Mutat Res Gen Tox En 825 (2018) 40–50

Contents lists available at ScienceDirect

Mutat Res Gen Tox En

journal homepage: www.elsevier.com/locate/gentox

Analysis of negative historical control group data from the *in vitro* micronucleus assay using TK6 cells

David P. Lovell[a,*], Mick Fellows[b], Francesco Marchetti[c], Joan Christiansen[d], Azeddine Elhajouji[e], Kiyohiro Hashimoto[f], Sawako Kasamoto[g], Yan Li[h,1], Ozaki Masayasu[i], Martha M. Moore[j], Maik Schuler[k], Robert Smith[l], Leon F. Stankowski Jr.[m], Jin Tanaka[g], Jennifer Y. Tanir[n], Veronique Thybaud[o], Freddy Van Goethem[p], James Whitwell[l]

[a] *St George's Medical School, University of London, London, SW17 0RE, UK*
[b] *Astra Zeneca, Drug Safety and Metabolism, Cambridge, CB4 0WG, UK*
[c] *Environmental Health Science Research Bureau, Health Canada, Ottawa, ON, K1A 0K9, Canada*
[d] *Department of Exploratory Toxicology, H. Lundbeck A/S, Ottiliavej 9, DK-2500, Valby, Denmark*
[e] *Novartis Institutes for Biomedical Research, Preclinical Safety, Basel, Switzerland*
[f] *Takeda Pharmaceutical Company Limited Drug Safety Research Laboratories, Pharmaceutical Research Division 26-1, Muraoka-Higashi 2-chome, Fujisawa, Kanagawa, 251-8555, Japan*
[g] *Genotoxicology Laboratory, Public Interest Incorporation Foundation Biosafety Research Center (BSRC), 582-2, Shioshinden, Iwata, Shizuoka, 437-1213, Japan*
[h] *U.S. Food and Drug Administration, National Center for Toxicological Research, USA*
[i] *Canon, Inc., Quality Management Headquarters, Chemical Safety Management Division, Chemical Safety Evaluation Department 1, Japan*
[j] *Ramboll Environ, Little Rock, AR, 72201, USA*
[k] *Pfizer Inc., Groton, CT, USA*
[l] *Covance Laboratories Ltd, Harrogate, HG3 1PY, UK*
[m] *Charles River Laboratories Skokie, LLC, Skokie, IL, USA*
[n] *ILSI Health and Environmental Sciences Institute (HESI), 1156 15th Street NW, 2nd Floor, Washington, DC 20005, USA*
[o] *Sanofi, Drug Disposition, Safety and Animal Research, Vitry-sur-Seine, France*
[p] *Discovery Toxicology & Translational Safety Sciences, Janssen R&D, Turnhoutseweg 30, 2340 Beerse, Belgium*

ARTICLE INFO

Keywords:
TK6 cells
In vitro micronucleus assay
Historical negative control data
Quality control statistics

ABSTRACT

The recent revisions of the Organisation for Economic Co-operation and Development (OECD) genetic toxicology test guidelines emphasize the importance of historical negative controls both for data quality and interpretation. The goal of a HESI Genetic Toxicology Technical Committee (GTTC) workgroup was to collect data from participating laboratories and to conduct a statistical analysis to understand and publish the range of values that are normally seen in experienced laboratories using TK6 cells to conduct the *in vitro* micronucleus assay. Data from negative control samples from *in vitro* micronucleus assays using TK6 cells from 13 laboratories were collected using a standard collection form. Although in some cases statistically significant differences can be seen within laboratories for different test conditions, they were very small. The mean incidence of micronucleated cells/1000 cells ranged from 3.2/1000 to 13.8/1000. These almost four-fold differences in micronucleus levels cannot be explained by differences in scoring method, presence or absence of exogenous metabolic activation (S9), length of treatment, presence or absence of cytochalasin B or different solvents used as vehicles. The range of means from the four laboratories using flow cytometry methods (3.7-fold: 3.5–12.9 micronucleated cells/1000 cells) was similar to that from the nine laboratories using other scoring methods (4.3-fold: 3.2–13.8 micronucleated cells/1000 cells). No laboratory could be identified as an outlier or as showing unacceptably high variability.

Quality Control (QC) methods applied to analyse the intra-laboratory variability showed that there was evidence of inter-experimental variability greater than would be expected by chance (*i.e.* over-dispersion). However, in general, this was low.

This study demonstrates the value of QC methods in helping to analyse the reproducibility of results, building

* Corresponding author.
E-mail address: dlovell@sgul.ac.uk (D.P. Lovell).
[1] Currently employed by Covance Central Laboratory Services, Indianapolis, IN 46214, USA.

https://doi.org/10.1016/j.mrgentox.2017.10.006
Received 8 August 2017; Received in revised form 23 October 2017; Accepted 24 October 2017
Available online 10 November 2017

up a 'normal' range of values, and as an aid to identify variability within a laboratory in order to implement processes to maintain and improve uniformity.

1. Introduction

The importance of historical control data was discussed by a working group of the International Workshop on Genotoxicity Testing (IWGT) at a meeting in Basel in 2009. Various recommendations were made by the group for the use of historical control data [1]. They focused mainly on historical negative control data pointing to its use in determining the acceptability of the experimental (concurrent) negative control in the test and as evidence of a laboratory's competency for conducting an assay. The recommendations made in this IWGT paper included: (1) the minimum sets of data needed for the creation of historical control datasets, (2) consideration of the distribution of the data rather than the simple ranges and (3) consideration of how such data can assist in the interpretation of results.

Recent revisions of the Organisation for Economic Co-operation and Development (OECD) Test Guidelines for the Testing of Chemicals now include an increased emphasis on the use of historical negative control data in the assessment of genotoxicity test results [2]. They include recommendations on how to build an historical control database. For example, the OECD TG 487 for the *in vitro* Mammalian Cell Micronucleus Test [3] states that:

> *"When first acquiring data for an historical negative control distribution, concurrent negative controls should be consistent with published negative control data where they exist. As more experimental data are added to the control distribution, concurrent negative controls should ideally be within the 95% control limits of that distribution. The laboratory's historical negative control database should initially be built with a minimum of 10 experiments but would preferably consist of at least 20 experiments conducted under comparable experimental conditions. Laboratories should use quality control methods, such as control charts (e.g. C-charts or X-bar charts), to identify how variable their positive and negative control data are, and to show that the methodology is 'under control' in their laboratory."*

The concurrent negative control data is used to assess whether the experiment meets acceptability criteria based on whether the negative control is considered to be acceptable for addition to the laboratory historical control database. The criteria for the evaluation and interpretation of results states that, in addition to an evaluation of the statistical significance and dose response, at least one or more data points must fall outside the distribution of the historical negative control. This later criterion provides a means to assess the biological significance of the results.

The OECD test guidelines also provide guidance on the acceptance criteria for inclusion of data into the historical negative control database. They recommend comparisons of the test results with the historical negative control data as one of the three conditions necessary for a definitive clear positive or negative result.

As a part of the deliberations during the last revision of the *in vitro* test guidelines, the OECD collected a small amount of historical control data for a number of *in vivo* and *in vitro* genotoxicity tests [4]. Data were collected from a number of laboratories following a call from the OECD for historical control data from different tests including the *in vitro* micronucleus test to help answer questions related to optimal cell numbers and sample size issues, based upon the expected statistical power and background incidence of genotoxic events in the negative controls. During the course of the OECD revision discussions, a number of limitations of this data set were identified including that reporting of the data was non-standardized and some of the data sets were quite variable. Furthermore, the analysis of those data sets showed what appeared to be appreciable within and between laboratory variability [4]. In view of these limitations, the HESI Genetic Toxicology Technical Committee (GTTC) Data Interpretation Workgroup started a project to collect and collate a set of data into a standardized database from well-established laboratories.

