Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

11-Jul-2014

English - Or. English

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

REPORT ON STATISTICAL ISSUES RELATED TO OECD TEST GUIDELINES (TGs) ON GENOTOXICITY

Series on Testing and Assessment

No. 198

JT03360419

ENV/JM/MONO(2014)12
Unclassified

English - Or. English

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

IOMC
INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS
A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT
Paris 2014

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 34 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in eleven different series: **Testing and Assessment; Good Laboratory Practice and Compliance Monitoring; Pesticides; Biocides; Risk Management; Harmonisation of Regulatory Oversight in Biotechnology; Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; and Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (http://www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organisations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

Also published in the Series on Testing and Assessment link

For this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/chemicalsafety/)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

Several projects for updating the set of OECD Test Guidelines (TG) on genotoxicity were submitted and included in the Test Guideline workplan in April 2011 by a group of four countries, Canada, France, the Netherlands and the US.

To support this update, statistical analyses were performed in 2012 and 2013, based on data collected from calls for data organised between September and December 2012. This document presents the Statistical analysis supporting the revision of the OECD Test Guidelines on Genotoxicity. It combines in **Section 1**, the power and sample size investigations carried out by Health Canada, for the *in vivo* rodent erythrocyte micronucleus assay (TG 474) and in **Section 2**, the statistical analysis performed by a consultant for the OECD Secretariat for the *in vitro* chromosomal aberration test (TG 473) and the *in vitro* micronucleus test (TG 487).

This statistical analysis report was endorsed by the Working Group of the National Coordinators of the Test Guidelines Programme (WNT) at its 26th meeting in April 2014. The Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology agreed to its declassification on 7th July, 2014.

This document is published under the responsibility of the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology.

TABLE OF CONTENTS

STATISTICAL ANALYSIS SUPPORTING THE REVISION OF THE OECD TEST GUIDELINES ON GENOTOXICITY

INTRODUCTION

1. This document combines the power and sample size investigations carried out by Health Canada (Section I), lead country for the update of TGs 474 and 475 and by David Lovell, St George's, University of London (Section II), as a consultant for the OECD Secretariat. The objective of this document was to help expert groups to identify the optimum experimental designs for *in vivo* and *in vitro* tests to be incorporated into the OECD Test Guidelines for *in vivo* and *in vitro* genetic toxicology tests.

2. The document is split into two sections. Firstly, the document produced by Health Canada discussing assay sensitivity, sample size and statistical power consideration for the rodent erythrocyte micronucleus assay. Secondly the document, produced by David Lovell, discussing statistical issues related to OECD *in vitro* genotoxicity Test Guidelines TG473 and 487.

3. The first section discusses issues arising from a meeting of the expert group on the review of OECD Test Guidelines on Genotoxicity, held in February 2012 [ENV/JM/TG/M(2012)5], where there was discussion on the number of cells that needed to be scored in the *in vivo* micronucleus test. The section reviews earlier work, carries out a simulation aimed at identifying the sample sizes in terms of animals and cells needed to detect fold increases with different spontaneous frequencies. It also provides information on the spontaneous frequencies from sets of data supplied by a number of laboratories. The emphasis on fold increases was based upon published recommendations that sufficient cells needed to be scored to detect a doubling of micronucleus frequency with 80% power.

4. The second section reviews and comments on work done in the context of the revision of *in vivo* genotoxicity Test Guidelines. A similar analysis is then carried out for the *in vitro* negative control data for the in vitro Test Guidelines TG 473 and 487. This work focuses on a number of cell types used in the *in vitro* cytogenetics and *in vitro* micronucleus assay and issues that were raised during expert group meetings with relation to the number of cells and replicates that need to be scored. Data were available from a series of experiments in a number of laboratories to provide data on which to base analyses. The section discusses the numbers of cells that need to be scored to get an approximate power of 80% for each assay to detect a 2 to 2.5 fold increase in the frequency of micronucleated cells/ frequency of cells with chromosomal aberrations/ mutant frequency.

5. The sections that make up this document have been used to support the revision of Test Guidelines 473, 474 and 487. Some of the discussion is relevant to the earlier draft versions of the TG and their evolution into the final version. Therefore, there may be some reference to text which no longer appears in the final version of the updated Test Guidelines.

6. The Annex at the end of the document presents the rationale for the recommendations made in the updated Test Guidelines. They are based on the statistical analysis included in this report and the discussion among the expert group on the review of existing Test Guidelines on Genotoxicity.

SECTION 1

Discussion of Assay Sensitivity, Sample Size and Statistical Power for the Rodent Erythrocyte Micronucleus Assay

Discussion Document on Assay Sensitivity, Sample Size and Statistical Power for the Rodent Erythrocyte Micronucleus Assay

Context

7. At the Expert Group Meeting for the Review of the Set of OECD Test Guidelines on Genotoxicity held in February 2012, there was some initial discussion on the number of cells that should be scored in the micronucleus assay and whether this should be increased from the current guideline. Since the last revision of Test Guideline 474 in 1997, advances in automated scoring methods, including image analysis systems and flow cytometry, have made counting large number of cells (up to 20,000 in many cases) relatively straightforward. Yet, these technologies are not found universally in all testing or research labs.

8. Micronucleated immature erythrocytes (MN-IEs) arise spontaneously at low frequencies, typically from 0.02 to 0.2% (1)(2); unpublished results), which has focused attention on the question of whether the current guideline recommendation to score 2000 IEs per animal will provide sufficient statistical power to detect treatment-induced increases in the MN-IE frequency.

9. Recommendations on the number of IEs to score have been made as an outcome of several past International Workshops on Genotoxicity Testing (IWGT). The 2nd IWGT expert meeting of March 25-26, 1999 recommended that, "Sufficient cells should be analyzed to assure the detection of a doubling in micronucleus frequency with 80% power at a significance level of less than 0.05" (3). Subsequently, the 4th IWGT meeting of September 8-9, 2005 further recommended that, "…it is desirable that each laboratory should determine the minimum cell sample size required to ensure that scoring error is maintained below the level of animal-to-animal variation" (4).

Recent Analyses

10. Several recent publications have focused attention on the problem of sample size and statistical power.

11. Kissling *et al* (5) considered inter-animal variability and the binomial counting error in relation to the observed spontaneous frequency of MN-IEs in several species. The inter-animal coefficients of variation from experimentally determined MN-IE values were reported to range from 26% (rat bone marrow) to 41% (rat peripheral blood). When considering the binomial error in the count of MN-IEs in an increasing series of IEs counted per animal from 2000 to 20,000 at spontaneous MN-IE frequencies of 0.05% to 0.30%, it was indicated that the counting error of individual animals values was substantially greater than the variation between animals.

12. The Kissing *et al* calculations of the number of IEs to be enumerated to reduce counting error to the equal to, or less, than the inter-animal coefficient of variation are reproduced below. Based on these calculations, about a doubling to tripling of the number of cells currently scored would be required for the binomial counting error to at least equal the inter-animal variation for spontaneous frequencies in the 0.1 to 0.2% range, as is common in rodent models.

Spontaneous frequency (% MN-IE)	Species/tissue with this approximate spontaneous frequency	Inter-animal %CV	No. of IEs to be scored to reduce counting error %CV to		
			Equal the inter-animal %CV	50% of the inter-animal %CV	20% of the inter-animal %CV
0.10	Rat PB	41	5943	23,772	148,573
0.20	Mouse BM & PB	35	4074	16,294	101,837
0.30	Dog	30	3693	14,770	92,315

13. Kissling *et al* also considered the minimum detectible increases in MN-IE in groups of 5 animals as a function of spontaneous frequency and the number of IEs scored based on Monte Carlo simulations. Their results are reproduced below.

Spontaneous frequency (%MN-IE)	No. of IEs scored	Minimum detectable fold-increase in spontaneous frequency			
		With 90% probability		With 95% probability	
		At $p \leq 0.05$	At $p \leq 0.01$	At $p \leq 0.05$	At $p \leq 0.01$
0.05 (S.D. = 0.020)	2,000	4.5	6.8	5.6	9.3
	4,000	3.5	5.5	4.0	6.3
	20,000	2.3	3.1	2.4	3.4
	∞	1.8	2.1	1.9	2.2
0.10 (S.D. = 0.045) (rat PB)	2,000	3.3	4.8	4.1	6.4
	4,000	2.9	4.2	3.2	4.7
	20,000	2.2	3.0	2.4	3.2
	∞	1.8	2.0	2.1	2.4
0.20 (S.D. = 0.070) (mouse PB)	2,000	2.7	3.9	3.0	4.5
	4,000	2.3	3.2	2.4	3.5
	20,000	1.9	2.5	2.2	2.7
	∞	1.7	2.0	1.8	2.1
0.20 (S.D. = 0.059) (rat BM)	2,000	2.7	3.9	2.9	4.4
	4,000	2.2	3.1	2.4	3.3
	20,000	1.8	2.3	1.9	2.5
	∞	1.6	1.8	1.7	1.9
0.30 (S.D. = 0.092) (dog PB)	2,000	2.4	3.4	2.6	3.7
	4,000	2.1	2.8	2.2	3.0
	20,000	1.8	2.3	1.9	2.4
	∞	1.6	1.8	1.7	1.9

14. The Kissling results were confirmed and extended by Health Canada. A simulation study was conducted in R to identify the minimal detectable fold change for varying number of cells scored (2,000, 4,000, 8,000, 20,000) such that a power of 0.8, 0.9 and 0.95 was achieved. Random samples of size five animals per group were generated assuming a binomial distribution. Simulations were conducted with spontaneous frequencies of 0.05%, 0.1%, 0.2% and 0.3% for controls and at significance levels of 0.05 and 0.01. Generalized Estimating Equations were used to test for differences between treated and controls. The bisection method was used to identify the minimal detectable fold change to obtain a given power.

Spontaneous Frequency (%MN-IE)	# of IEs scored	Minimum Detectable Fold-Increase					
		Power 0.80		Power 0.90		Power 0.95	
		$p < 0.05$	$p < 0.01$	$p < 0.05$	$p < 0.01$	$p < 0.05$	$p < 0.01$
0.05	2,000	3.42	4.30	4.10	5.16	4.64	5.82
	4,000	2.51	3.04	2.87	3.50	3.23	3.86
	8,000	2.00	2.29	2.24	2.59	2.42	2.80
	20,000	1.59	1.75	1.73	1.89	1.82	2.03
0.10	2,000	2.57	3.09	2.94	3.52	3.34	3.76
	4,000	2.03	2.31	2.26	2.55	2.46	2.84
	8,000	1.69	1.86	1.81	2.03	1.93	2.14
	20,000	1.40	1.50	1.48	1.60	1.55	1.66
0.20	2,000	2.01	2.32	2.25	2.62	2.39	2.77
	4,000	1.67	1.87	1.81	2.02	1.91	2.16
	8,000	1.46	1.58	1.54	1.68	1.62	1.79
	20,000	1.27	1.35	1.33	1.40	1.38	1.46
0.30	2,000	1.79	2.04	1.97	2.22	2.13	2.35
	4,000	1.54	1.70	1.66	1.79	1.74	1.90
	8,000	1.37	1.44	1.45	1.53	1.51	1.60
	20,000	1.24	1.27	1.27	1.33	1.31	1.36

15. Based on the IWGT recommendation that a sufficient number of cells be scored to detect a doubling in the MN-IE frequency with 80% power at a confidence level of $p < 0.05$, upwards of 8000 IEs (or more) may need to be scored to achieve this. Considering this analysis, it is evident that at spontaneous frequencies of 0.05% or lower, the assay is insensitive: if 2000 IEs are scored in a lab with spontaneous frequencies of 0.05%, then a 3.4-fold increase would fail to be detected in 20% of experiments with a confidence level of $p < 0.05$. Only when spontaneous frequencies approach 0.2% would scoring 2000 IEs meet the IWGT recommendation.

16. Sample size has also been considered by Hayes et al (6). Based on a spontaneous frequency of 0.048% from the Han Wistar rats used in their laboratory, the power to detect a doubling at a significance level of $p < 0.05$ for a group size of 5 animals was determined to be 47, 68, 80, 86 or 91% when scoring 2000, 4000, 6000, 8000 or 10,000 IEs, respectively. Hayes *et al* also proposed a sequential scoring strategy whereby treatment group results that are marginally statistically significant and/or outside the acceptance range for control values are confirmed by re-coding the slides and scoring an additional 4000 IEs, bringing the total to 6000 IEs.

