature' cells with micronuclei.

The main endpoints analysed were the number of MN/1000 PCE or MN/1000 immature RET (both were termed MN/1000 cells in the analyses), the %PCE and the %RETs.

*Statistical methods*

33 separate combinations (C1-C33) of sex, strain and tissue were assessed. All laboratories except Lab D carried out studies using both males and females, in some cases in the same experiment.

Appropriate statistical methods used in analysing data from the *in vivo* micronucleus assay are cited by the OECD (Lovell *et al*, 1989; Hayashi *et al*, 1989; Hayashi *et al*, 1994; Kim *et al*, 2000 and Jarvis *et al*, 2011). The methods used in this analysis were broadly based upon those reported in previous papers (Lovell *et al*, 2018a and Lovell *et al*, 2018b). Data were analysed using the statistical procedures available in Minitab (Minitab 16 Statistical Software. Minitab, Inc., State College, PA) and the R statistical programming language (R Development Core Team, 2011).

The methods used included: one-way and nested analyses of variances, tests for extra-binomial variation (i.e. Goodness of fit to a Poisson distribution) and calculation of tolerance intervals.

Several QC methods were used: I-Charts for individual replicate values of counts or proportions of micronucleated cells, X-bar Charts when there were several replicates per experiment and C-Charts for Poisson counts (See Lovell *et al* (2018a) for details on the interpretation of these charts). Detailed results of many of these analyses can be found in the Supplementary Material.

The analyses of variances (anovas) were carried out on untransformed 'counts' or proportions. These gave satisfactory results for the types of exploratory analyses carried out in this paper. The Box-Cox method was used to explore possible optimal transformations of the data (see later re: Box-Cox method).

**Results and discussion**

Ten laboratories (identified as A to J) provided data in Table 2. Data from all the participating laboratories were considered acceptable for inclusion in the analyses. The quality of the data provided was good and the laboratories were diligent in their presentation of the data in a standard format for all the sets of conditions. Some laboratories reported independent sets of data from studies using different rat strains, different scoring methods or different treatment regimens such that there were 15 separate datasets (S1–S15).

In all, there were 33 different combinations of sexes, strains and scoring methods (C1-C33) representing over 600 separate experiments. These datasets were made up of: 14 combinations with data from female rats and 18 from male rats and 1 ‘unknown’ (sex of rats not specified); 17 of these combinations were ‘Wistar’ and 16 ‘Sprague Dawley’ rats. 10 were from peripheral blood samples and 23 from bone marrow samples. The results submitted provided an opportunity to assess the relative variability within and between experiments.

*Overall range of values*

The ‘primary result was that the means of the 33 combinations ranged from 0.44 MN/1000 cells to 2.22 MN/1000 cells: a 5-fold range. Figures 1-4 show the means with their associated standard deviations (Fig 1) or their associated 95% confidence intervals (CIs) (Figs 2-4). The width of the CI’s varies considerably because of the differences in the sample sizes.

These results were obtained from many experiments across several variables (e.g. laboratories, strains, sexes and methods) carried out by proficient laboratories and provide a range for new laboratories to aim for. The mean from experiments carried out by a new laboratory should fall within this range (approximately 0.5 to 2.5 MN/1000 cells) but their between-experiment variability within their laboratory should be appreciably smaller than the between-laboratory variability reported here.

There was no overall major difference between males and females (Fig 2), between Wistar and Sprague Dawley (Fig 3) or between the means obtained from the different tissues and methodologies (BM and PB) (Fig 4). However, statistical analyses (see below) applied within the datasets showed that, in some cases, small but statistically significant differences between some of these variables were found.

*Variability between laboratories*

Fig 1 shows that there were clear differences in counts of micronucleated cells between the different laboratories, with Labs E (C11&C12) and H (C22-C25) having low frequencies and Lab C (C7&C8) and Lab F (C15-C17) having higher frequencies. Further analyses to try and identify factors differing between the laboratories which might explain this variability is difficult because there is a wide range of variables reported but only a small sample of laboratories available to use in more advanced statistical modelling.

*Variability within laboratories*

This variability can be explored further using control charts of the individual (I-chart) and group mean (Xbar-charts) values. Representative I-Charts are presented in the Appendix as Figs I1-I15. The X-bar Charts (Figs X1-X15) and a more complete set of charts and further analyses can be found in the Supplementary Material section.