To begin the project, the workgroup chose the *in vitro* micronucleus test using the human lymphoblastic (B-cell) TK6 cells as the prototype for collection of standardized control data and to develop methods to assess these data. The TK6 cell line was chosen because of the potentially large amount of relevant data that could be collected. The TK6 line is a standard and widely used cell line for the *in vitro* micronucleus assay as it can be grown in suspension, does not need trypsinization, has acceptable growth, is of human origin, has good karyotypic stability and has retained its 'wild-type' p53 competency [5]. Lorge et al. [6] provides a detailed description of the origin and establishment of the TK6 cell line and the subsequent deposit of well characterized stocks into cell banks. The paper provides guidelines for TK6 cell maintenance and characteristics of the cells such as karyotype, p53 status, cell growth and doubling times. The authors make recommendations for TK6 cell culture conditions, preservation and quality checks for the cell line. 'Banks' of the cells have been set up at the Japanese Collection of Research Bioresources (JCRB) Cell Bank, Japan and subsequently at the European Collection of Authenticated Cell Cultures (ECACC), UK.

The HESI GTTC Data Interpretation Workgroup put out calls in 2013 and 2014 for interested laboratories to submit data. A specially designed Excel spreadsheet was created for the collection of the data in a defined and uniform way. The Excel spreadsheet was successfully trialled by one of the participating laboratories before being sent out to the other participants. Participating laboratories were asked to complete the spreadsheet and provide answers to a series of specific questions developed by a Management Team on aspects of the conditions used in their studies.

The goals of the project were: (1) to identify the range of data collected by proficient laboratories; (2) to see if any of the different 'factors' in the conduct of studies affect the variability; (3) to see whether the historical negative control data could provide information to make a recommendation of the appropriate number of experiments needed to build a historical negative control database; and (4) to establish acceptable ranges of negative control database. This analysis and publication are intended to provide high quality negative control data for the *in vitro* micronucleus assay in TK6 cells carried out under OECD guidelines. These will assist with the interpretation of test results and with considerations of the type of data needed for the assessment and application of expert judgement in the interpretation.

2. Materials and methods

Thirteen laboratories (four from the USA, three from Japan and six from Europe) participated. Four (C, D, J, M) used flow cytometry methods while nine (A, B, E, F, G, H, I, K, L) used other scoring methods including manual counting using light (4) or fluorescent microscopes (3), automated fluorescence microscopy using the Cellomics ArrayScan® VTi HCS Reader (1) and what was described as microscopy using image analysis (1). No laboratory provided data using more than one method on the same material. Data were collected from experiments carried out between March 2003 and August 2014 although the time frame was, in general, much narrower for the individual laboratory. For instance, one laboratory has a range from 2003 to 2014 while others spanned less than a year.

2.1. Data collection and management

Data were collected using an Excel template with specific fields to be populated (Table 1). In general, few problems were experienced in the collection and management of the data. A small number of laboratories provided data in a non-standard format (these data were not included because the data on individual experiments could not be extracted).

Results were obtained for three main types of experiments: short treatment in the absence of S9 followed by a recovery period (−S9 short), short treatment in the presence of S9 followed by a recovery period (+S9 short) and long treatment in the absence of S9 without recovery (−S9 long). In most cases the duration of treatment is 3 h for the short treatment and 24 h for the long treatment, and the most commonly used recovery period is 24 h, but this did vary from laboratory to laboratory. A small amount of data were collected from 24 h treatment in the absence of S9 followed by a recovery period longer than 24 h (−S9 longer). Two of the four laboratories (D, M) using flow cytometry methods only reported data from studies without the addition of S9.

The data sets from all 13 laboratories (currently anonymized as A to M) were broken down into 55 separate combinations of the various conditions (*e.g.* ± S9, time, scoring methods and vehicles). Most laboratories provided information at the individual replicate culture level allowing assessment of the variability between replicate cultures within a laboratory. Other laboratories reported just one replicate per experiment. Two laboratories (F, I) pooled counts over two replicates and consequently there was no opportunity to assess variability between replicate cultures for these laboratories. Data on 4642 replicates were provided by the 13 laboratories, mainly in the form of defined multiples of 1000 cells. Of these replicates, 1090 were from the four laboratories using flow cytometry. In some cases, there were a variable number of replicates within a single experiment. In one experiment, Laboratory A reported there were 25 replicates.

Data, in general, were reported in one of the following formats:

- Number of micronucleated cells for exactly 1000 (in a small number of cases 2000) cells;
- Number of micronucleated cells from approximately 1000 cells;
- Number of micronucleated cells from a variable number of cells > 1000 (non-flow cytometry methods);
- Number of micronucleated cells from a variable number of cells > 1000 (flow cytometer method).

Different solvents were used as vehicles including: water, dimethyl sulfoxide (DMSO), phosphate buffered saline (PBS), saline, 10% Donor Horse Serum − Roswell Park Memorial Institute (DHS-RPMI) and a vehicle just described as “Medium”. The number of replicates/experimental condition ranged from 2 to 795 (40–461 for the laboratories using flow cytometry). The number of cells/replicate scored by the nine laboratories using non-flow cytometry methods ranged from 1000 to > 2000 and 1238 to 21,772 for the four laboratories using flow cytometry methods.

No restrictions were placed upon the amount of data that could be submitted except that laboratories were expected to have data from a minimum of 20 experiments, which is often considered the lower limit for the use of Quality Control methods [7]. In general, this was achieved, although there were a number of cases where data were derived from a smaller number of experiments because the number of experiments carried out using particular combinations of conditions (such as using a specific vehicle) was small. In all there were eight combinations where data from only a small number of experiments (less than 10 replicates including four combinations with just two replicates) were available. These small numbers were accepted, however, because these could be combined to provide higher numbers of replicates useful in other analyses. However, combinations with just one experiment with two replicates were not included in any of the analyses.

2.2. Statistical methods

Data were analysed using the statistical procedures available in Minitab (Minitab 16 Statistical Software. Minitab, Inc., State College, PA) and R [8]. Data from replicates of exactly 1000 cells were analysed as integer counts. Data from replicates with variable numbers of cells (especially much greater than 1000) were analysed as proportions of micronucleated cells. In some analyses, proportions were multiplied by 1000 to give estimates of the number of micronucleated cells/1000 cells scored.

The specific procedures were: one-way and nested analyses of variances, tests for extra-binomial variation (*i.e.* Goodness of fit to a Poisson distribution), correlations between −S9 short and +S9 short experiments (where applicable) and calculation of tolerance intervals.

A number of QC methods were used: C-Charts for Poisson counts, I-Charts for individual replicate values of counts or proportions of micronucleated cells, p-Charts for proportions when the number of cells counted differed between replicates and X-bar Charts when there were a number of replicates per experiment. Control charts are plots of data collected over a period of time with decision lines added. They are long-established and widely-used methods in industry to monitor the variability of samples and to show that their processes are ‘under control’ rather than drifting over time. There are a number of textbooks describing these methods [7,9–11].

C-Charts are a version of the control chart specifically used for count data. They are based upon the Poisson distribution but with a constant denominator (n) and make use of the relationship in a Poisson distribution between the mean and the variance. The upper and lower control limits (UCL and LCL) are derived from the equation: cbar ± 3√cbar, where cbar is the estimate of the “long-term process mean” derived from the initial development of the control chart. Fig. 1 is an example of a C-Chart for Laboratory L (L1). C-Charts are presented here with the mean of all replicates designated by a horizontal continuous line. Hashed lines represent the mean ± 1SD, ± 2SD and ± 3SD respectively based upon the Poisson variance which is equivalent to the

Table 1
Specific information requested from participating laboratories.