Inter-laboratory comparison of negative control frequencies

17. A request was made via members of the expert group for negative control data. Data were sent by GSK, Huntingdon, NTP (ILS), DuPont, Litron Labs, Servier, Covance, Sanofi and others; however, most of these laboratories wished that their data be presented anonymously. Since the format in which the data were presented differed greatly between sources, the extent of the analysis that can be done at present with all datasets was limited to a tabulation of the mean frequencies for each species, strain and tissue. These data are presented in the following table.

Species/tissue	Source of data	Spontaneous frequency (% MN-IE)	Standard deviation (%MN-IE)
Muta™ mouse male (BALB/c x DBA2) peripheral blood	Health Canada	0.25	0.10
B6C3F1 mouse male peripheral blood	National Toxicology Program	0.26	0.05
B6C3F1 mouse female peripheral blood	National Toxicology Program	0.21	0.05
ICR-CD-1 mouse male + female bone marrow	Lab 1	0.11	0.031
ICR-CD-1 mouse bone marrow	Lab 2	0.055	0.05
ICR-CD-1 mouse male bone marrow	Lab 4	0.074	0.062
ICR-CD-1 mouse female bone marrow	Lab 4	0.08	0.065
ICR-CD-1 mouse male peripheral blood	Lab 4	0.16	0.04
ICR-CD-1 mouse female peripheral blood	Lab 4	0.13	0.05
Wistar rat male peripheral blood	National Toxicology Program	0.096	0.034
Wistar rat female peripheral blood	National Toxicology Program	0.088	0.044
Wistar rat male + female bone marrow	Lab 1	0.103	0.043
Wistar rat male + female bone marrow	Lab 1	0.086	0.03
Wistar rat bone marrow	Lab 2	0.065	0.03
Wistar rat male bone marrow	Lab 3	0.081	0.066
Fischer 344 male peripheral blood	National Toxicology Program	0.042	0.02
Fischer 344 female peripheral blood	National Toxicology Program	0.038	0.018

Additional Analyses

18. Restricting the analysis to only those labs providing individual animal data allowed further exploration of the inter-animal variability. For the purposes of determining inter-animal coefficients of variation, only those labs supplying frequencies determined by flow cytometric analysis of ~20,000 IEs were considered, since the high count numbers are expected to minimize the binomial counting error, leaving mostly inter-animal variability as the source of variation. Examining the number of IEs required for the counting error to equal the inter-animal coefficient of variation yielded the following results.

Source of data	Species/tissue	Spontaneous frequency (% MN-IE)	Standard deviation (%MN-IE)	Number of animals	Inter-animal %CV	No. of IEs to be scored to reduce counting error %CV to equal the inter-animal %CV
Health Canada	Muta™ mouse male (BALB/c x DBA2) peripheral blood	0.25	0.10	109	38	2755
National Toxicology Program	B6C3F1 mouse male peripheral blood	0.26	0.05	245	19	10,626
	B6C3F1 mouse female peripheral blood	0.21	0.05	131	24	8382
	Sprague Dawley rat male peripheral blood	0.055	0.021	100	38	12,465
	Sprague Dawley rat female peripheral blood	0.060	0.023	65	38	11,335
	Wistar rat male peripheral blood	0.096	0.034	65	35	8297
	Wistar rat female peripheral blood	0.088	0.044	50	50	4541
	Fischer 344 male peripheral blood	0.042	0.02	62	48	10,407
	Fischer 344 female peripheral blood	0.038	0.018	29	49	10,978
Litron Laboratories	Sprague Dawley rat male peripheral blood	0.11	0.033	156	30	10,167

References

(1) Krishna, G., Urda, G. and Paulissen, J. (2000). Historical vehicle and positive control micronucleus data in mice and rats. Mutat Res, 453, 45-50.

(2) Salamone, M.F. and Mavournin, K.H. (1994). Bone marrow micronucleus assay: a review of the mouse stocks used and their published mean spontaneous micronucleus frequencies. Environ Mol Mutagen, 23, 239-273.

(3) Hayashi, M., MacGregor, J.T., Gatehouse, D.G., Adler, I.D., Blakey, D.H., Dertinger, S.D., Krishna, G., Morita, T., Russo, A. and Sutou, S. (2000). In vivo rodent erythrocyte micronucleus assay. II. Some aspects of protocol design including repeated treatments, integration with toxicity testing, and automated scoring. Environ Mol Mutagen, 35, 234-252.

(4) Hayashi, M., MacGregor, J.T., Gatehouse, D.G., Blakey, D.H., Dertinger, S.D., Abramsson-Zetterberg, L., Krishna, G., Morita, T., Russo, A., Asano, N., Suzuki, H., Ohyama, W., Gibson, D. and In Vivo Micronucleus Assay Working Group, IWGT. (2007). In vivo erythrocyte micronucleus assay III. Validation and regulatory acceptance of automated scoring and the use of rat peripheral blood reticulocytes, with discussion of non-hematopoietic target cells and a single dose-level limit test. Mutat Res, 627, 10-30.

(5) Kissling, G.E., Dertinger, S.D., Hayashi, M. and MacGregor, J.T. (2007).
Sensitivity of the erythrocyte micronucleus assay: dependence on number of cells scored and inter-animal variability. Mutat Res, 634, 235-240.

(6) Hayes, J., Doherty, A.T., Adkins, D.J., Oldman, K. and O'Donovan, M.R. (2009).
The rat bone marrow micronucleus test--study design and statistical power. Mutagenesis, 24, 419-424.

SECTION 2

Report on Statistical Issues Related to OECD *in vitro* Genotoxicity Test Guidelines

Summary

19. This report explores a number of statistical issues related to OECD guidelines for genotoxicity. Firstly, a review of previous work investigating the optimal number of cells and animals needed for the *in vivo* micronucleus test (TG 474). Secondly, analyses related to statistical issues associated with TG 473 (*in vitro* chromosomal aberration test) and TG 487 (*in vitro* micronucleus test).

20. The results of two investigations by Kissling *et al* and Health Canada (HC) were reviewed and together with various further simulations it was concluded that the methods were in broad agreement, that studies with sample sizes of 5 and scoring 4000 cells per animal had approximately 80% power to detect about 2-3 fold increases when the negative control incidence was high but that the power was smaller for lower background levels.

21. A number of laboratories provided historical negative control data for the *in vitro* cytogenetics (21 laboratories) and micronuclei (15 laboratories) tests. The distributions of the aberrations and micronuclei were derived for these samples.

22. There was no clear distinction between the mean scores for the different cell types. The range of variability in the chromosomal aberrations was relatively narrow but wider for the *in vitro* micronucleus tests. While there was variability between laboratories, there was little evidence for differences in incidence depending upon the presence or absence of S9 mix or different treatment times.

23. Combining data from replicate negative control cultures would provide an increase in power. Scoring more cells from the concurrent negative control would also provide slightly more power and more consistent results. Using more concentration levels would provide a more accurate description of the concentration-response relationship. A trend test is more powerful than pair-wise comparisons especially if multiple comparison procedures are applied to the pair-wise comparisons.

24. Similar power calculations were done for the chromosomal aberration and micronucleus tests. These showed that the *in vitro* chromosomal aberration test had low power to detect a 2-fold increase. The power curves that were derived provide indications of the number of cells that would be needed to identify a fold increase or, alternatively, the power associated with various designs.

25. There was no evidence in the datasets provided that there was heterogeneity between replicate cultures but a number of laboratories showed appreciable between experiment variability.

26. In the absence of inter-replicate variability the *in vitro* micronucleus test has power comparable with the *in vivo* micronucleus test but in the presence of inter-replicate variability it may be vulnerable to artifactual results. The use of more replicate cultures may provide some protection against possible artifactual results if the number of cell counted per concentration levels were to be greatly increased.

27. Although it was concluded that the datasets were adequate for carrying out the investigations it was noted that some improvements could be made to the development of databases of negative control data.

28. In response to further questions, the numbers of cells needed to be scored in some of the tests to avoid zero counts was explored. Based upon a Poisson distribution, samples where sufficient cells were measured such that the mean culture counts were approximately 5 would result in only a few cultures having scores of zero and the data would be approximately normally distributed. This would provide data that would be convenient for statistical analysis but this is not a prerequisite to achieving an adequate statistical analysis.

29. Eight scenarios for the use of historical negative control data to help determine a positive result and in the assessment of the biological importance of the results were explored. Most methods have some limitations. The GEF (Global Equivalence Factor) approach used in the Mouse Lymphoma Assay (MLA) test may have some potential. Another approach that might warrant further investigation is to use C-charts, developed for quality control purposes. These together with control limits based upon the Poisson distribution could provide an indication of results that would not be considered unusual as well as those which would indicate an effect had occurred.

Introduction

30. The purpose of this work was to perform statistical analyses to support the determination of the optimal number of cells to be scored in Test Guideline (TG) 473 (*in vitro* chromosomal aberration test), TG 476 and TG 476 bis (*in vitro* gene mutation tests) and TG 487 (*in vitro* micronucleus test), after reviewing and taking into consideration approaches already developed for determining the statistical power of the Erythrocyte Micronucleus Assay (TG 474) and *in vivo* chromosomal aberration assay (TG 475).

31. This report reviews analyses previously carried out with respect to the *in vivo* micronucleus test (TG 474) and then discusses statistical issues associated with Test Guidelines (TGs 473 and 487).

Some general points related to *in vitro* study designs and their statistical analysis

Hierarchical / nested designs

32. Many genotoxicity designs such as the *in vivo* and *in vitro* cytogenetics or micronucleus assays are hierarchical or nested designs in that there are cultures or animals 'nested' within treatment groups and a number of cells from each animal or culture are scored. The different levels in the design have differing degrees of variation which ideally should be taken into account in any statistical analysis.

33. A key concept in the statistical analysis of a hierarchical design is the identification of the experimental unit. This is defined as the "smallest amount of experimental material that can be randomly assigned to a treatment". In principle, this should be the culture or the animal as it is impractical in the test to assign cells randomly from the same animal or culture to different treatment groups.

34. There may be differences between cultures within a concentration level and between cells within a culture. The cells within a culture (sometimes called the observational unit), however, are not independent and correlations between cells can lead to biases. In general, statistical analyses should take into account these different levels of variability otherwise 'hidden' levels of variability may distort estimates of variability and lead to errors in interpretation.

35. Statistical tests which appear more powerful may be carried out by analyzing the data at the cell level rather than the culture level. This apparent increase in sample size is an example of pseudo-replication where there are no true replications of the observations because repeated measurements of the same sample are wrongly treated as independent. This is a violation of an important assumption underlying statistical analyses such as the Chi-square and Fisher's exact tests of 2 x 2 tables which assumes that the individual data points to be analysed are independent of one another. In effect, measuring more and more cells from the same animal or culture can result in a more precise description of that unit and the apparent identification of increasingly significant differences between animals or cultures as the number of cells is increased, leading to an overestimation of the statistical significance of comparisons

36. Clearly, in many *in vitro* tests the experimental unit would be the culture and the test could, therefore, be severely underpowered. However, the genotoxicity 'community' is aware of these potential limitations as well as the degree of inter-culture variability in their tests. It uses its considerable experience and understanding of the test systems to take a pragmatic approach to the analysis and interpretation of the assay results. The power calculations described below are, therefore, carried out at the cell level.

Limitations of Power calculations

37. The power of a study is the probability of detecting a true effect of a specific size or larger using a particular statistical test at a specific probability level. The power is $(1-\beta)$ where β is the Type II error associated with a hypothesis test. (The Type II error is the probability of wrongly accepting the null hypothesis as true when it is actually false while the Type I error is the probability of rejecting the null hypothesis when it is actually true.)

38. Power and sample size calculations are an important part of the design of experiments and have an important role in identifying the feasibility of a design and identifying the resources needed to ensure a successful outcome.

39. These calculations depend upon the formulation of a null and alternative hypothesis and can be easy to carry out when there are just one or two treatment groups using closed form equations which are

now available in many software packages. More sophisticated designs require simulations involving the use of pseudo-random numbers to mimic experimental results.