Table 3 summarises the results of one-way anovas carried out on factors that varied within laboratories. The results of the anovas are shown as the P-values associated with the F-test for between factor levels, together with the percentage of the total variability within the dataset 'explained' by the variability in the factor levels. It is important to appreciate that small effects identified by highly statistically significant small P-values may be a consequence of the statistical power of a test based upon large sample sizes, but may 'explain' little of the total variability.

1. *Between experiments*

In those laboratories where it was possible to test the between-experiment versus the within-experiment variability the following laboratories showed significant (P<0.05) between-experiment variability compared with within-laboratory variability: Lab A (S1) (P<0.001, 14%); Lab C (S3) (P<0.001, 13%); Lab E (S7) (P<0.001, 16%); Lab F (S8) (P<0.001, 27%); Lab H (S11) (P<0.001, 38%); Lab I (S13) (P=0.020, 1%) and Lab J (S15) (P<0.001, 11%). The percentage of the total variability in scores ‘explained’ by the difference in variability between experiments is shown in brackets. The between-experiment variability also explained an appreciable amount of the total variability in these datasets. The variability between experiments over time can be seen in the control charts.

1. *Between sexes*

In two cases (Lab E (S7) and Lab I (S14)) there were significant differences between the scores for males and females (P=0.002 and P=0.017, respectively). However, the differences were small and in different directions (Male – Female difference: 0.16% and -0.23%, respectively). The between laboratory differences were appreciably greater than that between the sexes, within the same laboratory.

1. *Between strains*

Data were available from just two strains: Wistar and Sprague Dawley. Aside from the four laboratories (Lab D, E, H and I) which reported data on both strains; comparisons between the two stocks in the other laboratories were confounded with other sources of variability between the laboratories. There were some significant differences in MN/1000 in some of the comparisons, with Wistars having slightly higher counts: Lab E (S7) 0.40% difference, P<0.001 (8% of variability); Lab I (S12) 0.23%, P=0.006 (1%) and Lab I (S13) 0.50%, P=0.019 (2%). The differences were quite small and more reflective of the large sample sizes and the consequent power of the statistical test. However, the comparison between the two strains for Lab D (S5&S6, Table 2) showed Sprague Dawley rats with a higher incidence than Wistar rats (0.25%, P<0.001 (6%)). Lab D investigators attributed this small but statistically significant difference to their use of flow cytometry, which facilitated analysis of 20,000 RET for the presence of MN per blood sample.

The Wistar samples were designated as Han:Wistar by five laboratories but were obtained from different suppliers and it was not known when these colonies diverged. The Sprague Dawley samples were designated as Crl:Sprague Dawley from Charles River laboratories but were obtained from several different Charles River facilities. The relationship, if any, between the two strains is not clear.

Both Wistar and Sprague Dawley rats are produced in outbred colonies (usually designated as 'stocks' as opposed to 'strains') and are potentially genetically variable, but from closed colonies (where the breeding programme aims to maintain genetic variability and to reduce the degree of inbreeding). It has long been known that the names of outbred rat stocks provide no information on their genetic profiles or the relatedness of colonies with the same name (Lovell & Festing, 1981). Medawar & Medawar (1983) stated that 'Because they look alike, albino rats and mice are often assumed to be genetically uniform; albinism may conceal what would otherwise be flagrant genetic diversity.' It is not known how much inter-individual variability in such outbred stocks has a genetic component. Inbred strains would provide a better model if some indication of the potential contribution of genetic differences to inter-individual variability was needed.

*Other factors*

The one-way anovas carried out (Table 3) to investigate the effects of the different vehicles used showed results similar to those obtained in the comparisons between experiments. It is difficult to identify which, if any of the many different vehicles and combination of vehicles used in these studies may have contributed to the between-laboratory variability. These differences in vehicles may also be confounded with other differences in the protocols used such as age at initiation of the experiment, numbers of days of and type of dosing carried out. Uncontrolled variation such as changes in personnel and seasonal variation over the length of time covered by these experiments may also have contributed. The relatively small number of independent datasets means that it is difficult to tease out the effects of these variables using more advanced statistical modelling methods.

The lack of major effects of strain, sex and tissues/counting methods on micronuclei counts points to other factors that differ between laboratories and between experiments within laboratories which are contributing to the variability in micronuclei.

Three laboratories (C (S3), D (S5) and E (S7)) showed significantly different MN/1000 values depending upon the age of the animal at the start of the experiment. It is important, however, to appreciate that these differences may result from confounding factors rather than the animal's age. It should also be noted that, although statistically significant, the factor 'age' only 'accounted' for between 3 - 6% of the variability in the values in these samples. Analyses with the number of days of dosing also showed evidence of confounding and the difficulty of disentangling the effects of the two variables using these datasets.