General information
Laboratory
Guideline
Time period during which above data was collected
Cells
Cell line
Origin of the cell line
Maintenance of the cell line
Donor information
Stimulation conditions in case of lymphocytes
Treatment conditions
Treatment schedule short-term treatment +S9
Treatment schedule short-term treatment −S9
Treatment schedule long-term treatment −S9
Time of incubation with cytochalasin B
Cytochalasin B concentration
Solvent
Number of cultures per experiment
Metabolic activation
With or without
Origin
Concentration
Scoring
Method
Staining
Number of cells scored per treatment condition
Number of cells scored/culture
Scoring method

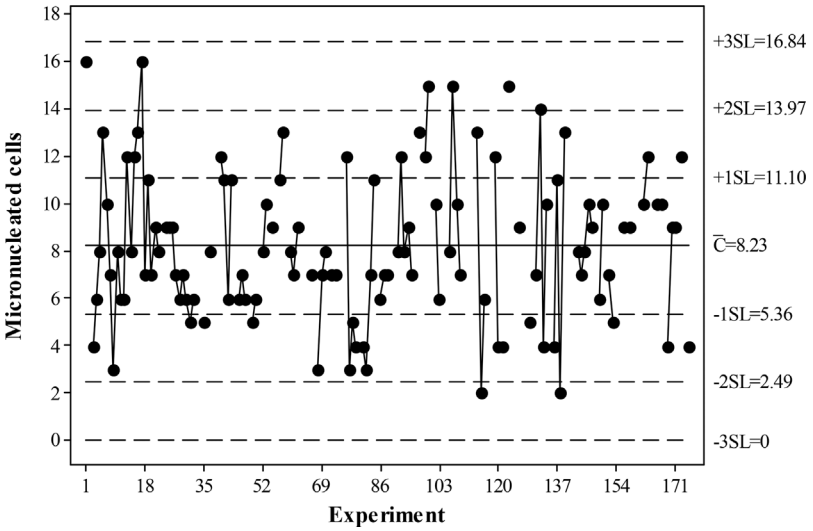


Fig. 1. Example of a C-Chart for Laboratory L (L1).
Plot of counts of micronucleted cells from 175 experiments (−S9 short, most with single replicate). Lack of connecting lines between points in places relate to cases where particular combination was not included in experiment.

mean. (Note that the Minitab package which produces the horizontal lines on the figures refers to these lines as SL1, SL2, and SL3 respectively.) The UCL and LCL are equivalent to the ± 3SD lines and the upper and lower warning limits (UWL and LWL) are equivalent to the ± 2SD lines.

I-Charts are a "time-ordered sequence" plotting individual values along the Y axis and the order of the individuals on the X axis. Fig. 2 is an example of an I-Chart for Laboratory B (B1). **X-bar Charts** are plots of the means of sets of replicates while **p-Charts** are plots of the proportions of 'events' in a series of replicates where the cell numbers in the replicate may differ. Fig. 3 is an X-bar Chart for Laboratory C (C1) and Fig. 4 is a p-Chart for Laboratory L (L2). Similar control and warning limits to those derived for the C-charts can be produced for these other QC charts.

In the QC charts, the points on the chart marked in grey are points which have been 'flagged' as 'out of control' based upon either the Western Electric or Nelson rules which highlight points outside the UCL or LCL or where various numbers of points are, for instance, in an increasing progression. In the case of C- and p-Charts there are four tests (Western Electric rules) while for the I- and X-Bar Charts there are eight tests (Nelson rules) [10]. The scales for QC-charts are based upon the number given to the replicate rather than dates so they do not reflect the relative length of time the laboratory was undertaking the test.

3. Results

All the participating laboratories that provided data in the standard format were diligent in their presentation of useable data. The data from all of these laboratories (and all sets of conditions) were considered acceptable to include in the analyses. The quality of the data was good.

Table 2 shows the mean and standard deviations (SD) for laboratories A to M for each of the 55 separate combinations of the various ± S9, time, scoring methods and vehicles. The Supplementary Results provide the summary results in different formats to facilitate comparisons.

Fig. 5 shows the mean and standard deviation for each combination giving an indication of the relative variability in each set of replicates whereas Fig. 6 shows the mean and its associated 95% confidence interval (CI) for each set giving an indication of the precision of the mean values. The varying numbers of replicates is reflected in the widths of the CIs (which provide an estimate of precision and are a function of sqrt(n) in the calculation of the standard errors) so that those combinations with very small n's will have very wide CIs (note that in Table 2, 21 combinations had less than 20 replicates and 12 of these had 10 or less replicates.) Non-overlapping CIs between laboratories with small differences in means can be identified when these CIs are based upon a large number of replicates (n); however, this is more a consequence of these large n's than representing biologically important differences.

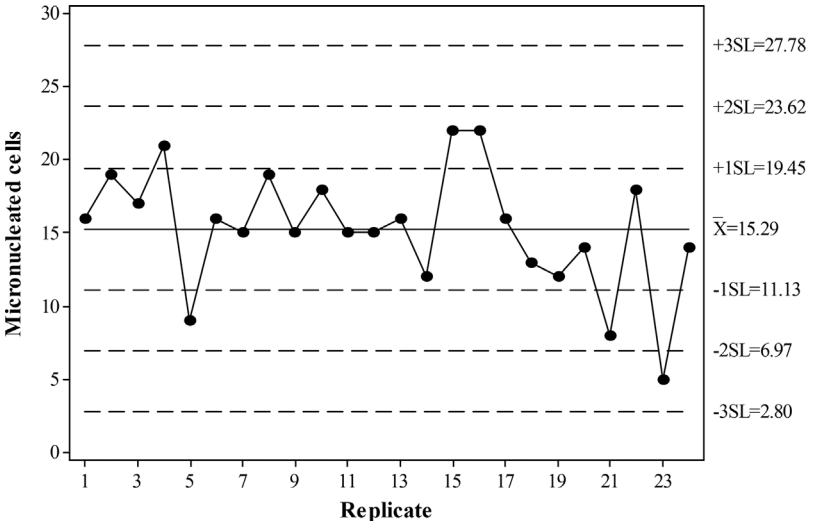


Fig. 2. Example of an I-Chart for Laboratory B (B1).
Plot of counts of micronucleated cells from 12 experiments where 2000 cells were scored from each of two replicates.

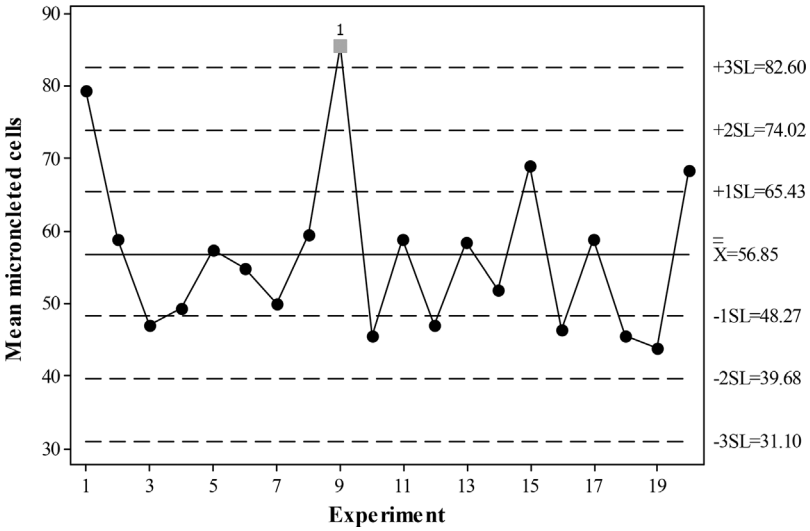


Fig. 3. Example of an X-bar Chart for Laboratory C (C1).
Plot of mean counts of micronucleated cells from 20 experiments where 10000 cells were scored from each of two replicates (scoring by flow cytometer).