40. It is, though, important to appreciate the limitations of these methods which are dependent upon the assumptions and inputs into the equations and simulations. They are based around the use of P-values for statistical inference and the identification of 'statistically significant' results, an approach which is under considerable criticism in the statistical literature and the increasing appreciation that statistical significance does not equate to biological importance.

41. Power calculations are also very dependent upon the estimate of the standard deviations included in the calculations. Estimates of the measure of variability have their own variability and the variance of an estimate of the variance/standard deviation can be large so that estimates of sample sizes are vulnerable to these assumptions**.**

42. Table 1 shows the 95% confidence intervals for the estimates of variability in the Kissling *et al*'s (2007) paper. This shows that when the sample sizes are smallish (i.e. less than 30) that the upper confidence limit can be twice as big as the lower one. The implication here is that this variability can affect the sample sizes needed to equate the binomial counting error to the inter-animal variability.

43. An experimental result is a sample of one from the population of all the possible experiments of the same design that could have been done. This single result is, therefore, somewhere on the hypothetical distribution of possible results and gives the 'best' estimate of the true state. In practice, it could be very close to the true or expected value, but it could also be near the upper extreme of the distribution of a low effect or the lower end of a high difference. It is not possible to distinguish from this single result which is correct.

44. One consequence of a design with 80% power is that it will also be able to detect effects smaller than the specified effect but with lower power. The design will then be capable of detecting real but small effects, which while being statistically significant are considered biologically unimportant (i.e. a negative 'call'). A study which is designed to have 80% power to detect a specified effect will also have, for instance, 50% power to detect a smaller effect. This particular effect is also the effect that will be just statistically significant at the chosen critical value say P=0.05.

1) Consideration of approaches developed for determining statistical power of the Erythrocyte Micronucleus Assay (TG 474)

45. The objective of the section is to review and comment on the analyses that have been done by Kissling *et al* (2007) and Health Canada (HC) (2012) in the context of the revision of *in vivo* genotoxicity Test Guidelines.

46. In the *in vivo* micronucleus assay (TG 474) some discussions has centred on recommendations to design studies with two specific characteristics. Firstly, to have sufficient power to be able to detect a fold increase (or doubling) of the endpoint. Secondly, to identify the number of cells that needs to be scored to reduce the binomial error to that of the inter-animal variability.

47. An objective of Kissling *et al's* (2007) work was to equate the "counting error" to the "inter-animal variability". They combined the binomial counting error expressed as a coefficient of variation (CV) with the inter-animal CV to obtain an expression giving n, the sample sizes of cells needed to equate the two measures. These two terms, respectively, are equated to the %CVs in their Table 2 and the % CVs in their Table 1.

Kissling *et al's* (Table 1) of mean and SD of various groups

	%	n	SD	CV%
Rat	0.11	15	0.05	41
Rat	0.23	190	0.06	26
Mouse	0.20	79	0.20	35
Dog	0.31	22	0.31	30

48. Kissling *et al* (2007) equate the "binomial counting error" to the standard error (also called the standard deviation) of the binomial distribution of a proportion assuming such a distribution.

49. Kissling *et al* (2007) also carried out a Monte Carlo simulation to investigate the properties of the design in terms of statistical power. One-tailed non-parametric Mann-Whitney tests were carried out to make comparisons between the 'control' and 'treated' groups.

50. They introduced inter-animal variability by simulating inter-animal differences from a distribution with a defined mean and SD based upon 4 combinations of representative values of percentages and SD (from their Table 1) and identified fold increases that could be identified. Individual micronuclei frequencies for each of the 10 animals were derived based upon the binomial distribution for 2000, 4000 and 20000 cells. 3000 runs were carried out with increasing levels of fold increase (f) until the 'power' exceeded either 90 or 95%. (Similar results were obtained using an alternative formulation based upon the beta distribution.)

51. A small set of simulations were also carried out at St George's using the R software with a similar set of assumptions and generally confirmed the results obtained in the simulation.

52. The HC discussion paper reproduces some of the tables included in Kissing *et al* (2007) paper and came to broadly similar conclusions that, say, 8,000 cells per animal are needed to have a reasonably power for detecting a 2-fold (or doubling) increase in micronuclei counts when the incidence level is low.

53. HC's simulation used the R software/language. The simulation used different numbers of cells (2000, 4000, 8000 & 20000) and negative control frequencies of 0.05, 0.1, 0.2 and 0.3%. Significance (alpha) levels of 0.05 and 0.01 were applied and the fold changes associated with power of 0.8, 0.9 and 0.95 were identified using the 'bisection method'. HC used Generalized Estimating Equations (GEE) to test the difference between the control and treated groups.

54. Running the R code kindly provided by HC gave similar results to those in the HC paper on the performance of the simulations.

55. Although broadly similar, there were some differences between the power associated with the different approaches (Kissling *et al* and HC). This is illustrated in Table 2 which shows the differences between the fold changes identified by the two methods at the two power levels (90 and 95%) that were common to the two studies. In general, the HC simulations showed more power (smaller sample sizes needed to detect a given fold change) than those in the Kissling *et al* paper.

56. These differences may arise from the different statistical tests used with the non-parametric Mann-Whitney tests having lower power but may also relate to how the inter-animal variability is included in the simulations. Kissling *et al* (2007) explicitly introduce an inter-individual component into their simulation. It is not clear whether the HC approach incorporates an inter-individual component in the simulation. Initial review of the R code suggests that there is no extra added variation.

57. Kissling *et al* (2007) also derived an equation for identifying the number of cells that need to be scored for the counting error to equal the within animal variability. HC appear to have used this equation to carry out a similar analysis for their historical control data but only 5 of the 10 estimates matched those obtained using the Kissling *et al* formula. The other estimates are similar but are sufficiently different to suggest that either some estimates had been rounded before carrying out the calculation or a modified equation has been used.

58. Power can be increased by increasing the number of cells and/or cultures/animals. The R code from HC was run with sample sizes of 4, 5 and 6 to get some indication of the power associated with differences in sample size (Table 3). This shows, as expected, that increasing the sample size to 6 raised the power of the design while decreasing it to 4 lowered the power. Tables such as this can provide help in deciding on the relative benefit of altering the number of animals in the group and the number of cells scored per animal. If the costs of including an extra animal or scoring extra cells are available then an analysis could be carried out to find an optimum design in terms of cost for a specific level of variability in the estimates of the means.

59. A further analysis was carried out (Table 4) using the nQuery Advisor package to identify the fold changes that would be detectable using a one-sided pair-wise comparison at a power of 0.80 with very large sample sizes of cells (equivalent to $n = \infty$ used by Kissling *et al* in their Table 3).

60. In general, the results of these earlier studies are robust and broadly confirm the conclusions drawn previously. They confirm that large numbers of cells are needed to have a high power (80% or higher) to detect a fold increase or a doubling when the negative control incidence is low (0.05% or 0.5/1000) and that the power increases as more cells are scored as well as when the background incidence is higher.

61. They are also in broad agreement with the estimates produced by Hayes *et al* (2009). Hayes *et al* estimated that, using their stock of rat, there would be 74% power to detect a doubling based upon a background incidence of about 0.1% (1/1000) using a design with 2000 cells scored in each of 7 rats in

each group. They estimated the power would be 97% when scoring 6000 cells from each of 7 rats. They concluded that "no meaningful increase in power is gained by scoring >6000" cells.

62. Galloway *et al* (2012) in a poster to the 2012 GTA meeting reported various simulations based upon *in vivo* micronucleus data collected in their laboratory. They concluded that with an incidence of 0.18% in rats that the power was good for detecting 2-fold increases, probably based upon the use of a trend test, when either 10 rats and 2000 cells were scored or 5 rats and 4000 cells/ animal were scored. There was also good power to detect 2.5-fold increases with 4 groups of 5 mice with an incidence of 0.14%. These results broadly agree with the findings here.

A note on relative or absolute differences

63. An important question for any power calculation aimed at identifying an appropriate experimental design and sample size is what size of effect should be considered biologically important.

64. Differences between group means for an endpoint can be either absolute or relative. An absolute difference might, for instance, be an increase in the incidence of chromosomal aberrations by 2% from, say, 2% to 4%. A relative difference might be a 2- or 3-fold change. In one case an increase from 1% to 3% would be the same as from 2% to 4% but as a fold change the first would be 3-fold, the second 2-fold. If proportional changes are relevant then fold changes may be appropriate but if specific changes from the underlying background control incidence (noise) are of interest then an absolute difference may be more relevant. This is a discussion about the biological relevance or importance of the changes rather than about statistical significance. It is important to stress that attaining statistical significance should not be the primary objective of a statistical analysis and that statistical significance does not equate with biological importance.

2) Historical Negative Control data

65. A call was made in September 2012 from the OECD Secretariat for negative historical control data with a further call made in December 2012. The call allowed laboratories to restrict their return to recent data from upto their last 20 experiments.

66. A number of laboratories responded with data from their *in vitro* chromosomal aberration and *in vitro* micronuclei negative control cultures. Although individual culture data were requested not all laboratories provided their data in this form. In practice, the reporting of this information varied from laboratory to laboratory.

67. The objective of the data collection process was to obtain a manageable set of data from over 12 and no more than 20 laboratories for each test with a geographical spread and representation of the users of the tests. Some laboratories provided their most recent data from, say, their last 20 experiments; others provided their full historic control database. Laboratories were asked to provide some limited information on the conditions that the tests were carried out on. The amount of extra information provided was somewhat limited. For instance, in the case of the micronucleus assay only two laboratories specified that bi-nucleated cells were scored in the documentation provided and only three laboratories specified whether or not cytochalasin B was used.

68. As some laboratories requested their data should remain anonymous the laboratories have been given codes to designate them.

69. In the case of both tests (micronuclei and chromosomal aberration) the basic data is whether a cell has chromosomal aberrations or a micronucleus. The number of cells 'scored' from the culture (the denominator) is also required. A number of issues described below for each test complicated the analysis of these data.

70. The datasets provided were broadly suitable for investigation of the degree of variability across the negative control samples. Consideration should, however, be given to whether the laboratories who volunteered their data are representative of all the users of the tests and whether the call for data may result in an editing of datasets to exclude data which was problematic. Anecdotally, assessors of regulatory submissions have pointed out that some of the data seen by them have values appreciably outside the values reported here. The appreciation that compiling data for submission might be onerous and therefore allowing flexibility in the amount and time-span of data that could be submitted may create some differences in the quantity of data supplied by different laboratories. The reporting of results was somewhat variable between laboratories with not all laboratories providing data from individual cultures. If a more extensive database was to be created some standardization of the endpoints and the reporting format should be considered as well as whether an expert panel should be convened to provide guidance on whether to include or exclude datasets. A more detailed questionnaire covering more of the detailed experimental conditions might also provide useful extra information.

2.1) The *in vitro* cytogenetics test

71. This section discusses experimental design and statistical analysis issues associated with the *in vitro* chromosomal aberration test guidelines (TG 473) such as numbers of cells to score, single or duplicate cultures, numbers of concentration levels and the power of the designs.

72. The *in vitro* chromosomal aberration test is a 'mature' test having been in use for at least 30 years; the first OECD guidelines having been produced in 1983. The basic design consists of negative control cultures (sometimes referred to as flasks) and a number of treated cultures. Separate experiments are conducted with and without an S9 fraction to mimic metabolism. (Negative controls can be both vehicle and solvent controls.) Laboratories score 100 or 200 (sometimes 50) cells for aberrations. Some laboratories set up replicate cultures at each concentration levels and score 100 cells per replicates. Others set up a single culture and score 200 cells per culture.

73. Statistical analysis is usually by Fisher's exact test (usually one-sided) between the concentration levels and the negative control. Various Chi-square tests may also be used to show heterogeneity between the treatment results or to test for linear trends using the Cochran-Armitage trend test. It has long been recognized that the *in vitro* chromosomal aberration test has low power, meaning that is difficult to detect quite large real effects when they are present.

Summary of Results provided

74. Data were obtained from 21 laboratories. Some laboratories provided data from more than one cell type.

75. The variables in the data set included: different cell types - Human peripheral blood lymphocytes (HPBL), CHU/IL, and Chinese Hamster Ovary (CHO) WBL cells - and different combinations of the presence and absence of S9 fractions and experimental exposure times: - 3hr, 24hr, 48hr as well as different recovery times and multiple vehicles (which were not always defined). Experimental conditions, where stated, differed between laboratories. Some laboratories, for instance, included extra cultures with different concentrations of S9 fraction in their design. Data from these laboratories were separated where possible into 65 combinations made up of different S9 fractions and exposure time combinations.