There was no strong evidence that the route of administration affected the incidence of micronuclei. The difference between p.o. and s.c administration in Lab I (S14) was just significant (P=0.028 with s.c. 0.25% higher than p.o.); this was the only significant effect out of ten comparisons and, then, only explaining 4% of variability.

*Zero scores*

The proportions of samples which had no micronucleated cells varied between laboratories with an appreciable number of zeros counts reported in some laboratories especially those with low means. Lab H (S11) reported 315 zero scores out of 1295 rats (24%), Lab E (S7) reported 81 out of 399 (20%); and Lab A (S1) 53 out of 630 (8.4%). On the other hand, a few laboratories (Labs B, D and I) reported no zero values. The frequency of zero scores can be clearly seen in the I-Charts. The presence of many zero values can complicate the interpretation of control charts as the Lower Control Limit may be estimated as a negative value which results in it not being possible to distinguish a zero value as being 'normal' from a value which might cause 'concern'. Such a high incidence of zero values may affect the performance of the statistical tests applied to the data. A run of zero values would, however 'flag up' a possible problem with the data such as a problem with a scorer. These results suggest that those laboratories with low mean micronuclei counts should increase the number of cells they 'score' to reduce the proportion of zeros in their experiments.

*Outliers*

There was one striking ‘outlier’ (105/20000 or 5.25/1000 MN) rat #192 in Lab D's Sprague Dawley sample S5; whereas Lab D mean = 1.08/1000). The respective QC charts show a large ‘spike’. Including or excluding the value made only minor differences to the descriptive statistics. It is not possible to identify from the dataset why this animal is an outlier. There are no obvious explanations of why a rat might be a 'jackpot' animal with a very high frequency of micronuclei: it might have been ill or it might reflect an artefact of some sort in the count of micronuclei. Lab D investigators, however, indicated that a blood sample obtained from the same vehicle-control animal several weeks later was highly elevated as well (4.00/1000 MNT), which suggests this was a biological outlier of an unusual size as opposed to a technical scoring artefact.

*Other data*

Supplementary data were reported on micronuclei counts in other cells and on the toxicity measure, %PE. These data allowed the relationships between these different measures to be explored.

1. *Degree of agreement between micronuclei from PCEs or immature RETS and micronuclei from NCE or mature RETs*

In the case of peripheral blood samples, MN/1000 values derived from the immature RETs were compared with those from the mature ME's in the same animals in five datasets. Four correlations were significant at P<0.001. For datasets Lab D (S5), Lab D (S6), Lab H (S12) and Lab I (S14) the correlations, r, were 0.48, 0.43, 0.85 and, 0.42 respectively (all P<0.001). Lab C (S3), r was -0.03, P = 0.66. In the case of micronuclei in mature cells from the bone marrow samples the correlations with between the micronuclei counts in PCE and NCE cells were significant: Lab C (S4) r= 0.23, P=0.015 and Lab F (S8), r=0.56, P= <0.001. These results show that there was some but not strong degree of agreement between the micronuclei counts from the two sets of cells collected in tests. The lack of strong agreement between MN-RET and MN-NCE values is to be expected because MN-NCE are unlikely to accumulate while splenic filtration is also a factor that limits the association between MN-NCE and MN-RET values.

1. *Differences between sexes and animal’s age on %PCE*

There was significant between-experiment variability in the %PE data in all the laboratories. There was a highly significant difference in %PCE between the sexes in one laboratory (Lab F; P=0.004; males 49.6% v. females 45.7% females). None of the other comparisons between the sexes for %PCE were significant. There was an inconsistent pattern of changes in %PE values in animals of different ages, possibly because of the confounding effects of other variables. The size of these differences was, however, appreciably greater than any effects seen for MN/1000 cell values.

1. *Association between MNPCE and %PCE*

The association between the MN/1000 values derived from PCEs and the %PCE values from the same animals was investigated in 9 bone marrow sample datasets. In seven cases the correlations were small and not significant. There were significant correlations between the two measures in two cases: Lab E (S7) (r= 0.13, P=0.011) and Lab F (S8) (r=0.18, P=0.031). Although significant, the correlations 'explain' less than 5% of variation between the two measures suggesting that there is no strong association between %PE and MN/1000 values within a laboratory's set of results.