Fig. 5 also shows the 'breakdown' of the means by scoring method with the flow cytometry-based methods being shown in grey and the other methods shown in black. Fig. 6 shows the breakdown based upon the presence or absence of S9 (+S9 or −S9). The mean values for the 55 combinations were 7.06/1000 (SD 2.35/1000). Fig. 7 shows the distribution of the means. There was no evidence of any differences in the means of the flow cytometry scoring laboratories compared with the others, or between the presence and absence of S9 (results shown in Supplementary Results).

Table 2 and Fig. 7 show that there was inter-laboratory variability in estimates of micronucleated cells/1000 cells, with statistically significant ($P < 0.001$) greater variability between laboratories than within based upon analyses of variance (results shown in Supplementary Results). However, there were no dramatic outliers. The mean values for the 55 combinations ranged from 3.20/1000 for one combination for Laboratory E to 13.83/1000 for one combination for Laboratory K. In general, mean values for different conditions were similar within the same laboratory. An exception was Laboratory K where the mean values from the five conditions ranged from 5.50/1000 to 13.83/1000. There were variable levels of intra-laboratory variability in the participating laboratories. Laboratory J showed appreciably more variability than the other laboratories. All the laboratories had mean scores of over 3.2/1000. Based upon a Poisson distribution very few zero values would be expected with these levels of mean scores. In fact, only two replicates (from E6 and G4) had zero counts.

Formal comparison between the laboratories using flow cytometry or other scoring methods was difficult to assess because of the small number of laboratories participating. Furthermore, no laboratory used more than one method for the same series of experiments. There was some overlap between the results from the flow cytometry and the other scoring methods. The ranges seen with the flow cytometry-based scoring (3.50/1000 to 12.88/1000) and the other scoring methods (3.20/1000 to 13.83/1000) were comparable.

Representative examples of some of the QC graphs are shown in the Supplementary Results section. Examination of the QC plots indicates that, although there are some differences in the means and the variability, each laboratory has a distinctive pattern of results with a clear and relatively narrow range of negative control/baseline values.

Some of these graphs illustrate 'interesting' aspects of the data provided. The degree to which the intra-laboratory variability differed between laboratories can be seen from these charts such as those cases where there seemed to be 'shifts' or 'step-changes' in mean levels in the numbers of micronucleated cells (for example, Laboratories J and M). A more complete set of analyses is provided in the Supplementary Results. In several laboratories, some of the early replicates in the QC charts are outside the control limits which suggest increased variability in the early stages of the development of the assay in the laboratory which may be expected.

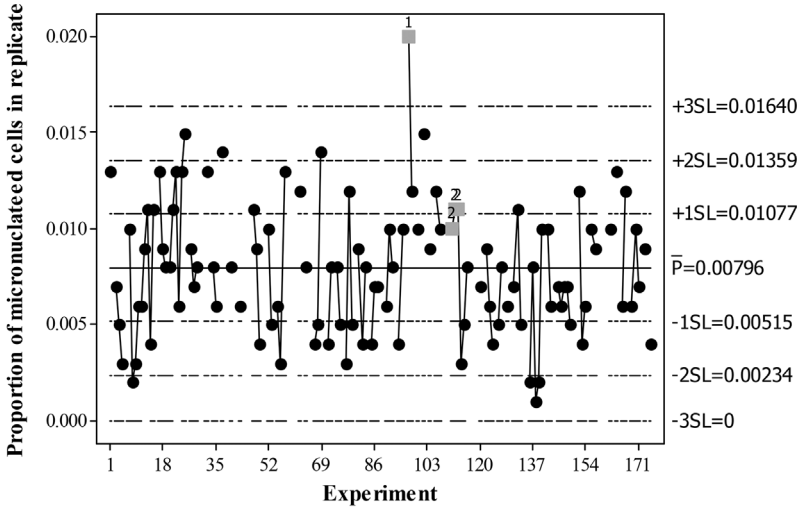


Fig. 4. Example of a p-chart for Laboratory L (L2).
Plot of proportion of counts of micronucleted cells from 175 experiments (+S9 short, most with single replicate). Lack of connecting lines between points in places relate to cases where particular combination was not included in experiment. Numerical values linked to grey points relate to points 'flagged up' by Western Electric rules (see text).

Tests performed with unequal sample sizes

Table 2
Means and SDs of micronucleated cells/1000 cells for 55 combinations for 13 laboratories (split between scoring methods).

Row	Lab	Comb	N	Mean	SD	Meth	Vehicle	Cyto-B	S9	Time
1	A	A1	14	7.43	3.86	M	DMSO	Y	+S9	Short
2	A	A2	10	8.30	4.47	M	Water	Y	−S9	Short
3	A	A3	2	6.50	0.71	M	Water	Y	+S9	Short
4	A	A4	22	10.59	3.63	M	Water	Y	−S9	Long
5	A	A5	24	8.58	4.65	M	Water	Y	−S9	Longer
6	A	A6	59	8.32	3.31	M	DMSO	N	−S9	Short
7	A	A7	72	8.15	3.91	M	DMSO	N	+S9	Short
8	A	A8	48	7.31	3.08	M	DMSO	N	−S9	Long
9	A	A9	8	7.25	2.25	M	Water	N	−S9	Short
10	A	A10	26	7.58	3.58	M	Water	N	−S9	Long
11	A	A11	22	7.96	3.34	M	Water	N	−S9	Longer
12	B	B1	24	7.65	2.08	M	DMSO	N	−S9	Short
13	B	B2	24	7.21	1.89	M	DMSO	N	+S9	Short
14	B	B3	24	7.88	2.51	M	DMSO	N	−S9	Long
19	E	E1	21	4.76	1.84	M	Water	N	−S9	Short
20	E	E2	19	3.90	1.29	M	Water	N	+S9	Short
21	E	E3	21	4.14	1.68	M	Water	N	−S9	Long
22	E	E4	27	4.70	1.66	M	DMSO	N	−S9	Short
23	E	E5	32	4.63	2.17	M	DMSO	N	+S9	Short
24	E	E6	39	3.62	1.93	M	DMSO	N	−S9	Long
25	E	E7	6	4.17	1.47	M	Medium	N	−S9	Short
26	E	E8	11	4.09	1.58	M	Medium	N	+S9	Short
27	E	E9	10	3.20	1.62	M	Medium	N	−S9	Long
28	F	F1	26	7.77	2.30	M	10%DHS-RPMI	Y	−S9	Short
29	F	F2	12	6.14	1.71	M	10%DHS-RPMI	Y	+S9	Short
30	F	F3	16	6.88	1.48	M	10%DHS-RPMI	Y	−S9	Long
31	F	F4	2	9.72	6.68	M	10%DHS-RPMI	Y	−S9	Longer
32	G	G1	20	7.30	2.56	M	Saline	Y	−S9	Short
33	G	G2	20	6.60	3.36	M	Saline	Y	+S9	Short
34	G	G3	14	5.29	3.60	M	Saline	Y	−S9	Long
35	G	G4	14	3.50	2.03	M	Water	Y	−S9	Long
36	G	G5	2	5.00	1.41	M	DMSO	Y	−S9	Short
37	G	G6	6	5.83	2.86	M	DMSO	Y	+S9	Short
38	G	G7	10	4.80	2.35	M	DMSO	Y	−S9	Long
39	H	H1	10	10.20	2.32	M	DMSO	N	−S9	Short
40	H	H2	117	9.04	2.75	M	DMSO	N	+S9	Short
41	H	H3	132	8.83	2.58	M	DMSO	N	−S9	Long
42	I	I1	623	9.26	2.16	M	DMSO	N	−S9	Short
43	I	I2	795	9.01	1.99	M	DMSO	N	+S9	Short
44	I	I3	765	8.69	1.92	M	DMSO	N	−S9	Long
47	K	K1	11	9.73	5.39	M	DMSO	N	−S9	Short
48	K	K2	28	9.86	7.03	M	DMSO	N	−S9	Long
49	K	K3	2	5.50	0.00	M	DMSO	N	−S9	Longer
50	K	K4	12	13.83	6.18	M	Water	N	−S9	Long
51	K	K5	4	5.75	3.10	M	Water	N	−S9	Longer
52	L	L1	117	8.23	3.15	M	DMSO	N	−S9	Short
53	L	L2	111	7.96	3.43	M	DMSO	N	+S9	Short
54	L	L3	118	8.80	3.28	M	DMSO	N	−S9	Long
15	C	C1	40	5.68	1.42	F	DMSO	N	−S9	Short
16	C	C2	44	7.46	2.60	F	PBS	N	−S9	Short
17	C	C3	42	6.16	1.99	F	DMSO	N	+S9	Short
18	D	D1	461	3.54	1.59	F	DMSO	Y	−S9	Long
45	J	J1	135	12.88	8.68	F	DMSO	N	+S9	Short
46	J	J2	251	7.88	4.82	F	DMSO	N	−S9	Long
55	M	M1	212	3.50	1.49	F	DMSO	N	−S9	Long