76. Table 5 summarizes the results obtained from the laboratories.

77. All the results collated here are for chromosomal aberrations excluding gaps. The presentation of results varied between laboratories. Some results were from individual cultures, other were the results expressed based on two cultures. Data were provided as counts of the numbers of aberration in the culture, as aberrant cells as a % of cells scored or as aberrations /100 cells.

78. Lab CA and CB counts were based upon 200 cells combined over replicate cultures by multiplying % aberrations by 2.

79. For Labs CC, CD, CG, CH CI, CJ, CK, CO, CQ, CR and CS counts were based upon 100 cells from individual cultures. For Labs CL and CP counts were based upon 100 cells from individual cultures (including a number with just 50 cells, mainly 0s).

80. Labs CE and CF counts were based up 100 cells from individual cultures n= 100-102 and 100-106 with a few cases where just over 100 cells were scored (i.e. 102 etc). The reported counts were included as counts/100 based upon a 'standard' 100 to ease statistical analysis and because the level of inaccuracy was small.

81. Lab CU count was based upon 2 x % assumed combined from replicate cultures. Lab CM and CN counts were based up 200 cells combined over replicate cultures.

82. Lab CT counts were summary data of the number of colonies with 0, 1, 2 etc in 100 cells.

83. Five other laboratories provided data. One laboratory (CX) contributed summary data from 2155 cultures (over a third of the total) with a mean of 1.41%. Another laboratory (CW) provided means and standard deviations from 238 cultures with different S9 conditions and times. One (CV) provided data expressed as %Cabs. The number of cells scored for the presence of structural aberrations per culture and 100 cells per culture was based upon between 2-6 cultures. Data included many non-integer numbers. Data from CW were not included in the analyses but are included in the footnotes to Table 5. Two other laboratories (CY and CZ) provided minimal information and were not included in the analyses.

84. No laboratory gave information on whether any negative control cultures had been 'rejected' because the chromosomal aberration levels were too high or too variable or for other Quality Control (QC) issues. Exact protocols were not provided. At least one laboratory only used single cultures of 200 cells.

85. In all, data were provided from 6287 cultures (mainly from samples of 50, 100, 200 cells scored per culture). The mean of the mean aberration % for the 65 combinations was 0.73% with an SD of 0.50%. The maximum mean was 2.5% and the minimum mean was 0.11% aberrant cells. The lowest value for an individual culture was 0%, of which there were many examples and the highest value was 6%.

86. Table 5 provides a summary of the descriptive statistics of data from laboratories for the *in vitro* cytogenetics tests. This lists the mean counts combined over different conditions (presence or absence of S9 and times), maximum and minimum culture counts and the number of cultures with zero aberrations based upon either the 100 or 200 cells scored. The mean % chromosome aberrations (aberrations/ 100 cells) are also provided. The table also provides the P-values associated with a chi-square goodness of fit test of the aberration counts to a Poisson distribution. A significant deviation ($P<0.05$) from goodness of fit is an indication of significant between experiment variability in culture counts for the laboratory.

87. As would be expected for data with small mean counts there were a large number of cultures with zero values. Variability within and between cultures and between labs could be assessed to some extent. While there was appreciable between laboratory variability there was no evidence for appreciable variability between replicate cultures within the same study. No heterogeneity or over-dispersion was detected between the replicate cultures in this dataset. In no case did the difference between the replicate cultures reach statistical significance in a two-sided Fisher's exact test.

88. These findings agree both with Soper & Galloway's (1994) results and with Margolin *et al*'s (1986) earlier findings that the 'true' Type 1 error rate of the Fisher's exact test is appreciably lower than the nominal 0.05 level when the incidence is low. No evidence of appreciable differences in aberration counts were apparent between replicates from male and female donors, vehicles or between the presence of absence of S9 fraction or different exposure times (3 or 24hr).

89. There was no marked evidence of differences in the incidence of aberrations between the different cells lines while inter-laboratory variability was pronounced. In some of the datasets where an analysis was possible there was evidence of heterogeneity between experiments with the variability between experiments being more than might be expected if all the data were from the same Poisson distribution.

90. Figure 1 shows a histogram of the distribution of the means of the 65 combinations and Figure 2 the means and with the associated 95% confidence intervals with the different cell types shown by

different colours. Figure 3 show the same means but this time with the associated standard deviations. In this case the different S9 fraction characteristics are shown. In a few cases it was either difficult to separate the datasets into sub-categories or the numbers were small so the combined values are included. Figure 2 shows there is appreciable difference in incidences between laboratories but no clear differences are obvious between the different cell types.

91. Figures 4-8 show the distribution of the counts for each of the different cell lines. The footnotes indicate how the counts were derived and whether they represent counts based upon 100 or 200 cells. The figures illustrate the appreciable inter-experimental variability associated with some laboratories and the high number of zero counts associated with those laboratories that have a low incidence of aberrations. The range of mean values (% aberrations) combined over all conditions for each of the cell lines was 0.20 to 1.07 for HPBL cells, 0.53 to 0.81 for CHO cells and 0.09 to 1.06 (with an outlying laboratory with a mean of 2.16%) for CHL/IU cells. The laboratory providing the largest sample, but of just summary data, reported 1.73% for HPBL and 1.41% for CHO cells.

Limitations of the Fisher's exact test

92. As noted above the actual alpha level of the Fisher's exact test is less than the nominal value of 0.05 when the incidence of chromosomal aberrations is small. This is illustrated by a simulation using R where 100 cells are simulated for two replicate cultures with the same proportion of aberrations (i.e. no difference between the two groups). The expectation is that the Type 1 error (false positive rate) should be 5%. (A two-sided Fisher's exact test is appropriate for these comparisons.)

93. In fact, the number of significant results based upon the $P<0.05$ criterion is:

% Aberrations	No. sig. out of 100,000 'pairs'	Type I error rate
1%	126	0.0013
2%	769	0.0077
3%	1403	0.0140
5%	2079	0.0208

94. This means that when there are low proportions of aberrations in the cultures there will be few significant differences detected between replicate negative control cultures (i.e. false positives). This is seen both with the data set available here and in Soper & Galloway's (1994) analysis of their data where few replicates cultures are statistically significantly different from one another.

95. A significant result with a two-sided Fisher's exact test at $P <0.05$ with sample sizes of 100 cells would have been 0 v. 6, 1 v. 8 and 2 v.10 for 100 cells per culture. (With 200 cells per culture the comparable results would also have been 0 v. 6, 1 v. 8 and 2 v.10). There were no examples of these or more extreme results in the data provided. (It is not known, though, whether experiments with significant differences between replicate cultures were excluded from the database.)

Power of the designs

96. Alternative designs have slightly different statistical properties such as the power to detect effects of a specific size as significant.

i) Power for detecting doubling: n=100 and n =200 with negative control frequencies of 0.5%, 1.0% and 2.0%

97. The three graphs (Figures 9-11) show the power associated with a pair-wise comparisons between a control and treated group using a one-sided Fisher's exact test at P=0.05.

98. The graphs illustrate the increased power obtained with an increase in sample size from 50 to 500 cells per treatment group and that the power to detect a doubling is increased with an increased background/control levels. A similar pattern of results could be obtained by running a simulation study.

ii) Duplicate vehicle control cultures

99. The draft test guidance states that "Because of the importance of the negative controls, duplicate vehicle control cultures should be used." In the absence of appreciable inter-culture variability 200 cells from one culture are approximately equivalent to 100 cells from two cultures.

100. There have been suggestions for scoring 400 cells (2 x 200 cells or 4 x 100 cells) for the negative control. This has the advantage of providing more accurate estimates of the negative control value for comparing treatment groups and increases the power of the test somewhat. Some laboratories have achieved 400 cells by pooling the negative and solvent control cultures.

101. It could be argued that replicate cultures provide some protection against the effect of 'rogue cultures' and give some protection against false significant results. Soper & Galloway (1994) point out, though, that the "Use of replicate flasks has a *theoretical* advantage for controlling over-dispersion compared with experiments using a single flask per treatment. When real data are examined, however, there is little practical improvement."

102. The table below (from nQuery Advisor) shows low power for sample sizes of 200 cell at the concentration levels to detect a doubling with background incidences as a proportion (p1) from 0.005 (0.5%) to 0.03 (3%). The table illustrates the low power of the design with only about 33% power to detect a doubling from 3% to 6% in the incidence of aberrations.

	1	**2**	**3**	**4**	**5**	**6**
Test significance level, a1 or 2 sided test?	**0.050**	**0.050**	**0.050**	**0.050**	**0.050**	**0.050**
Group 1	**1**	**1**	**1**	**1**	**1**	**1**
proportion, p1	**0.005**	**0.010**	**0.015**	**0.020**	**0.025**	**0.030**
Group 2	**2**	**2**	**2**	**2**	**2**	**2**
proportion, p2	**0.010**	**0.020**	**0.030**	**0.040**	**0.050**	**0.060**
Power (%)	**2**	**8**	**14**	**21**	**27**	**33**
n1	**200**	**200**	**200**	**200**	**200**	**200**
n2	**200**	**200**	**200**	**200**	**200**	**200**

103. The table below shows the power if the control group is made up of 400 cells from, for instance, two replicate cultures. The power is slightly higher if the negative control group has twice as many cells (e.g. 400 cells) as the concentration levels (e.g. 200 cells).

	1	2	3	4	5	6
Test significance level, a1 or 2 sided test?	**0.050**	**0.050**	**0.050**	**0.050**	**0.050**	**0.050**
Group 1	**1**	**1**	**1**	**1**	**1**	**1**
proportion, p1	**0.005**	**0.010**	**0.015**	**0.020**	**0.025**	**0.030**
Group 2	**2**	**2**	**2**	**2**	**2**	**2**
proportion, p2	**0.010**	**0.020**	**0.030**	**0.040**	**0.050**	**0.060**
Power (%)	**9**	**18**	**24**	**31**	**38**	**44**
n1	**400**	**400**	**400**	**400**	**400**	**400**
n2	**200**	**200**	**200**	**200**	**200**	**200**

104. Duplicate negative controls (200 v 400) cells gives slightly more power and a better estimate of the negative control incidence. This provides more robust comparisons when multiple concentrations are tested but increases the complexity of the study.

iii) Number of concentration levels: 3 or 6 concentrations

105. Different types of designs are possible such as a control and 3 concentration levels (200 cells per concentration level) or more concentrations such as a control and 6 concentrations (100 cells per concentration level). The more experimental/concentration groups the more precise will be the description of the concentration-response relationship especially when there are multiple concentration levels in the region of interest.

106. Margolin *et al* (1986), however, recommended the former design. "For *in vitro* chromosome aberrations, however, three concentrations and a control with 100 cells/concentration point appears to produce too insensitive an assay; an increase to 200 cells/concentration point in the Galloway *et al* protocol seems worthy of serious consideration." (Margolin *et al*, 1986).

107. More concentration levels should be used when there is specific interest in the concentration-response.

iv) Trend tests v. pair-wise tests

108. Multiple concentration levels provide the opportunity for different type of comparisons between the concentration levels: a test for a linear trend or pair-wise comparisons.

109. Tests of a linear trend in a concentration-response in a design are statistically more powerful than pair-wise comparisons because of the natural or inherent ordering imposed on it by the experimenter but need a more specific null hypothesis. The greater power of tests of these specific hypotheses can result in a shallow but real, concentration- response relationship being detected by the linear trend test although none of the pair-wise tests with the negative control are significant. On the other hand, if the concentration response is curvi-linear rather than linear the Cochran-Armitage trend test may fail to detect it, although an extension of the Cochran-Armitage test to test for curvature of the response could.

110. It should be recognised, though, that it is possible to obtain reproducible and statistically significant concentration-related responses but that these concentration responses have a shallow slope. The response might be considered a weak positive but biologically unimportant. Defining the criteria for

dichotomizing a result as either positive or negative requires expert judgement. It should be appreciated, though, that some misclassification will occur and that this is a consequence in part of the statistical issues around the 'power' of a design.