*Historical data from other sources*

Most data for the *in vivo* micronucleus assay relate to studies using mice, with fewer publications using rats. The results from the rat strains studied here can be compared with those in mice reported by Hayashi (2016). However, Kasamoto *et al* (2017) reported negative control *in vivo* micronucleus data from male 7-week old Sprague Dawley rats from three series of experiments designed to investigate combining the micronucleus and comet assay. The mean micronuclei combined from 11 experiments in one series was 0.15% (SD 0.11%, n=66). In a second series of just 2 experiments the mean was 0.14% (SD 0 .09%, n=8). Micronuclei data from a third set were only shown in a figure with a mean of approx. 0.1% and no SD or n provided. An unpublished study by ITR Laboratories Canada Inc (ITR) on "Insea2" found on the internet (<http://insea2.com/wp-content/uploads/2017/11/inSea2-in-vivo-rat-micronucleus-test.pdf>) reported a mean micronuclei value for negative (vehicle) control rats of approx. 0.14% (SD 0.035%, n=10). Plots of the company's historical negative control data were also included in the report.

*Implications for power and sample sizes.*

This dataset can be used to investigate the power and sample sizes required in the design of *in vivo* micronuclei studies. The within-group standard deviations in the datasets ranged from 0.25 to 1.16 %MN (Table 2) with a median value of 0.58% MN. Sample sizes of 4 to 5 animals have about an 80% probability (power) of detecting an approximately 2 standard deviation (SD) unit difference between two groups if this difference really exists. Here, based upon the median SD, the difference at the P=0.05 level in a two-sided two-sample t-test would be approximately 1.2 %MN. This gives an indication or 'rule of thumb' of the ‘resolution’ of the experimental design and can help in determining the size of effects which could be considered ‘biologically important’ and distinguishable from 'statistically significant' effects. This is a deliberately simplistic approach to power and sample size estimation and does not take into consideration the relationship between the mean and variance associated with count-based data. Resources for more sophisticated power calculations considering Poisson distributed data can be found in a range of statistical packages.

*Use of control charts*

This study provides a good illustration of the potential of QC charts to identify and, potentially, reduce the variability in results between experiments within a laboratory. The primary benefit of the control charts is transparency in that they provide a visualization of the laboratory's historical control data. They can be used, for instance, to help assess where, although there may be a statistically significant difference between the concurrent control sample and one or more of the treated groups, the historical control data indicates that the size of the effect is below what might be considered biologically relevant or important.

Control charts provide an assessor with a comprehensive overview of the performance of a laboratory. This can help, for instance, with the expert judgment needed to assess whether the variability within the laboratory's results is satisfactory and not so wide that it could lead to 'rewarding' laboratories with poor data with fewer positive results. The charts can also, although not the primary objective in this field, provide the laboratory with a set of techniques (and software tools) for monitoring performance and reducing the degree of between study/experiment variability.

*Transformations*

The datasets in this study can be used with the Box-Cox method to investigate the most appropriate transformation to use. This method is related to Tukey's 'ladder of powers' which applies a power transformation to the 'raw' data to try to change a skewed distribution into one that is nearly normally distributed. The value of the exponent (or power) called lambda (λ) that is applied to the data is directly linked to common transformations: a λ of 2 is a transformation by squaring the original data; of 1, no transformation; of 0.5, the square root; of 0, the logarithmic and, of -1, the reciprocal. These are the 'stepped' changes in Tukey's 'ladder'. The Box-Cox method is an approach for finding the optimum estimate of λ. The data from each of the datasets were analysed by the Box-Cox option in Minitab. This gives a best estimate of λ, lambda, the 95% confidence interval for the estimate and an approximation to the nearest point on Tukey's 'ladder' (Table 4).

A range of λ from -0.07 to 0.54 was obtained for the eight datasets where a Box-Cox analysis was possible. The rounded λ was 0.5 for 4 datasets, indicating, a square root transform, and another 3 the rounded λ was 0 indicating a logarithmic transform. The eighth dataset provided an estimate of λ of 0.23 (equivalent to the 4th root). The Box-Cox transform cannot be carried out with zero or negative values so a small value (0.1) was added to the other seven datasets where there were zero values. The λ's for the seven datasets where there were zero values are also shown in Table 4 in italics.

These transformations were applied to the datasets and used to produce I- and X-bar Charts and did not drastically change the pattern of results. In some cases, individual values close to the upper and lower control lines switched position relative to the line depending upon whether transformed or not. In a sense, this illustrates the importance of using the control charts and control limits as a tool to help with assessment rather than as a formal testing approach where a different decision is made solely depending upon whether a value is just above or below a specific line. The charts here have been produced using the untransformed data. Control charts are not affected appreciably by moderate non-normality. This is especially so for X-bar charts which use group means and benefit from the central limit theorem. Transformation also requires that the data are plotted with the Y axis on the transformed scale and it may be difficult to relate the transformed values that are plotted to the values usually obtained in the laboratory.