Comb: Combination; N: number of replicates; SD, Standard Deviation; Meth: scoring method F (Flow cytometry); M (other methods); S9; −S9 or +S9; Time; short, long, longer. (See text for more details of combinations).

4. Discussion

4.1. Study design, data collection and analysis

The role of historical control data is becoming more important in the analysis and interpretation of genetic toxicology tests as illustrated by the latest revision of the OECD Test Guidelines [2]. Historical control data can be useful provided the data chosen are comparable with the study being investigated (*i.e.* obtained under the same experimental conditions). They can also provide information on differences detected between negative control groups which have either received a vehicle or not (*i.e.* absolute negative controls) in those cases where a less commonly used vehicle is chosen. It should, though, be stressed that, in any discussion of historical control data, comparisons

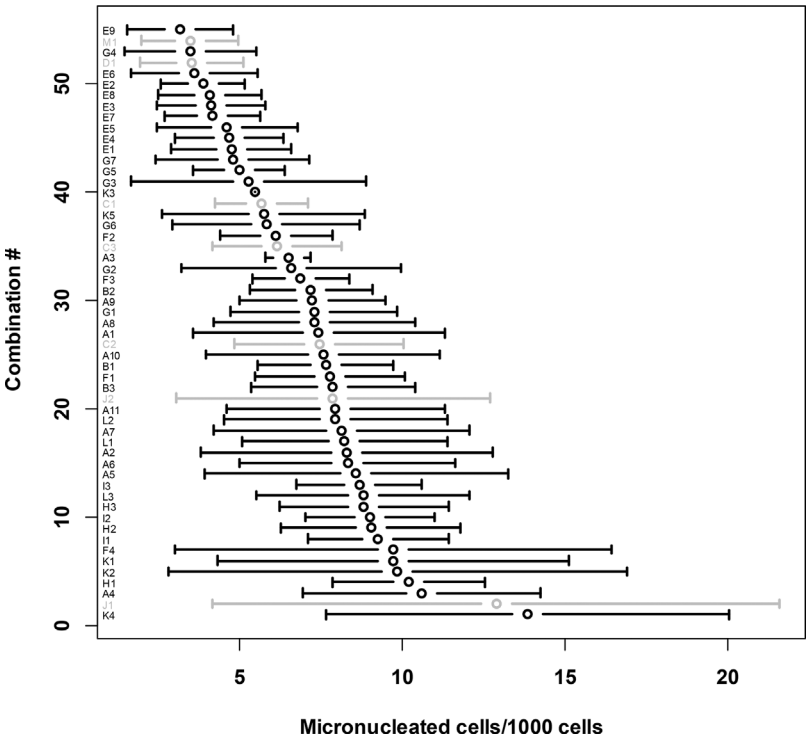


Fig. 5. Means and SD of 55 combinations: Grey: Flow; Black: Other Methods.

of test groups with the concurrent negative control group is the primary consideration in testing for genotoxicity testing.

The current GTTC workgroup analysis of negative control data from 13 laboratories using the TK6 cell line micronucleus assay provides important guidance values to improve the use of negative control data both for quality control and in the interpretation of test results. This assay and cell line was selected based on the likelihood that a sufficient amount of data could be obtained from a relatively small group of proficient laboratories using well-defined protocols from companies and institutes in the field. Some effort was put into securing their participation. The number of participating laboratories is considered adequate and is comparable with other inter-laboratory comparison studies. It was a requirement for participation that laboratories were able to provide data from at least 20 individual experiments. The results can, therefore, be considered a 'best case scenario' and a successful 'proof of principle' for further studies.

The selected laboratories included those using accepted methods of scoring including flow cytometry, light and fluorescence microscopy. [6]. Details of the characteristics of the TK6 cell line and recommendations for its maintenance can be found in another HESI GTTC

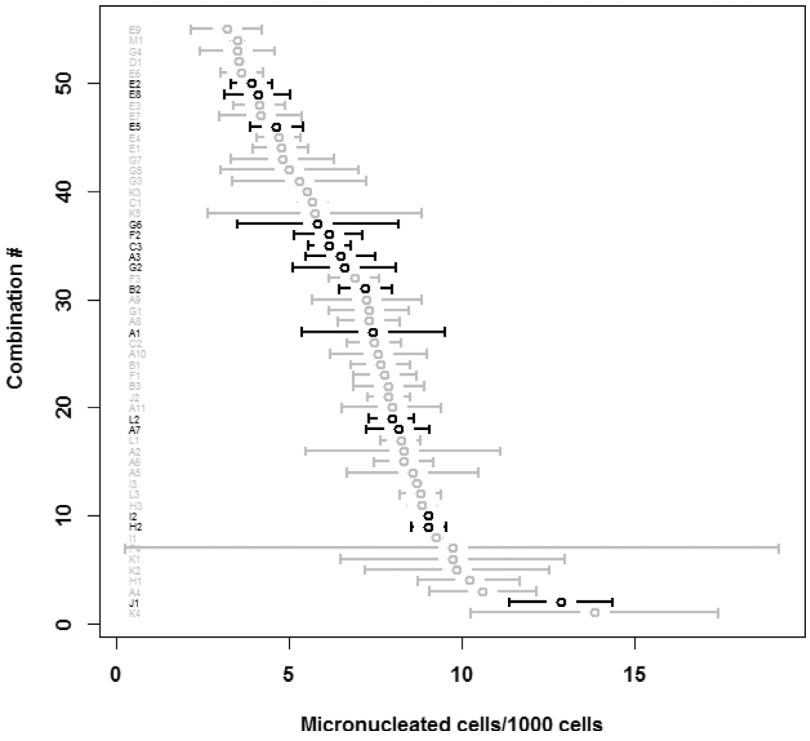
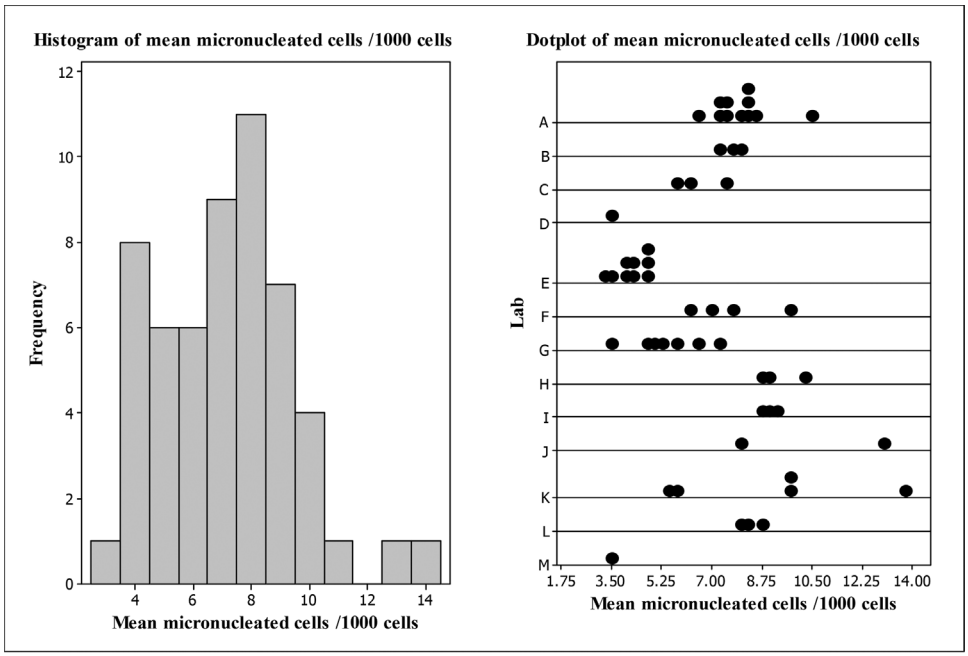