111. Complexity may arise over the use of multiple comparison methods. Multiple comparison approaches address concerns that there is a dramatically increased risk of Type 1 errors (declaring results significant when they are not) compared with the nominal level when a large number of comparisons (e.g. between pairs of treatments) are made. A multiple comparison procedure in effect, 'dampens' down the number of significant results reported.

112. For example, one commonly used method, the Bonferroni correction, could reduce the power of the design appreciably. A Bonferroni adjustment would, in effect, be testing for significance at P=0.0167 for a 3 concentration design and at P=0.0083 for a 6 concentration design. Other less conservative approaches that may be used include Dunnett's test for comparisons with a control, the Ryan-Einot-Gabriel-Welsh (REGWR) or the Student-Newman-Keuls (SNK) range tests.

113. The use of multiple comparison methods remains a controversial topic with considerable debate amongst statisticians over their use. A criticism of the multiple comparison approaches is that they ignore the structure of a carefully designed experiment where the concentrations and groups sizes have been chosen to have a high probability of identifying an effect of a certain size which is biologically important or which explicitly includes a concentration-response component.

114. The Cochran-Arbitrage test for linear trend is 'powerful' while Fisher's exact test pair-wise tests with a multiple comparison correction are 'conservative'. It is important that the specific test used to produce a P-value and declare a result as statistically significant or not is explicitly stated.

v) Power associated with the Cochran Armitage trend test.

115. The power associated with the linear trend test for various experimental designs can be derived from a method and equation developed by Nam (1987).

For example for the scenario:

Conc.on 0, 1, 2, 3, 4, 5, 6
% abs 1, 2, 3, 4, 5, 6, 7
with 200 cells in the negative control (nc) and 100 cells at 6 concentration levels would have 80% power to detect a linear trend (1% increase/unit concentration) at P<0.05.

The power would be 55% if the control incidence is 2% (and 1% increase/unit concentration) and 35% if the control incidence is 3% (and 1% increase/unit concentration)

A design with 200 nc cells and 200 cells at 3 concentration levels
Conc 0, 2, 4, 6
% abs 1, 3, 5, 7
Would have 95% power to detect a linear trend (P<0.05)

A design with 100 nc cells and 100 cells at 6 concentration levels
Conc 0, 1, 2, 3, 4, 5, 6
% abs 1, 2, 3, 4, 5, 6, 7
Would have 85% power to detect a linear trend (P<0.05)

A design with 200 nc cells and 200 cells at 3 concentration levels
Conc 0, 2, 4, 6
% abs 2, 4, 6, 8
Would have 90% power to detect a linear trend (P<0.05)

A design with 100 nc cells and 100 cells at 6 concentration levels
Conc 0, 1, 2, 3, 4, 5, 6
% abs 2, 3, 4, 5, 6, 7, 8
Would have 78% power to detect a linear trend (P<0.05)

116. Trend tests with a smaller number of concentration levels have slightly higher power based upon the Nam equation approach than those with more levels. Power is quite high for detecting linear trends but will be reduced and the interpretation more complex if the response is not linear or there are quadratic or cubic components to the dose-response relationship. .

2.2) The *in vitro* micronucleus test

117. This section discusses experimental design and statistical analysis issues associated with the *in vitro* micronucleus test guidelines (TG 473) such as numbers of cells to score, single or duplicate culture, numbers of does levels and the power of the designs.

118. The basic design is similar to the *in vitro* chromosomal aberrations test in that there are negative control cultures and a number of treated cultures. Separate experiments are conducted with and without an S9 fraction to mimic metabolism. Some laboratories set up replicate cultures at each concentration levels and scored 1000 cells per replicates. Others set up a single culture and scored 2000 cells per culture. The potential exists through the use of flow cytometry to score appreciably more cells per culture. Statistical analysis is usually by Fisher's exact test and/or various Chi-square tests.

Summary of results provided

119. Data were obtained from 15 laboratories. Data from three laboratories (MN, MO and MP) were not used in the analyses. Some laboratories provided data from more than one cell type.

120. The variables in the data set included: different cell type - Human peripheral blood lymphocytes (HPBL), CHL/IU, L5178Y and TK6 cells - and different combinations of presence and absence of S9 mix and experimental times: - 3hr, 24hr, 48hr and different exposure times and multiple vehicles (which were not always defined). Some laboratories provided data from more than one cell type. Experimental conditions, where stated, varied from laboratory to laboratory. Data for these laboratories were separated into 53 combinations based upon different cell types, presence or absence of S9 fractions and length of exposures.

121. Table 6 summarizes the results obtained from the laboratories

122. Only two laboratories (MA and ME) specified that bi-nucleated cells were scored in the documentation provided. Only a few laboratories specified whether or not cytochalasin B was used. Laboratories ME and MF stated that it was; Laboratory MB stated that it was not.

123. There was variable presentation of results. Data were supplied as counts of the numbers of micronuclei in the culture, others as cell counts reported as % micronucleated cells and micronucleated cells /1000 cells. Some results were from individual cultures, others were the results expressed based on two or more cultures. Replicate culture data were available from some laboratories which allowed an exploration of the degree of between and within laboratory variability.

124. Some laboratories (MA, MK) provided the counts of the number of cells with micronuclei from individual cultures clearly stating how many cells (usually 1000 or 2000 cells) were scored per culture. In these laboratories it was easy to identify the individual culture scores.

125. A number of other laboratories (MB, MC, ME, MI, MJ) reported their data as % micronuclei. These data points could be clearly linked to individual cultures where n=1000: in these cases integer counts could be obtained by multiplying by 10.

126. Counts were used for both the TK6 and L5178Y samples from MD based upon 1000 cells (with 2000 cells from three experiment where, for simplicity, numbers were halved to provide counts equivalent to those scored based upon 1000 cells). The data were presented as counts and as %MN. These two measures did not quite agree but were close to those expected if the number of cells scored were not exactly 1000.

127. MF presented the data as %MN based on scoring 1000 cells in a culture but multiplying by 10 did not always produced integer counts.

128. MG & MH presented the data as %MN based on scoring 2000 cells in a culture but multiplying by 20 did not always produced integer counts.

129. ML1 & ML2 provided data as %MN but no details were provided of how many cells these was based upon (or how many replicate cultures). Data were not easy to translate into integers and the number of cells scored per culture could not be estimated.

130. MM2 provided individual culture %MN for 2-6 cultures and were based upon 2000 cells/culture but the results reported were not integers (because counts could not be created). The MM1 provided data from duplicate samples with the %MN apparently based upon 2000 cells/culture but again values were not integers (and counts could not be created).

131. MK provided HPBL culture data as %MNs based upon both 1000 (expt 3) and 500 (expt 1& 2) cells/cultures. Converting the data back to counts for the 500 cell cultures gave non-integer counts. The data are included in the result table as if there were all based upon 1000 cells (which gave integer counts) (For example, 0.7% is 3.5 micronucleated cells if 500 cells are scored.)

132. One laboratory (MP) provided %MN based upon sample n's of between 1245-3383 cells. These data have not been included in these analyses. These data looked odd as the counts do not seem to be integers. For example, data were presented as in the table below:

No. of culture per experiment: 1

Incidence of micronucleated cell in the negative control group (%)	No. of cells per culture
0.457	2000
0.411	2296

133. One laboratory (MN) provided the mean and SD for an unknown number of CHL/IU cultures. These data were not included in the analysis.

134. Another laboratory (MO) provided summary data of the mean and SD from 2 cultures (500 binucleated cells / culture) from a small number of studies. These data were not included in the analysis.

135. No laboratory gave information on whether any negative control cultures had been 'rejected' because the micronuclei values were too high or too variable or for other Quality Control issues. Exact protocols were not provided.

136. In all, data were provided for 3316 cultures (mainly from samples of 1000 or 2000 cells). The mean of the mean micronuclei counts for the 53 combinations was 0.52% with an SD 0.28%. The maximum mean was 1.49% and the minimum mean was 0.14%. The lowest value for an individual culture was 0%, of which there were many examples and the highest value was 2.8% (28/1000).

137. Table 6 provides a summary of the descriptive statistics of data from laboratories for the micronucleus tests. This lists the mean counts combined over different conditions (presence or absence of S9 and times), maximum and minimum culture counts and the number of cultures with zero aberrations based upon either the 1000 or 2000 cells scored. The mean % micronuclei (micronuclei/ 100 cells) are also provided. The table also provides the P-value associated with a chi-square goodness of fit test of the

aberration counts to a Poisson distribution. A significant deviation ($P<0.05$) from goodness of fit is an indication of significant between experiment variability in culture counts for the laboratory.

138. Figure 12 shows a histogram of the distribution of the means of the 53 combinations and Figure 13 the means with the 95% confidence intervals with the different cell types shown by different colours. Figure 14 shows the same means but in this case with the standard deviations. In this figure the different S9 fraction conditions are shown. In a few cases it was either difficult to separate the datasets into sub-categories or the numbers were small so the combined values are included.

139. Figures 15-17 show the distribution of the counts for each of the different cell lines. The footnotes indicate how the counts were derived and whether they represent counts based upon 1000 or 2000 cells. The figures illustrate the appreciable inter-experimental variability associated with some laboratories and the high number of zero counts associated with those laboratories that have a low incidence of micronuclei. The range of mean values (% micronuclei) combined over all conditions for each of the cell lines was 0.45 to 1.39 for HPBL cells, 0.21 to 0.59 for L5178Y cells and 0.42 to 1.13 for TK6 cells.

140. There was appreciable between laboratory variability as can be seen from Figure 13 but no clear differences are obvious in incidence of micronuclei between the different cell lines or conditions. Inter-laboratory variability was much more pronounced than the effect of the other variables. There was little evidence for appreciable variability between replicate cultures within the same study.

Power of the test

i) Power for doubling n=1000 and n =2000 and negative control frequency 0.5% 1.0% 2%

141. The *in vitro* micronucleus test is carried out with either single or duplicate cultures at each concentration level. In this sense it is similar to the *in vitro* cytogenetics study. In theory, the culture would be the experimental unit and analysing the results as if the cell (the observational unit) were the experimental unit could lead to problems of pseudo-replication. In practice, there appears to be little evidence of appreciable between culture variability within studies in the negative control cultures. A pragmatic approach has been to analyse the data using Fisher's exact tests but to appreciate that this may lead to an increase in the false positive rate as a consequence of inter-culture variability.

142. Power calculations have been carried out for the *in vitro* micronucleus test based upon various negative control incidences (0.05%, 0.1%, 0.2% and 0.5%) and different sample sizes (1000, 2000, 4000, 5000, & 10000) cells from a single culture. Table 7 shows the fold change that would be detected for different control incidences and sample sizes. Table 8 shows the power associated with the design for detecting a fold increase or doubling over the negative control incidence.

143. Figures 18-21 shows the corresponding power curves associated with pair-wise comparisons between a control and treated group using a one-sided Fisher's exact test at P=0.05 with various negative control incidences (0.05%, 0.1%, 0.2%,.0.5%) and different sample sizes (1000, 2000, 4000, 5000, & 10000)

144. These graphs illustrate that the power associated with the *in vitro* micronucleus with a control incidence of 0.1% (1.0 per thousand cells) and sampling 1000 cells is very similar to that for the *in vitro* cytogenetic sassy with a negative control incidence of 1% and sampling 100 cells. Similar results are obtained with other incidences and cell numbers. The power curves of the two tests are nearly (but not quite) super- imposable. As before the power to detect a doubling or fold increase over the control incidence is increased with higher background/control incidences.

145. Tables 7 and 8 can be adapted to provide approximate fold-changes and power for different control incidences and numbers of cells for *in vitro* cytogenetics studies by multiplying the background incidence percentage by 10 and dividing the sample sizes n by 10.) The values in these tables can also be read directly off Figures 9 - 12 and 18 -21.

146. The incidence of micronuclei in negative control cultures is likely to be distributed as a Poisson distribution. With a low incidence (0.1%) and a relatively small number of cells (1000 cells /culture) a sizable proportion of zero counts would be expected. In this example it would be expected that about 37% of colony counts would be zeros and a further 37% would be expected to be 1. When the background incidence is higher (5/1000) and more cells are scored the distribution becomes more like a normal distribution (see later).