*Hyper-Poisson variation*

Count data are often described as Poisson distributed with the property that the variance is equal to the mean. In this case this variability is random. In most animal studies it is not unexpected for laboratories to show between-experiment variability resulting in hyper-Poisson distributions. Although there have been suggestions, particularly with respect to *in vitro* studies, that the between-experiment variability can be effectively removed so that all the remaining variability is 'pure' random this is unlikely to be achieved in animal experiments where many factors can affect results. Given that there will be inter-animal and between-experimental variability it is difficult to specify what is an acceptable level for such variability. It would, for instance, be quite possible for a series of experiment to be 'in control' based on control charts and with no 'unacceptable' results being identified. However, this might be achieved because the variability of the results was large, producing unacceptably wide, from a regulatory perspective, control limits. This problem is not solved by an algorithmic approach but rather by expert judgement (and possibly consensus) that the level of variability produced by a laboratory might be too wide to be acceptable. Collections of data such as the one reported here provide an indication of what can routinely be achieved by good laboratories in the field. This does not mean that attempts should not be made to reduce the variability rather that the objective should be to maintain it an acceptable level. This is somewhat different to how QC is used in some other areas, such as manufacturing, where the object is to reduce variability to a minimum to ensure, for instance, that some precise critical measurement is always attained.

**Conclusion**

This study has identified appreciable variability between and within laboratories in the number of micronuclei detected in the negative (vehicle) control rats used in *in vivo* micronucleus tests. No specific factors have been found that account for this variability. Differences were detected in some laboratories between the sexes and the strains used but these effects were small. Large differences between experiments were identified but could not be traced to any specific factors. Some differences may be related to the different vehicles used. The study indicates that there remain factors either in the animal's environment or the study protocol still to be identified that vary over time so contributing to the variability. The increasing use of the rat in this assay especially in conjunction with the investigation of other toxicological endpoints suggests that it is important to characterize the background levels of the endpoint especially as results from this assay forms part of the profile of a test material. The data reported here should provide both the new entrant into the use of the assay and the assessor of results with a range of values where control values would be expected to fall and the degree of within-laboratory variability that is found using the test.

**Acknowledgments**

The authors wish to thank the participating laboratories and their staff who provided data. The authors also thank the HESI GTTC and the Data Interpretation Workgroup for intellectual and financial support. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**References**

# COM (2011) A strategy for testing of chemicals for genotoxicity: 5 guidance statements on testing chemicals with no or inadequate genotoxicity data. UK Committee on Mutagenicity of Chemicals in Food, Consumer Products and the Environment. <https://www.gov.uk/government/publications/a-strategy-for-testing-of-chemicals-for-genotoxicity>

ICH Guideline (2011) *ICH S2(R1)* Guidance on genotoxicity testing and data interpretation for pharmaceuticals intended for human use. From: <http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Safety/S2_R1/Step4/S2R1_Step4.pdf>.

Jarvis, P., Saul, J., Aylott, M., Bate, S., Geys, H. & Sherington, J. (2011) An assessment of the statistical methods used to analyse toxicology studies. Pharmaceutical Statistics, 10 477-484

Hayashi, M. (2016) The micronucleus test - most widely used in vivo genotoxicity test - Genes and Environment*.* 38 18.

Hayashi, M., Yoshimura, I., Sofuni, T. & [Ishidate Jr, M.](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Ishidate%2C+M) (1989) A procedure for data analysis of the rodent micronucleus test involving a historical control. Environmental and Molecular Mutagenesis 13 347-356

Hayashi, M., Hashimoto, S., Sakamoto, Y., Hamada, C., Sofuni, T. and Yoshimura, I. (1994). Statistical analysis of data in mutagenicity assays: rodent micronucleus assay. Environmental Health Perspectives, 102 Suppl 1, 49-52.