Fig. 6. Means and 95%CI of 55 combinations: Grey: −S9; Black: +S9.



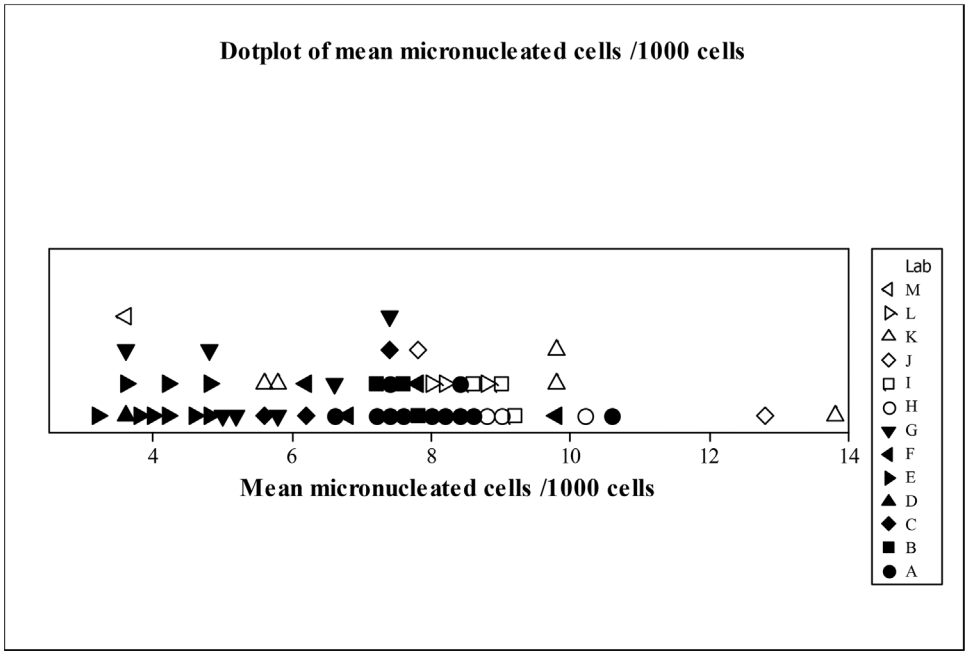


Fig. 7. Distribution of the mean micronucleated cells/1000 cells from the 55 combinations from the 13 participating laboratories.

paper [6]. Many laboratories (10) reported that they obtained their cell lines from ATCC (American Type Culture Collection). No information on the original donors was provided by any laboratory.

The data collected in this study were analysed by standard statistical methods and using QC tools which are widely used in industry to monitor processes and to ensure maintenance of quality control. This is now recommended by the OECD test guidelines. Although the length of data collection varied over time (range six months to 10 years) these data show that QC methods have a role in improving a laboratory's performance and that a further reduction of intra-laboratory variability is feasible. It is not known whether laboratories provided data from all their experiments or whether those experiments that were considered to have failed to meet the laboratory's own acceptance criteria were excluded from the datasets. Laboratories may vary in the degree to which they 'self-edited' their datasets. If this was the case, then such a laboratory might be expected to report less variability than actually occurred.

One of the complications in building up negative control databases is whether data can be combined across the different conditions to increase the number of replicates and achieve the 'rule of thumb' of 20 experiments for QC methods. In most cases here, there were no appreciable differences between the different sets of replicates within a laboratory. One difficulty is developing the database so that the QC charts can be based upon the order the experiments were conducted in (*i.e.* in a time-ordered format). In practice, it should be the responsibility of the laboratory to make the case for combining the results and doing this in a transparent way so that others can judge whether this is appropriate.

One of the complexities of cross laboratory comparisons is the standardization of the endpoints. Different laboratories use different denominators (*e.g.* 1000, 2000 or variable numbers of cells.) Results presented in the form of the number of micronucleated cells/cells scored expressed as proportions are also common; sometimes with different numbers of cells per replicate. In this case the precision of the

estimate of the mean for those conditions will vary from combination to combination. Estimates derived from flow cytometry are likely to be more precise than those from manual scoring due to the increased number of cells scored.

4.2. Between laboratory variability

The between laboratory variability is significantly greater than the within laboratory variability ($P < 0.001$ in an analysis of variance) (Figs. 8 and 9). The percentage variability in the replicate counts explained by the between laboratories component was nearly 70%. Less than 2% of the variability can be explained by including other factors in a multiple regression analysis. The inter-laboratory variability cannot be explained by the variables outlined in the methods section. There does not seem to be any large or systematic effects of experimental conditions such as choice of vehicle, short or long treatment and recovery times or the presence or absence of cytochalasin B (three laboratories reported using cytochalasin B with their TK6 cells) on the size of the means within and between the laboratories. Some inter-laboratory variability might be due to variables such as staining methods (seven laboratories used acridine orange and two Giemsa stains), types of visualization and scoring methods which differed between laboratories. Direct comparisons is difficult because of the relatively small numbers of laboratories using each method. However, there was no indication that these studies gave unusual or very different results. None of the variability that might be associated with these different experimental conditions exceeded that which could be due to other uncontrolled factors. This indicates that factors which differ between laboratories in the protocol have not been identified and contribute to the variability.

One extra source of variability between laboratories is the range of the lengths of the treatment and recovery time. Treatment times defined by the laboratories ranged from 3 to 4 h with recovery times varying from 21 to 40 h for both with and without S9 mix. The corresponding times for the −S9 long protocol were 24–30 h with no recovery time; the −S9 longer protocols had 24–48 h treatment times. These are the nominal times but variability may have been introduced if the actual times that the cells experienced in treatments in the laboratories differed somewhat from these.

Another explanation for the inter-laboratory variability could be the sources of the cell lines. Most of the laboratories reported that their cell lines originated from a reputable source, the ATCC. Lorge et al. describe the need for well-standardized cell lines [6] and noted that different results can arise, even when experienced laboratories test a specific chemical using batches of cells believed to have been derived from the same origin [5]. Other possible sources of variability (heterogeneity) between laboratories could include passage number, the original donors and pooling cells.

4.3. Within laboratory variability

In those cases where it was possible to assess intra-laboratory variability, this was low with little difference between the different study designs (−S9 short, +S9 short and −S9 long). In some cases (where this was testable), there was some statistically significant variability between experiments within a laboratory. In these cases, however, the mean differences were small and unlikely to be biologically important to the extent that they would impact the results of conducting the micronucleus test.