147. The analysis of the *in vivo* micronucleus test differs from the *in vitro* micronucleus test because, in the former, the animal is the experimental unit while the culture would take this role in the *in vitro* design. If the number of cultures within concentrations were increased then the simulations carried out by Kissling *et al* and HC could be adapted to assess the power associated with various levels of inter-culture variability. Note that it was not possible to run the R program of HC successfully with sample sizes of 2 (or 3 replicates*)*. A small simulation was carried out using the same general approach that Kissling *et al* had used for n=2 and n=3 and, again, difficulties were experienced possibly because the methods were unable to handle cases where there was no within group variability.

ii) Single or duplicate vehicle control cultures

148. Similar arguments can be made for the choice of single or duplicate control cultures as in the case of the *in vitro* cytogenetics test. Duplicate cultures could be useful while the test is being developed so that a larger database is created to provide a more accurate estimate of the inter-culture variability to include in power calculations or simulations.

iii) Number of concentration levels: 3 or 6 concentrations

149. Similar arguments to those used with the *in vitro* cytogenetics apply to the choice of either 3 or 6 concentrations levels. Between culture variability could increase the potential for significant, but artifactual, non-concentration-related responses to arise.

150. Using the Nam (1987) method the power associated with the Cochran-Armitage trend test with various scenarios is as follow (nc = negative control):

A design with 2000 nc cells and 2000 cells at 3 concentration levels
Conc 0, 2, 4, 6
% 0.1, 0.3, 0.5, 0.7
Would have 94% power to detect a linear trend (P<0.05)

A design with 1000 nc cells and 1000 cells at 6 concentration levels
Conc 0, 1, 2, 3, 4, 5, 6
% 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7
Would have 84% power to detect a linear trend (P<0.05)

A design with 2000 nc cells and 2000 cells at 3 concentration levels
Conc 0, 2, 4, 6
% 0.2, 0.4, 0.6, 0.8
Would have 89% power to detect a linear trend (P<0.05)

A design with 1000 nc cells and 1000 cells at 6 concentration levels
Conc 0, 1, 2, 3, 4, 5, 6
% 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8
Would have 77% power to detect a linear trend (P<0.05)

Again the power is very similar, but not quite identical, to that for the *in vitro* chromosomal aberrations test when the incidence is 10 times higher but the number of cells counted is 10 times lower.

151. In the absence of inter-culture variability the *in vitro* micronucleus test has power comparable with the *in vivo* micronuclei test but in the presence of inter-culture variability it may be vulnerable to artifactual results. The use of more replicate cultures may provide some protection against possible artifactual result if the number of cell counted per concentration levels were to be greatly increased.

3) Some other points within and arising from the Guidelines
(Note: some of the points in this section relate to specific text in an earlier draft of the guidelines which is no longer in the current draft)

i) Some comments on paragraph 38 in the section on the evaluation and interpretation of results.

38. Providing that all acceptability criteria are fulfilled, the following criteria are considered for the evaluation of results:

(1) the increase is dose-related,

Q. Definition of 'dose-related' needed? Monotonic?

(2) at least one of the test concentrations exhibits a statistically significant increase compared to the concurrent negative control,

Q .At least one or more? Definition of significant result? P<0.05 in Fisher's exact test without multiple correction? Depends upon sample size and number of concentrations.

(3) the statistically significant result is reproducible (e.g. between duplicates or between experiments),

Q. Reproducibility and replication

The statement "…but results between duplicate cultures can be useful for evaluation of reproducibility" is potentially misleading. It is important to distinguish between reproducibility and replication. Replication is not reproducibility but rather provides evidence of internal validity while reproducibility is a repeat study which provides evidence of external validity.

Soper & Galloway (1994) argue that the advantages of replication in practice may be less pronounced than might be expected. "Replicate flasks provide some reassurance that an experiment will not be completely lost if an individual flask is lost due to microbial contamination or processing error, for example. From a "quality control" point of view, replicates do not increase confidence in the integrity of the data because many technical problems affect the whole assay. Far better reassurance is obtained by replicating the experiment on a different day."

(4) the statistically significant result is outside the distribution of the historical negative control data (e.g. 95% confidence interval) (see paragraph 37).

This is discussed later.

ii) Reduced scoring when high numbers of aberrations are observed

152. Paragraph 30} states: "At least 200 well-spread metaphases should be scored per concentration and control equally divided amongst the duplicates, if applicable. In case of single culture per concentration (see paragraph 20 and 27), at least 200 well spread metaphases should be scored in this single culture. *This number can be reduced when high numbers of aberrations are observed.* "

153. This is not an approach to be recommended as it potentially leads to biased estimates.

iii) Positive Control data

154. Positive control data should not be used as part of the formal statistical analysis of the study. They should though be compared with the laboratory's historical positive control data base as part of the assessment of the adequacy of the study.

iv) Replicates are from different cultures not the same culture

155. There is a contradiction in the *in vitro* micronucleus guideline is that in para 44 there is a reference to two replicates of the same culture but later (para 48, 49) the references are to duplicate cultures (two cultures per concentration).

156. The advantage of replicates is that they come from different cultures and provide some internal validity. Cultures should be replicates not two samples from the same culture.

v) A comment on Poisson distributed data and number of cells to count to avoid zero counts

157. Data follow a Poisson distribution when an event is rare (the probability of the event, p, is small) and the events are independent of one another but the sample size, n is large. In the case of the Poisson distribution where the sample size, n, is large only the number of events (or counts) needs to be considered.

158. The data produced in both the *in vitro* chromosomal aberration and the *in vitro* micronucleus tests have an underlying Poisson component. Table 9 gives the values for the 90%, 95% and 99% confidence intervals assuming a Poisson distribution for different values of an observed count. In the case of a count of 1 in a 1000 (micronuclei) or 1 in a 100 (chromosomal aberrations) the 95% confidence interval of the population mean that the individual came from would be between 0.03 and 5.57. The comparable figures for an observation of a count of 5 would be 1.62 and 11.67. Note that versions of these tables and figures can be found in various statistical texts and web sites.
http://www.ucl.ac.uk/english-usage/staff/sean/resources/binomialpoisson.pdf
http://corplingstats.wordpress.com/

159. A small mean and a large n will produce many zero counts. If the mean count is one then, for Poisson distributed data, 37% of the counts will be zero and a further 37% will be zeros. If the mean is 0.5 than more than 60% of counts will be zero and 30% will be one. Table 10 lists the proportions of counts for different means. Figure 22 shows the percentage of zero scores for Poisson distributed data with increasing population mean. As the mean increases the percentage of zero counts falls. Figure 23 and 24 illustrates the Poisson distributions of mean counts from 1 to 10 and 1 to 25.When the population mean is 5 the expected proportion of zero counts is 0.7%. Sample sizes that will produce mean counts of 5 will also provide data which approximates to a normal distribution.

160. The Binomial and Poisson distributions have the property that they approximate to the normal distribution when np >5 (for example, when n=5000 and p =0.001). To avoid zero counts, enough cells (n) should be scored from cultures with mean proportions of cells with aberrations or micronuclei (p) such that np ≥5. To obtain np = 5 from cultures with a mean of 1% chromosomal aberrations, 500 cells need to be scored; for cultures with 0.1% micronucleated cells 5000 cells would need to be scored (Table 11).

161. The micronuclei counts for individual samples in the micronucleus test are a good example of how Poisson distributed data approximates to a binomial distribution and subsequently approximates to a normal distribution as the number of cells scored increases. Simulations can shows this very clearly. This property is one of the advantages of flow cytometry methods as large counts become approximately normally distributed and tractable for conventional statistical analyses.

162. A property of Poisson distributed data is that the mean is equal to the variance. This relationship between the mean and variance can be 'handled' by a square root transformation of the data (usually by sqrt(X+0.5) to avoid problems with zeros).

4) Scenarios for the use of historical negative control data in helping the interpretation of results from in vitro micronucleus and in vitro cytogenetic test

163. Criterion (4) for the evaluation of and interpretation of results (see paragraph 38).was that "the statistically significant result is outside the distribution of the historical negative control data (e.g. 95% confidence interval)"

164. In any discussion about historical control data, it should be stressed that the concurrent control group is always the most important consideration in the testing for genotoxicity. The historical control data can, though, be useful provided that the data chosen are from studies that are comparable with the study being investigated. They can also provide useful information on differences between vehicle and no vehicle control data.

165. The historical control data have two main uses. Firstly, to check whether the concurrent control data are consistent with the historical control data as a check on whether the experiment can be considered acceptable and, secondly, for use in considering the biological relevance or importance of any increases found in the treated groups.

166. The historical control database can be used as a reference to ensure that the concurrent control is consistent with previous studies.

167. The distribution of the historical negative control data consists of two components. Firstly, the range of historical control frequencies across the set of laboratories carrying out the tests. Secondly, the distribution of data generated within individual laboratories.

S1) Use of the range

168. Values within a specified range might be considered sufficiently 'small' to be considered to be compatible with negative control data. The range might be a generic one associated with the test as a whole or empirically determined based upon data derived previously by the laboratory carrying out the test.

169. Guidance could be given on what is considered an acceptable range of results based upon the range that a set of experienced laboratories have achieved in the past and reported in the published literature. In the data set here the range of laboratory means extend from 0.11% to 2.5% for the *in vitro* chromosomal aberrations test and from 0.14% to 1.49% for the *in vitro* micronucleus test. It was suggested, though, that *in vitro* chromosomal aberration tests submitted to regulators may have a range between 2% and 5%.

170. The disadvantage is that the range will get wider over time the more data are generated especially if there is appreciable between laboratory and between experiment variability. This makes it more difficult to identify treatment-related effects by them being outside the range. A similar comment has been made with respect to tumour data in long-term rodent carcinogenicity bioassays by Haseman *et al* (1984).

171. An implication of defining values within a range as 'normal' could be that few studies would be classified as positive. A wide range of negative control incidences would also mean that the interpretation of a fold change would be dependent upon the background incidence in the laboratories where, for example, the responses might range from an increase from 1% to 2% upto from 5% to 10%. The power of the design to detect such fold-changes would also differ between laboratories based upon their negative control incidence.

172. In the case of data from this set of laboratory the largest number of cells with micronuclei out of 1000 cells scored in the micronucleus test was 28 and the range for this laboratory was from 1 to 28 (Lab MF) . The smallest range was for laboratory MJ which was from 0 to 5 micronuclei in samples of 1000 cells. This compares with range suggested in TG 487 which states "Solvent/vehicle control and untreated cultures should give reproducibly low and consistent micronuclei frequencies (typically 5-25 micronuclei/1000 cells for the cell types identified in paragraph 11)".

173. In the case of the chromosomal aberrations the largest number of cells with chromosomal aberrations out of 100 cells scored in the cytogenetics test was 6 and the range for this laboratory was from 0 to 6 (Lab CC). The smallest range was from Laboratories CS and CQ where the range was from 0 to 1.

174. In general, "the use of the range of the historical data (i.e., minimum and maximum value observed during the data accumulation period) is not considered appropriate" Hayashi *et a*l (2011).

S2) A defined value

175. A value could be defined above which a result could be considered biologically important and below not. The choice of such a value would be somewhat arbitrary but would be based upon considerable experience with data sets from the test. An example here for the chromosomal aberration might be to choose a value of 5 cells with aberrations in a culture of 100 cells as being biologically important given that very few cultures in this data set exceed 5 aberrations / 100 cells for any of the laboratories.

176. The data produced in both the *in vitro* chromosomal aberration and the *in vitro* micronucleus tests have an underlying Poisson component. Based upon Poisson distributed data only a small proportion of cultures would be expected to have counts of 5 or more by chance for 100 cells with background incidences of 0.5 to 2.5/100 cells. (See Figure 25). Note that an increase from 1 to 7 is just statistically significant in a one-sided Fisher's exact test. Even though not statistically significant a value of 5 in a culture when the background incidence is 1 could represent a 5-fold increase as a consequence of the treatment. This is, again, an indication of the low statistical power of the test. If a laboratory had a higher background incidence a higher value might need to be derived.

177. The maximum values between the laboratories for the micronucleus tests are much more variable, from 5 to 28, so that identifying a value common for all of the laboratories represented here does not seem feasible.