Heddle, J.A. (1973), A rapid *in vivo* test for chromosomal damage. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 18 187-190

## Kasamoto, S., Masumori, S., Tanaka, J., Ueda, M., Fukumuro, M., Nagai, M., Yamate, J. & Hayashi, M. (2017) Reference control data obtained from an *in vivo* comet-micronucleus combination assay using Sprague Dawley rats. Experimental and Toxicologic Pathology [69](https://www.sciencedirect.com/science/journal/09402993/69/4)  187-191

Kim, B.S., Cho, M. and Kim, H.J. (2000). Statistical analysis of *in vivo* rodent micronucleus assay. Mutation Research 469 233-241

Kilkenny C, Browne W.J., Cuthill, I.C., Emerson, M. & Altman, D.G. (2010) Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. PLoS Biol 8 e1000412. <https://doi.org/10.1371/journal.pbio.1000412>

Lovell, D.P., Anderson, D., Albanese, R., Amphlett, G.E., Claire, G., Ferguson, R., Richold, M., Papworth, D.G. and Savage, J.R.K. (1989). Statistical Analysis of *In vivo* Cytogenetic Assays In: D.J. Kirkland (Ed.). Statistical Evaluation of Mutagenicity Test Data. UKEMS Sub-Committee on Guidelines for Mutagenicity Testing, Report, Part III. Cambridge University Press, Cambridge, UK, pp. 184-232

Lovell, D.P. & Festing, M.F.W. (1981) Relationships among colonies of the laboratory rat (Rattus norvegicus) based on morphological analysis of mandible shape. Journal of Heredity. 73 81-82

Lovell, D.P., Fellows, M., Marchetti, F., Christiansen, J., Elhajouji, A., Hashimoto, K.,  Kasamoto, S., Li, Y., Masayasu, O.,Moore, M.M., Schuler, M., Smith,R., Stankowski, Jr., L.F., Tanaka, J., Tanir, J.Y., Thybaud, V., Van Goethem, F., & Whitwell, J. (2018a) Analysis of negative historical control group data from the *in vitro* micronucleus assay using TK6 cells. Mutation Research/Genetic Toxicology and Environmental Mutagenesis 825 40-50

Lovell, D.P., Fellows, M., Elhajouji, A., Farabaugh, C.S. Gilby, B.G., Hashimoto, K., Li, Y., Roy, S., Schuler, M., Whitwell, J. & Tanir, J.Y. (2018b) Analysis of historical negative control group data from the in *vitro* micronucleus assay using human lymphocytes. Mutation Research/Genetic Toxicology and Environmental Mutagenesis (In press)

Medawar, P. & Medawar, J. (1983) Aristotle to Zoos. An authoritative, witty guide to central topics in biology. Oxford University Press, Oxford

OECD (2016a) TG 474 Mammalian erythrocyte micronucleus test. OECD Test Guideline TG-474. 2014. <http://www.oecd.org/env/ehs/testing/draft_tg474_second_commenting_round.pdf>

OECD (2016b*)* Overview of the set of OECD Genetic Toxicology Test Guidelines and updates performed in 2014-2015Series on Testing & Assessment No. 238, OECD Publishing, Paris. http://www.oecd-ilibrary.org/docserver/download/9717131e.pdf?expires=1520263341&id=id&accname=guest&checksum=1CBCB4880A409F8F46AD7994AFDB13EE

Schmid, W. (1975), The micronucleus test. Mutation Research 31 9-15

**Table titles and Legends to Figures**

Table 1 Information requested from participating laboratories on animal husbandry, experimental variables and laboratory procedures

Table 2 Table of means and SD for each combination (C1-C33) of sex, strain and tissue from 15 datasets (S1-S15) derived from 10 laboratories (A-J)

Table 3 Results of one-way analyses of variance carried out to identify effects of various factors on variability within 15 datasets (S1-S15)

Table 4 Box -Cox lambda values for datasets S1 – S15

Figure 1 to 4 Figures plotting mean and standard deviation (or 95% Confidence Intervals) for 33 combinations of laboratories, sex, strain and site of sampling. Graphs use different colour to provide contrasts for comparisons,

Fig 1 Between Laboratories: Means ± SD

Fig 2 Mean ±95%CI: Black, Female; Grey, Male; Open Circle, Sex unknown

Fig 3 Mean ±95%CI: Black, Wistar; Grey, Sprague Dawley

Fig 4 Mean ±95%CI: Black, Bone Marrow; Grey, Peripheral Blood

Appendix 1 Fig I.1 – Fig I.15 I-Charts

**Table 1 Information requested from participating laboratories on animal husbandry, experimental variables and laboratory procedures**

Information was requested on:

Age of rat

Strain of rat: e.g. Sprague Dawley / Wistar

Sex of rat

Solvent/vehicle

Scoring method

Route of administration

Number of doses

Sample day

Filtered v. unfiltered

Age at start of treatment

Slide staining method (AO Giemsa))