Heterogeneity within a laboratory can be introduced by different scorers over time. A number of laboratories noted, however, that all the scoring had been done by a single scorer, in some cases, blinded to the treatment groups, which would also blind the scorer to the negative controls.

In nearly all laboratories there was evidence of over-dispersion where the between replicates variability was greater than what would be expected just by sampling error. This indicates that there are variables other that those recorded in the study that can lead to the variability in the results. Over-dispersion is an issue in the derivation of control limits. The relationship between the mean and SD of counts provides some indication of the degree of over-dispersion. The laboratories varied with respect to the uniformity of their results. Laboratory L, for instance, reported results from 346 replicates but showed only the occasional values with extreme variability in the QC charts. Laboratory D, on the other hand, with 461 replicates showed an appreciable number of replicates over the UCL (see Supplementary Results, Figs. S59 and S24).

The significant variability between experiments within laboratories also suggests that it should be possible to achieve a further reduction in the degree of inter-replicate variability. To achieve this, a laboratory would need to check its procedures to see whether there were potential aspects of its protocol that might introduce variability between replicates and between experiments. Ryan [7] discusses, in general, approaches that can be used to do this.

One possible innovation for the counts obtained from the flow cytometry method would be a check that the proportion of micronuclei in each 1000 cells scored consecutively conforms to a Poisson distribution. This does not have to be done but it would provide a QC check on the

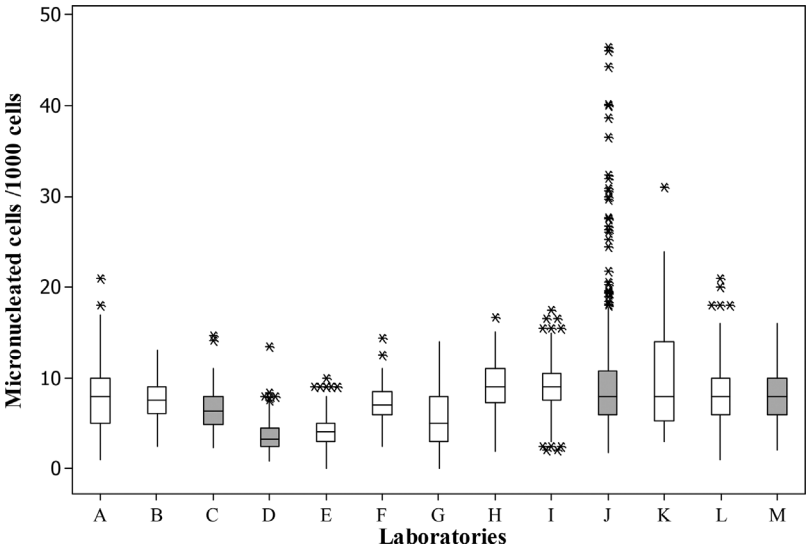


Fig. 8. Box plots of the distribution of the micronucleated cells/1000 cells from the 13 participating laboratories.
Shaded box plot indicate laboratories using flow cytometry, clear box plots indicate laboratories using other counting methods. Asterisks indicate potential outliers.

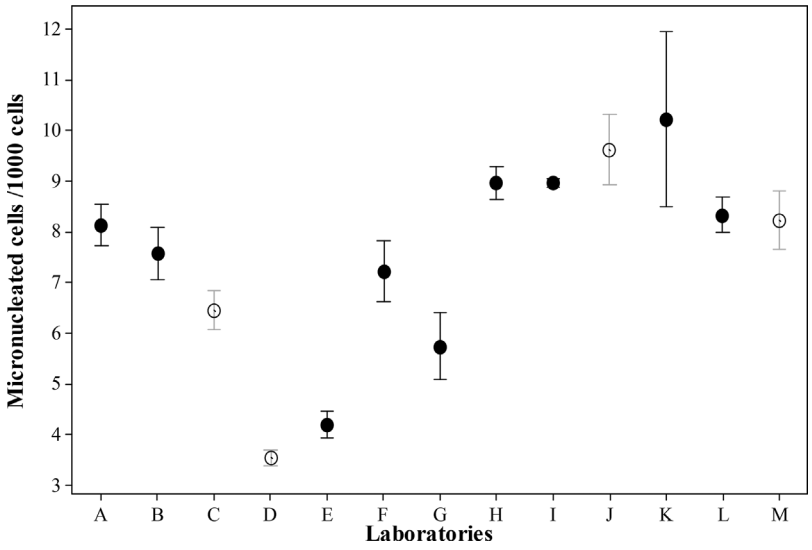


Fig. 9. Mean and 95% CI for the mean micronucleated cells/1000 cells from the 13 participating laboratories.
Open circle indicates laboratory used flow cytometry and filled circle indicates laboratory using other counting methods.

possibility that there was variability in the selection of cells over the longer run of sampling.

4.4. QC charts and acceptable ranges

Counts on the same number of cells are readily analysed using methods based upon the Poisson distribution. C-Charts have the advantage that the width of the control limits depends upon the theoretical relationship between the mean and the variance of the Poisson distribution and it is possible to derive Poisson-based control limits. These allow a degree of visual representation of how much over-dispersion there was within the laboratories.

The control limits, therefore, represent the 'pure error' associated with the endpoint and what can be achieved in the absence of inter-replicate or inter-experiment variability. This can, therefore, represent the 'normal', baseline, or irreducible variability associated with the endpoint. Counts that fall outside these limits either represent very rare random events or, more likely, the effect of some uncontrolled factor. Attempts to detect shifts in the mean within this 'normal range' would require large experiments to have sufficient power.

Based upon these results, it can be argued that an acceptable level of variability for the background control incidence of micronuclei would be one that fell within the control limits such as the ± 2SD range (*i.e.* the LWL and UWL) provided that the dataset does not show excessive variability. Obviously the mean level within a specific laboratory affects the size of effect detectable and the design of a study with the power to detect a doubling over the negative control level.

In the combined groups there were five cases (A, E, G, K and L) where the calculated LCL was 'below zero'. This means that there is no lower value that would be outside the limits. However, in all these cases the calculated 'negative' LCLs values were just below zero indicating that zero values, while not 'triggering a warning' were close to the LCL.

The range of estimates of micronucleated cells/1000 cells across all the 55 combinations in 13 laboratories was from 3.2 to 13.8/1000 cells. Only four of the combinations had means outside the range 3.5/1000 and 10.6/1000. This range might be considered an indicator of acceptable performance if accompanied by evidence that the level of intra-laboratory variability is satisfactory. A slightly larger range is obtained from the mean ± 2SD of the overall mean of the 55 combinations: 7.06 (2.37–11.76) or from the overall means of the 13 laboratories across combinations: 7.48 (3.45–11.51). These ranges, calculated as the mean ± 2SD, could, therefore, be considered as acceptable. Combinations J1 and K4 fall outside these ranges showing appreciable variability and K4 had a small numbers of replicates. By excluding these two combinations, an acceptable range from to 3.20 to 10.59 was achieved. Laboratories with means falling outside this range should review their technical procedures.

The range of values, although wide, seems an acceptable range and, therefore, advice could be given that laboratories carrying out the assay should be able to show both that their mean value lies within this range and show evidence of acceptable level of within-laboratory variability around their specific mean as demonstrated by QC methodology. Therefore, each laboratory should be able to set its own acceptable range which should be narrower than one based solely upon the spread of the laboratories' overall means in this study.