178. It is possible that data that falls below the defined value might show statistically significant results. For instance, some consideration should be paid to a result such as 0, 1, 2 and 3 aberrations (or 1, 2, 3 and 4) in a study with 3 increasing concentrations but below an upper limit of the normal range of , say, 5). A set of increasing effects with increasing concentration levels would only be expected to arise about 1 time in 24 times (or P=0.042). (Note that using a Cochran-Armitage trend test for n=100 such a result would have a P-value of 0.15 and if based upon n=200 a P-value of 0.043.)

S3) Choice of 'cut-offs' based upon expert judgement

179. This approach involves the identification of values which classify results into two or a small number of categories. The choice of cut-offs is a set of criteria based upon expert but nevertheless subjective judgement. An example is the cut-off values identified within Soper & Galloway (1994).

These are that:

An ABR (aberration rate) >7.5% is 'reasonably convincing'
An ABR <3.0% is 'clearly negative'
An ABR between 3.0 to 7.5% is 'marginal' and may require repeat testing.

180. Elsewhere, Hayashi *et al* (2011) noted that a level of 5% for chromosomal aberrations has been used.

181. Such cut-offs derive from experience of the test system and an appreciable amount of experimental data. The use of these cut-offs in the interpretation of data from other laboratories may not be suitable unless there has been careful consideration of the data from that specific laboratory.

S4) A fold change.

182. A fold change might be considered a change between a control and treated group sufficient to warrant further investigation. The choice of what level of fold to consider important is again a decision that needs to be taken based upon experience of the tests.

183. The importance of the effect may depend upon the background incidence. Would it, for instance, be appropriate to consider the doublings of 1 to 2%, 2 to 4% and, say, 5 to 10% to be biologically similar; or for a 3-fold increase, would 1 to 3%, 2 to 6% and, say, 5 to 15% to be considered biologically equivalent? As shown elsewhere the power of a design to detect a fold increase depends, in part, on the negative control incidence.

S5) An absolute change

184. An alternative to a fold increase would be an absolute increase. The question here is , for instance, whether a 2 unit increase is biologically equivalent when it is from a negative control level of 1 to 2/100 cells as from 5 to 6/100 cells. Similarly, a change from 1 to 6/100 in comparison with 5 to 10/100?

185. Identification of either a fold-change or an absolute change as being biologically important needs to be based upon an understanding of the test system rather than just an arithmetical convention.

S6) Global Equivalence Factor (GEF) approach

186. The GEF approach used for the Mouse Lymphoma Assay (MLA) test is, in effect, an absolute change. Based upon Moore *et al* (2003) paper, the GEF is defined as the mean plus one standard deviation based upon the distribution of the historical negative control data collected across laboratories. Provided the concurrent negative control falls within a predefined range, again based upon the historic negative control data, then an induced mutation frequency (IMF) value obtained from a treated group which equals or exceeds the GEF triggers a statistical analysis and a significant trend test signals a positive result.

187. It is not clear that there is sufficient data at present to derive similar results for the two tests here. As an illustration (Figure 26), using the same approach as carried out for the MLA test, for the chromosomal aberration test for the HPBL and CHL/IU cell lines, based upon the combined means and standard deviations of chromosomal aberration counts across laboratories, the GEF equivalent would be approximately 1.3/100 cells for the CHL/IU and 1.62/100 cells for the HPBL.

188. In the case of the micronucleus the inter-laboratory variability is quite large. For the three cell types L5178Y, HPBL and TK6 the GEFs would be 7.53/1000, 8.09/1000 and 12.06/1000.

189. These analyses are carried out on the counts/1000 cells. The distributions are quite skewed and it may be necessary to transform the data to get appropriate estimates of the means and standard deviations for the calculation of GEFs.

S7) Laboratory control limit

190. The negative control data from a single laboratory should have an important role in providing evidence of the technical capability of the laboratory. Quality Control (QC) methods can be applied to investigate variability in the negative control data. A range of methods such as the development of control charts with action levels provide criteria for checking on the performance of a method. Experienced laboratories working to OECD guidelines could be expected to provide such information to show that the test is performing appropriately.

191. Quality Control (QC) methods provide graphs of the overall mean of all the cultures from a laboratory and upper (UCL) and lower control limits (LCL) based upon the between culture standard deviations derived from the data from the laboratory. In QC charts various rules are used to see whether the results from a laboratory are 'in control' and that the variability over time is effectively random and not the consequence of systematic changes in the conditions a laboratory is working under.

192. A 'breach' of the rules is supposed to generate action to bring the process back into 'control'. A 'simple' rule is that a value that falls outside the 3SD Upper Control Limit indicates the need to take immediate action because such a result is expected to occur less that 0.3% times by chance. Results that fall between the 2SD and 3SD limits can be considered as warnings that a system may be in the process of going out of control. Such event have about a 5% chance of occurring by chance, two successive 'warnings'

would be a trigger for action (the probability of two such events occurring in succession is 0.0025) There are other rules such as an unusually large number of values on one or other side of the mean or a run of successively increasing results.

193. The control limits provide one possible range of values that can be considered as representing the normal range such that values which fall within the limits can be considered within the range of normal data. The 2SD confidence limits have been suggested previously as possible bounds outside which effects are considered biologically important.

194. The main limitation with the use of such bounds is that variability between experiments within the same laboratory (over-dispersion) will widen the control limits. This has the disadvantage of introducing a bias into the assessment in that those laboratories that have greater inter-experimental variability will be 'rewarded' by having wider control limits and more results from treated cultures that could be considered within the normal range while a laboratory with minimal inter-experimental variability would be 'penalized' by having more of their experimental results considered outside the normal range.

195. An example could be a scenario where there is a positive trend and/or a statistical significant increase in at least one concentration but in a laboratory with appreciable over-dispersion so that all the treated values are below the upper limit of the negative control range. However, in a laboratory with less over-dispersion some of the values might be above the upper limit.

196. Margolin *et al* (1986) addressed the issue of whether this over-dispersion could be removed based upon their analysis of cytogenetic studies. They argued that it could be removed and stated: "Achievement of Poisson sampling variability for replicate control flask is an achievable goal" and that "homogeneous binomial sampling is an achievable goal, even across experiments."

197. They further argued that this also has the advantage of improving the test by reducing the false-positive and false-negative rates and that "Laboratories that employ this protocol, but consistently exhibit variability between replicate flasks that is substantially in excess of the Poisson model, can infer that there is a controllable yet undetected source of random variability in the way in which they execute the protocol".

S8) Poisson control limits

198. One possible approach to the complications that the extra between-experiment variability introduces into the assessment of control charts is to base the control limits on a theoretical standard deviation for all laboratories rather than the empirically determined value for each individual laboratory. In the case of count data the *in vitro* chromosomal aberrations test and the *in vitro* micronucleus test the Poisson distribution provides the proportion of expected counts for a given mean value (Table 10, Figure 23) which have the property that the mean is equal to the square of the standard deviation.

199. Examples of 'C' charts, QC charts for counts of abnormalities (or 'defects') in samples, and based upon the Poisson distribution, were produced for laboratories providing historical control data). The charts are shown for those laboratories where data were available from those *in vitro* micronucleus and *in vitro* cytogenetic tests where counts from cultures with 100 or 1000 cells had been scored (see Appendices A and B).

200. Similar rules and tests are used to asses whether the process is in control. The advantage of this approach is that laboratories which have low variability are now 'rewarded' by fewer values outside the control limits than those which have appreciable between experiment variability.

201. Counts that lie within a chosen interval based upon these distributions would, therefore, not be considered unusual but those counts that lie outside would be. An argument could be made that those counts within the interval based upon the Poisson distribution for the laboratory's mean incidence are considered within the normal range. It would be narrower than limits based upon a laboratory's full data set which may be over-dispersed and includes inter-experiment variability. Any significant results from treated cultures which fell within the narrower Poisson-derived limits range might be considered biologically unimportant.

Conclusions

202. It should always be appreciated that finding statistical significance is not the main objective of a statistical analysis and that statistical significance does not automatically equate to biological importance.

203. Power calculations are important tools for designing experiments but are sensitive to assumptions especially on the estimate of the variability (standard deviation (SD)). Confidence intervals for an SD can be wide when the sample size that the SD is based upon is small.

204. The *in vivo* designs for micronuclei with n = 5 animals have the power to detect 2 to 3-fold effects with 80% power based upon counts of about 4000 cells per animal when the background incidences are relatively high (0.1% and higher). The power increases with higher background control incidences.

205. Larger sample sizes, either as more animals and/or many more cells, would be needed to have sufficient power to detect a 2-3 fold incidence when the background incidence is lower (i.e. <0.05%).

206. Power calculations derived by Kissling *et al* and Health Canada are broadly in agreement. Differences between them may derive from the use of different statistical tests and methods for incorporating inter-individual variability in the simulations. (They also broadly agree with those derived by other workers in the field.)

207. In the case of the *in vitro* tests, power calculations and simulations provide similar conclusions that sample sizes (cells or cultures/animals) would need to be increased appreciably to detect doubling. The power to detect doubling/fold increase is low especially when the background level is low (i.e. 0.5% for chromosomal aberrations or 0.05% for micronuclei.

208. The power of the *in vitro* chromosome aberration test to detect a fold increase is particularly low. The number of cells needed to detect a fold increase would need to be increased appreciably to detect doubling. The actual numbers depend upon the background incidence.

209. Replicate (duplicate) negative control samples are advisable on theoretical grounds (at least until it can be shown that there is evidence of no appreciable replicate variability). Increasing the number of replicate cultures increases the statistical 'credibility' of the result but needs to be weighted against any increased complexity in the conduct of the study introducing extra variability.

210. Increasing the number of cells in the negative control cultures (from say 200 to 400) provides a small increase in power and a better estimate of the negative control incidence. This provides more robust comparisons when multiple concentrations are tested but increases the complexity of the study.

211. If duplicate negative control cultures are homogeneous then the data can be pooled across cultures. This is a pragmatic approach to increasing the power a little but it is important to appreciate that this may affect (slightly) the error rates associated with the analysis. If duplicate cultures are significantly different from one another this probably introduces quality control issues and raises the question as to whether the experiment meets acceptance criteria.

212. Trend tests with a smaller number of concentration levels have slightly higher power based upon the Nam equation approach than those with more levels. Power is quite high for detecting linear trends but will be reduced and the interpretation more complex if the response is not linear. The use of 6 concentration levels instead of 3 concentration levels may reduce the power of the linear trend test slightly but may give more information on the shape of the concentration-response relationship. More concentration levels should be used when there is specific interest in the concentration-response.

213. Trends tests are statistically more powerful for detecting linear concentration-response relationships than pair-wise comparisons. Statistical tests are available for identifying concentration-response relationships which differ from linear.

214. Use of multiple comparison methods is controversial especially when the number of concentration groups differs. The Cochran-Armitage test for linear trend is a 'powerful' test while Fisher's exact tests with a multiple comparison correction for pairs of results are 'conservative'. It is important that the specific test used to produce a P-value and declare a result as statistically significant or not is explicitly stated.

215. The *in vitro* chromosomal aberration data showed no obvious differences between the different cell lines or different S9 mix conditions. The range of mean values (% aberrations) combined over all conditions for each of the cell lines was 0.20 to 1.07 for HPBL cells, 0.53 to 0.81 for CHO cells and 0.09 to 1.06 (with an outlying laboratory with a mean of 2.16%). The lowest value for an individual culture was 0%, of which there were many examples and the highest value was 6%.

216. The micronucleus data showed appreciable intra- and inter-laboratory variability making it difficult identify a range that could be considered a 'normal range' across all laboratories. The range of mean values (% micronuclei) combined over all conditions for each of the cell lines was 0.45 to 1.39 for HPBL cells, 0.21 to 0.59 for L5178Y cells and 0.42 to 1.13 for TK6 cells. The lowest value for an individual culture was 0%, of which there were many examples and the highest value was 2.8% (28/1000).

217. In the chromosome aberration dataset values ≥5 or aberrant cells rarely occurred in a 100 cell sample and, based upon the Poisson distribution, when the laboratory negative control mean is between 0.5% and 2.5% values ≥4 aberrant cells have low probability and so values between 0 and 4 could be considered to fall into the normal range.

218. Poisson distributed data have the property that the mean is equal to the variance. A square root transformation (or a logarithmic transformation) can be used to allow the data to be analysed by standard statistical methods. More sophisticated modelling methods are also suitable methods for analysis of the data.