Mountant (DPX, Wet mount)

Microscopy

Husbandry conditions: Temperature, Relative Humidity, air changes, Light/Dark cycle

**Table 2 Table of means and SD for each combination (C1-C33) of sex, strain and tissue from 15 datasets (S1-S15) derived from 10 laboratories (A-J)**

**Comb Lab Sex Gen Tiss N Mean/1000 SD Expts Dataset**

 1 A Female W BM 108 1.32 1.02 97 S1

 2 A Male W BM 522 1.42 1.02

 3 B Female SD PB 119 1.19 0.52 44 S2

 4 B Male SD PB 169 1.21 0.43

 5 C Female SD PB 90 1.37 0.48 17 S3

 6 C Male SD PB 166 1.44 0.55

 7 C Female SD BM 35 2.14 1.03 16 S4

 8 C Male SD BM 72 2.22 1.16

 9 D Male SD PB 212 1.08 0.44 35 S5

10 D Male W PB 63 0.83 0.25 12 S6

11 E Female SD BM 90 0.53 0.40 36 S7

12 E Male SD BM 247 0.44 0.36

13 E Female W BM 28 1.00 1.02 5

14 E Male W BM 34 0.75 0.56

15 F Female W BM 5 1.40 0.28 28 S8

16 F Male W BM 123 1.73 0.83

17 F Unknown W BM 20 1.59 0.44

18 G Female W BM 35 0.92 0.70 24 S9

19 G Male W BM 114 1.13 0.67

20 G Female W BM 29 0.77 0.49 23 S10

21 G Male W BM 80 0.93 0.68

22 H Male SD BM 1250 0.61 0.55 250 S11

23 H Male W BM 45 0.84 0.58 9

24 H Female SD PB 15 0.50 0.25 13 S12

25 H Male SD PB 48 0.66 0.42

26 I Female W BM 6 1.50 0.84 25 S13

27 I Male W BM 3 2.00 0.50

28 I Female SD BM 92 1.22 0.62

29 I Male SD BM 91 1.12 0.59

30 I Female SD PB 45 0.93 0.51 9 S14

31 I Male SD PB 45 1.16 0.35

32 J Female W BM 90 1.86 1.16 50 S15

33 J Male W BM 264 1.67 1.04

Gen (Strain): W, Wistar; SD, Sprague Dawley

Tiss(ue): BM, Bone Marrow; PM, Peripheral Blood

**Table 3 Results of one-way analyses of variance carried out to identify effects of various factors on variability within 15 datasets (S1-S15)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Set | Lab | Sex | Strain | Route | Doses | Days | Vehicle | Age | Experiment |  Set |
|   |   | P | % | P | % | P | % | P | % | P | % | P | % | P | % | P | % |   |
| S1 | A | 0.36 | 0 | N/A |   | N/A |   | 0.001 | 2 | 0.32 | 0 | <0.001 | 9 | N/A |   | <0.001 | 14 | 1 |
| S2 | B | 0.70 | 0 | N/A |   | N/A |  | 0.14 | 1 | 0.24 | 1 | 0.10 | 4 | 0.18 | 1 | 0.15 | 4 | 2 |
| S3 | C | 0.27 | 0 | N/A |   | 0.13 | 1 | 0.14 | 2 | 0.06 | 2 | <0.001 | 11 | <0.001 | 6 | <0.001 | 13 | 3 |
| S4 | C | 0.71 | 0 | N/A |   | N/A |   | 0.15 | 1 | 0.08 | 4 | 0.35 | 1 | N/A |   | 0.67 | 0 | 4 |
| S5 | D | N/A |   | N/A |   | 0.23 | 0 | N/A |   | N/A |   | 0.38 | 0 | 0.008 | 3 | 0.56 | 0 | 5 |
| S6 | D | N/A |   | N/A |   | N/A |   | N/A |   | N/A |   | 0.05 | 8 | N/A |   | 0.28 | 4 | 6 |
| S7 | E | 0.002 | 2 | <0.001 | 8 | 0.81 | 0 | <0.001 | 9 | <0.001 | 9 | <0.001 | 15 | 0.002 | 3 | <0.001 | 16 | 7 |
| S8 | F | 0.50 | 0 | N/A |   | 0.53 | 0 | N/A |   | <0.001 | 24 | 0.24 | 71 | 0.11 | 2 | <0.001 | 27 | 8 |
| S9 | F | 0.11 | 1 | N/A |   | 0.30 | 0 | N/A |   | N/A |   | 0.16 | 4 | 0.87 | 0 | 0.17 | 5 | 9 |
| S10 | G | 0.24 | 0 | N/A |   | 0.17 | 2 | 0.20 | 2 | 0.20 | 2 | 0.20 | 6 | 0.51 | 0 | 0.21 | 5 | 10 |
| S11 | H | N/A |   | 0.006 | 1 | 0.15 | 0 | 0.001 | 1 | 0.10 | 0 | <0.001 | 24 | N/A |   | <0.001 | 38 | 11 |
| S12 | I | 0.18 | 1 | N/A |  | 0.83 | 0 | 0.24 | 1 | N/A |   | <0.001 | 63 | N/A |   | N/A |   | 12 |
| S13 | I | 0.34 | 0 | 0.019 | 2 | 0.65 | 0 | <0.001 | 12 | 0.014 | 10 | 0.011 | 8 | 0.08 | 2 | 0.02 | 9 | 13 |
| S14 | I | 0.017 | 5 | N/A |  | 0.028 | 4 | 0.10 | 2 | 0.25 | 3 | 0.20 | 3 | 0.12 | 2 | 0.33 | 1 | 14 |
| S15 | J | 0.14 | 0 | N/A |   | N/A |  | 0.02 | 2 | 0.08 | 1 | 0.013 | 6 | 0.70 | 0 | <0.001 | 12 | 15 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