4.5. Other data in the literature

The values reported here are broadly in line with those reported by other investigators but on a much larger set of data. Lorge et al. [6] report a value of 10.2 ± 4.8 scored on 1000 cells (without cytochalasin B) based upon data from Honma and Hayashi [12] and note "… that these ranges reflect values from experienced laboratories handling the cells under rather standardized procedures. They should be taken seriously as recommendations for acceptable values, but not as strict specifications for exact values for other laboratories and have no regulatory status."

Zhang et al. [13] published negative control incidences of between 6 and 15 mnt/1000 cells (based upon approximately 13 experiments with a single replicate per experiment). Honma and Hayashi [12] reported background negative control data in TK6 cells following a 4 h treatment +48 h recovery protocol without S9 mix and presumably without cytochalasin B (as there was no mention of cytochalasin B in their paper). They reported incidences of a mean of 10.2/1000 cells (SD 4.8/1000 cells). This mean was presumably based upon the negative control data from the 25 participating laboratories which tested 14 chemicals in their study. From their Fig. 3, the range of the values is from just over 0/1000 to approximately 25/1000 with a median of about 10/1000 (the values in their Fig. 3 do not seem, however, to relate to the results of the 30 experiments shown in their Table III where the range appears to be from 4 to 16 micronucleated cells/1000 cells). The vehicle used was either physiological saline or DMSO. A single negative control culture appears to have been scored in each experiment with at least 1000 cells scored for micronuclei (the range of values reported, therefore, in effect, includes both a between laboratory and a within laboratory component).

In the OECD report [4], five laboratories (with between 39 and 198 replicates) described micronuclei incidences with a range of from 4.2/

1000 to 11.3/1000. Two of the laboratories indicated that they had used cytochalasin-B and reported micronuclei frequencies of 4.2/1000 and 7.1/1000 cells.

Background levels of micronucleated TK6 cells in the presence or absence of cytochalasin B can also be found in the literature. Fellows and O'Donovan [14] reported, in a non-cytokinesis blocked assay with S9 mix (*i.e.* no cytochalasin B block), micronuclei frequencies of 7.5/1000 cells and 8.3/1000 cells in negative control TK6 cultures after 24- and 48-h 'treatments' respectively. Elhajouji [15] described negative control incidences (in Tables 1–4 of the paper) of 7–20.5/1000 cells in human TK6 cells in the presence and absence of cytochalasin B. These incidences appear to be based upon a single replicate negative control in each experiment. Fowler et al. [16] indicated negative control incidences based upon quadruplicate cultures (in Tables 1–3 of their paper) of 4.5–9.5/1000 cells in human TK6 cells in the presence and absence of cytochalasin B.

In conclusion, this study has identified the range of results collected by 13 proficient laboratories in conducting the *in vitro* micronucleus assay using TK6 cells. In addition, it demonstrates, that the variability between laboratories does not appear to be due to a number of differences in how the data were produced. This study supports the need for data to be produced from about 20 experiments to create a negative control database, it also provided evidence to justify the OECD guideline recommendation that at least 2000 cells (from two cultures) per concentration should be scored in order to provide an acceptable range for historical control databases. Furthermore each laboratory should be able to set its own acceptable range which should be narrower than one based solely upon the spread of the laboratories' overall means in this study.

Acknowledgements

The authors wish to thank the participating laboratories and their staff who provided data: AstraZeneca, BSRC, Canon Inc, Covance, Janssen A, Janssen B, Litron, Lundbeck, NCTR, Novartis, Pfizer, Sanofi, and Takeda. 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.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.mrgentox.2017.10.006.

References

[1] M. Hayashi, K. Dearfield, P. Kasper, D. Lovell, H.J. Martus, V. Thybaud, Compilation and use of genetic toxicity historical control data, Mutat. Res. 723 (2011) 87–90.
[2] OECD, Guidance Document on Revisions to OECD Genetic Toxicology Test Guidelines, Genetic Toxicology Guidance Document: Second Commenting Round. Nov 30, 2015, https://www.oecd.org/env/ehs/testing/Draft%20Guidance%20Document%20on%20OECD%20Genetic%20Toxicology%20Test%20Guidelines.pdf.
[3] OECD, Guideline for the testing of chemicals, Test No. 487, In vitro mammalian cell micronucleus Test, 2016. http://www.oecd-ilibrary.org/environment/test-no-487-in-vitro-mammalian-cell-micronucleus-test_9789264264861-en.
[4] OECD, Environment, Health and Safety Publications, Series on Testing and Assessment, No. 198, Report on statistical issues related to OECD test guidelines (TGs) on genotoxicity, ENV/JM/MONO(2014)12, 2014. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2014)12&doclanguage=en.
[5] S. Pfuhler, M. Fellows, J. van Benthem, R. Corvi, R. Curren, K. Dearfield, P. Fowler, R. Frotschl, A. Elhajouji, L. Le Hegarat, T. Kasamatsu, H. Kojima, G. Ouedraogo, A. Scott, G. Speit, In vitro genotoxicity test approaches with better predictivity: summary of an IWGT workshop, Mutat. Res. 723 (2011) 101–107.
[6] E. Lorge, M.M. Moore, J. Clements, M. O'Donovan, M.D. Fellows, M. Honma, A. Kohara, S. Galloway, M.J. Armstrong, V. Thybaud, B. Gollapudi, M.J. Aardema, J.Y. Tanir, Standardized cell sources and recommendations for good cell culture practices in genotoxicity testing, Mutat. Res. 809 (2016) 1–15.
[7] T.P. Ryan, Statistical Methods for Quality Improvement, John Wiley & Sons, Inc., 2011.
[8] RCore Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
[9] G.R. Henderson, Introduction, Six Sigma Quality Improvement with Minitab, John Wiley & Sons, Ltd, 2011, pp. 1–11.
[10] D. Montgomery, Introduction to Statistical Quality Control, 6th edition, John Wiley & Sons, Inc, Hoboken, New Jersey, 2009.
[11] E. Mullins, Statistics for the Quality Control Chemistry Laboratory, The Royal Society of Chemistry, Cambridge, 2003 ISBN 0-85404-671-2, xvii + 455 pp.
[12] M. Honma, M. Hayashi, Comparison of in vitro micronucleus and gene mutation assay results for p53-competent versus p53-deficient human lymphoblastoid cells, Environ. Mol. Mutagen. 52 (2011) 373–384.
[13] L.S. Zhang, M. Honma, M. Hayashi, T. Suzuki, A. Matsuoka, T. Sofuni, A comparative study of TK6 human lymphoblastoid and L5178Y mouse lymphoma cell lines in the in vitro micronucleus test, Mutat. Res. 347 (1995) 105–115.
[14] M.D. Fellows, M.R. O'Donovan, Etoposide, cadmium chloride, benzo[a]pyrene, cyclophosphamide and colchicine tested in the in vitro mammalian cell micronucleus test (MNvit) in the presence and absence of cytokinesis block using L5178Y mouse lymphoma cells and 2-aminoanthracene tested in MNvit in the absence of cytokinesis block using TK6 cells at AstraZeneca UK, in support of OECD draft Test Guideline 487, Mutat. Res. 702 (2010) 163–170.
[15] A. Elhajouji, C. Mitomycin, 5-fluoruracil, colchicine and etoposide tested in the in vitro mammalian cell micronucleus test (MNvit) in the human lymphoblastoid cell line TK6 at Novartis in support of OECD draft Test Guideline 487, Mutat. Res. 702 (2010) 157–162.
[16] P. Fowler, J. Whitwell, L. Jeffrey, J. Young, K. Smith, D. Kirkland, Cadmium chloride, benzo[a]pyrene and cyclophosphamide tested in the in vitro mammalian cell micronucleus test (MNvit) in the human lymphoblastoid cell line TK6 at Covance laboratories, Harrogate UK in support of OECD draft Test Guideline 487, Mutat. Res. 702 (2010) 171–174.