S1-S15 designates the different datasets. A-J designates the different laboratories. N/A Statistical test not applicable

**Table 4 Box -Cox lambda values for datasets S1 – S15**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
| **Dataset** | **Lambda λ** | **LCLλ** | **UCLλ** | **Rounded λ** |
| S1 | *0.40* | *0.31* | *0.50* | *0.40* |
| S2 | 0.31 | 0.07 | 0.54 | 0.50 |
| S3 | 0.21 | -0.01 | 0.47 | 0 |
| S4 | *0.51* | *0.25* | *0.84* | *0.50* |
| S5 | -0.07 | -0.29 | 0.13 | 0 |
| S6 | 0.04 | -0.55 | 0.74 | 0 |
| S7 | *0.36* | *0.25* | *0.49* | *0.36* |
| S8 | 0.28 | 0.05 | 0.51 | 0.50 |
| S9 | *0.52* | *0.30* | *0.72* | *0.50* |
| S10 | *0.54* | *0.30* | *0.77* | *0.50* |
| S11 | *0.25* | *0.19* | *0.20* | *0.25* |
| S12 | 0.28 | -0.08 | 0.65 | 0.50 |
| S13 | 0.23 | 0.01 | 0.47 | 0.23 |
| S14 | 0.44 | 0.15 | 0.69 | 0.50 |
| S15 | *0.51* | *0.39* | *0.64* | *0.50* |

Values in italics are based upon the value +0.1 to overcome that the Box-Cox method does not work with zero values

LCL lower 95% confidence limit for λ

UCL upper 95% confidence limit for λ

**Figure 1 to 4 Figures plotting mean and standard deviation (or 95% Confidence Intervals) for 33 combinations of laboratories, sex, strain and site of sampling. Graphs use different shades of black and grey to provide contrasts for comparisons.**

**Fig 1 Between Laboratories: Means ±SD**



**Fig 2 Mean ±CI: Black, Female; Grey, Male; Open circle, Sex unknown**



**Fig 3 Mean ±CI: Black, Wistar; Grey, Sprague Dawley**



**Fig 4) Mean ±CI: Black, Bone Marrow; Grey, Peripheral Blood**



**Appendix 1 Fig I.1 – Fig I.15 I-Charts**

**Fig I.1 Laboratory A S1 BM**



**Fig I.2 Laboratory B S2 PB**



**Fig I.3 Laboratory C S3 PB**



**Fig I.4 Laboratory C S4 BM**



**Fig I.5 Laboratory D S5 PB Sprague Dawley**



**Fig I.6 Laboratory D S6 PB Wistar**



**Fig I.7 Laboratory E S7 BM**



**Fig I.8 Laboratory F S8 BM**



**Fig I.9 Laboratory G S9 BM Single dose**



**Fig I.10 Laboratory G S10 BM Repeat dose**



**Fig I.11 Laboratory H S11 BM**



**Fig I.12 Laboratory H S12 PB**



**Fig I.13 Laboratory I S13 BM**



**Fig I.14 Laboratory I S14 PB**



**I.15 Laboratory J S15 BM**