219. The Poisson distribution approximates to a normal distribution when np ≥5 where n is the number of cells and the mean proportions of cells with aberrations or micronuclei is p. Scoring sufficient cells from a culture to produce mean counts of 5 should provide data which approximates to a normal distribution.

220. Zero counts can be avoided by counting enough cells which are Poisson distributed to obtain a mean number of counts/culture of 5. If this is achieved on average only 0.7% of cultures would have zero counts. To obtain np = 5 from cultures with a mean of 1% chromosomal aberrations, 500 cells need to be scored; for cultures with 0.1% micronucleated cells 5000 cells would need to be scored.

221. While counting sufficient cells to produce means of approximately 5 is convenient for avoiding zero counts and for obtaining approximately normally distributed data this is not a prerequisite for statistical analysis as statistical method are available for handling such situations.

222. Analyses using the concurrent control should be the primary analysis of the experimental data to test for treatment-related effects. Historical control data may help with the interpretation of the biological importance of the results. (Data to be included in the formal analysis should be within acceptance criteria. These criteria may need to be defined.)

223. Laboratories should provide historic negative control data using quality control methods such as control charts to identify how variable their data are and to show that the methodology is 'under control' in their laboratory. A process in 'control' "exhibits only random variation within the control limits"

224. It is important to ensure that experiments that 'fail' (i.e. do not meet acceptance criteria) are also included in the historical control dataset.

225. Most methods for using the historical control data to help determine a positive result and in the assessment of the biological importance of the results have some limitations. The GEF (Global Equivalence Factor) approach used in the Mouse Lymphoma Assay (MLA) test may have some potential.

226. Control limits based upon the between cultures (or between study) standard deviations can be used to specify upper and lower control limits. Use of these when there is appreciable between-laboratory variability can 'reward' laboratories that have high variability.

227. The use of C-charts, developed for quality control purposes with control limits based upon the Poisson distribution to set the UCL could provide an indication of results that would not be considered unusual as a way of identifying 'a normal range' for the historical control range as well as those which would indicate an effect had occurred.

References

Galloway, S., Holder, D.J. & Armstrong, M. J. (2012) Sensitivity and power of the bone marrow micronucleus (MN) test related to background level and sample sizes. Poster presented at the GTA meeting 2012.

Haseman, J.K., Huff, J. & Boorman, G.A (1984) Use of historical control data in carcinogenicity studies in rodents Toxicologic Pathology. 12 126-135.

Hayes, J., Doherty, A.T., Adkins, D.J., Oldman, K. & O'Donovan, M.R. (2009) The rat bone marrow micronucleus test—study design and statistical power. Mutagenesis 24 419-424.

Hayashi, M., Dearfield, K., Kasper P., Lovell D. & Thybaud, V. (2011) Compilation and use of genetic toxicity historical control data. Mutation Research 723 87-90.

Health Canada (2012) Discussion Document on Assay Sensitivity, Sample Size and Statistical Power for the Erythrocyte Micronucleus Assay (TG 474) produced for the OECD expert group meeting.

Kissling, G.E., Dertinger, S.D., Hayashi, M. & MacGregor, J.T. (2007) Sensitivity of the erythrocyte micronucleus assay: dependence on number of cells scored and inter-animal variability. Mutation Research 634 235–240.

Margolin B.H., Resnick, M.A., Rimpo, J.Y., Archer, P., Galloway, S.M., Bloom, A.D. & Zeiger, E. (1986) Statistical analyses for *in vitro* cytogenetic assays using Chinese hamster ovary cells. Environ Mutagen. 8 183-204.

Moore, M.M., Honma, M., Clements, J., Bolcsfoldi, G., Burlinson, B., Cifone, M., Clarke, J., Delongchamp, R., Durward, R., Fellows, M., Gollapudi, B., Hou, S., Jenkinson, P., Lloyd, M., Majeska, J., Myhr, B., O'Donovan, M., Omori, T., Riach, C., San, R., Stankowski, L.F. Jr., Thakur, A.K., Van Goethem, F., Wakuri, S. & Yoshimura, I. (2006). Mouse lymphoma thymidine kinase gene mutation assay: follow-up meeting of the international workshop on genotoxicity testing, Aberdeen, Scotland, 2003, assay acceptance criteria, positive controls, and data evaluation, Environmental and Molecular Mutagenesis 47 1-5.

Nam JM. (1987) A simple approximation for calculating sample sizes for detecting linear trend in proportions. Biometrics 43 701-705

Soper, K.A. & Galloway S.M. (1994). Replicate flasks are not necessary for in vitro chromosome aberration assays in CHO cells. Mutation Research 312 139-149.

Table 1. 95% upper and lower confidence limits of estimates of Kissling *et al*'s (2007) (Table 3) standard deviations

Mean	SD	n	LCL	UCL
0.05	0.045	15	0.0329	0.071
0.1	0.059	190	0.0536	0.0656
0.2	0.070	79	0.0605	0.083
0.3	0.093	22	0.0715	0.1329

Table 2. Fold changes detectable at 90% and 95% power for various negative control incidence and numbers of cells counted as derived by Kissling *et al* and Health Canada (group size =5)

%		90% Canada	0.05 Kissling	90% Canada	0.01 Kissling	95% Canada	0.05 Kissling	95% Canada	0.01 Kissling
	2000	4.1	4.5	5.2	6.8	4.6	5.6	5.8	9.3
0.05	4000	2.9	3.5	3.5	5.5	3.2	4.0	3.9	6.3
	20000	1.7	2.3	1.9	3.1	1.8	2.4	2.0	3.4
0.1	2000	2.9	3.3	3.5	4.8	3.3	4.1	3.8	6.4
	4000	2.3	2.9	2.6	4.2	2.5	3.2	2.8	4.7
	20000	1.5	2.2	1.6	3.0	1.6	2.4	1.7	3.2
0.2	2000	2.3	2.7	2.6	3.9	2.4	3.0	2.8	4.5
	4000	1.8	2.3	2.0	3.2	1.9	2.4	2.2	3.5
	20000	1.3	1.9	1.4	2.5	1.4	2.2	1.5	2.7
0.3	2000	2.0	2.4	2.2	3.4	2.1	2.6	2.4	3.7
	4000	1.7	2.1	1.8	2.8	1.7	2.2	1.9	3.0
	20000	1.3	1.8	1.3	2.3	1.3	1.9	1.4	2.4

Table 3. Results of simulations using Health Canada's R program to investigate the effect of varying numbers of animals and cells. Results are the fold increases detectable for 80% power at P=0.05 (for group sizes, n, = 4, 5 and 6)

%	n	4	5	6
0.05	2000	3.8	3.5	3.3
	4000	2.8	2.6	2.4
	8000	2.2	2.0	1.9
	20000	1.7	1.6	1.5
0.10	2000	2.7	2.6	2.4
	4000	2.2	2.0	1.9
	8000	1.8	1.7	1.6
	20000	1.4	1.4	1.4
0.20	2000	2.2	2.0	1.9
	4000	1.8	1.7	1.6
	8000	1.5	1.5	1.4
	20000	1.3	1.3	1.3
0.30	2000	1.9	1.8	1.7
	4000	1.6	1.6	1.5
	8000	1.4	1.4	1.3
	20000	1.2	1.2	1.2

Table 4. Power calculation using nQuery Advisor to identify fold changes detected with 80% power using the means and SDs from Kissling *et al* paper (and assuming number of cells per animal =∞ as in their Table 3 for sample sizes of 4, 5 and 6)

Mean SD	4	5	6
0.10 & 0.045	1.90	1.78	1.69
0.05 & 0.02	1.80	1.69	1.62
0.20 & 0.70	1.70	1.60	1.54
0.20 & 0.59	1.59	1.51	1.46
0.30 & 0.92	1.62	1.53	1.48

Table 5. Table of descriptive statistics of data from laboratories for *in vitro* cytogenetics tests

Lab ID	Cell type	Type	% abs	Zeros	Celle	N	Mean*	St Dev	Min	Max	P-value
CB	HPBL	Count	0.24	315	200	466	0.49	0.85	0	5	<0.001
CC	HPBL	Count	1.07	63	100	146	1.07	1.25	0	6	0.01
CD	HPBL	Count	0.78	107	100	208	0.78	0.99	0	5	0.003
CE	HPBL	Count	0.74	83	100	164	0.74	0.90	0	4	0.39
CN	HPBL	Count	1.02	7	200	41	2.05	1.41	0	5	0.69
CP	HPBL	Count	0.20	105	100	128	0.20	0.44	0	2	0.89
CS	HPBL	Count	0.21	60	100	76	0.21	0.41	0	1	ND
CU	HPBL	Count	0.60	36	200	120	1.21	1.11	0	5	0.97
CA	CHO	Count	0.53	116	200	255	1.07	1.35	0	6	<0.001
CF	CHO	Count	0.81	80	100	180	0.81	0.94	0	5	0.61
CG	CHL/IU	Count	0.13	134	100	150	0.13	0.39	0	2	0.043
CH	CHL/IU	Count	0.09	280	100	300	0.09	0.36	0	3	0.001
CI	CHL/IU	Count	0.40	131	100	200	0.40	0.59	0	2	0.45
CJ	CHL/IU	Count	0.78	163	100	424	0.78	0.74	0	3	<0.001
CK	CHL/IU	Count	0.27	37	100	48	0.27	0.54	0	2	0.59
CL	CHL/IU	Count	0.50	653	100	911	0.50	0.98	0	4	<0.001
CM	CHL/IU	Count	2.16	52	200	85	4.32	2.62	0	10	<0.001
CO	CHL/IU	Count	1.06	52	100	144	1.06	1.05	0	4	0.91
CQ	CHL/IU	Count	0.22	14	100	18	0.22	0.43	0	1	ND
CR	CHL/IU	Count	0.44	32	100	54	0.44	0.57	0	2	0.11
CT	CHL/IU	Count	0.22	190	100	236	0.22	0.48	0	2	0.33

N Number of independent cultures

P- value P-value of goodness of fit test for fit to Poisson distribution

* Mean of counts from 100 or 200 cells

Note three other laboratories reported data that could be summarized

CX	CHO	2115 cultures with a mean of 1.41% with an SD of 0.91%
CX	HPBL	68 cultures with a mean of 1.73% with an SD of 0.60%
CV	HPBL	73 cultures with a mean of 1.12% with an SD of 1.07%
CW	CHL/IU	238 cultures with a mean of 0.38% with an SD of 0.70%

Table 6. Table of descriptive statistics of data from laboratories for *in vitro* micronucleus tests

Lab ID	Cell type	BC	Cyto	Type	%mnt	Zeros	Cells	N	Mean*	St Dev	Min	Max	P value
MA	HPBL	Y	NI	Count	0.58	1	1000	171	5.81	3.13	0	21	0.003
MC	HPBL	N	NI	Count	0.45	2	1000	204	4.46	2.93	0	15	<0.001
MH	HPBL	N	NI	Count	1.39	0	2000	36	27.81	17.30	6	68	N/D
MJ	HPBL	N	NI	Count	0.17	10	1000	54	1.67	1.24	0	5	0.85
MK	HPBL	N	NI	Count	0.64	0	1000	68	6.38	3.12	1	14	0.017
MM2	HPBL	N	NI	Non-integer	0.70		2000	28	13.96	4.84	7	26.4	N/D
MB	L5178Y	N	N	Count	0.21	21	1000	138	2.11	1.51	0	7	0.66
MD1	L5178Y	N	NI	Count	0.54	2	1000	518	5.44	2.61	0	19	<0.001
ML2	L5178Y	N	NI	Non-integer	0.28		?1000	767	2.80	1.21	0	8	N/D
MM1	L5178Y	N	NI	Non-integer	0.59		2000	126	11.79	4.24	4	21.8	N/D
MI	CHL/IU	N	NI	Count	0.35	0	1000	154	3.51	1.01	1	6	<0.001
MD2	TK6	N	NI	Count	1.13	1	1000	98	11.31	3.42	2	21	0.26
ME	TK6	Y	Y	Count	0.42	0	1000	59	4.15	2.14	1	11	0.44
MF	TK6	N	Y	Non-integer	0.71		1000	198	7.06	3.95	1	28	N/D
MG	TK6	N	NI	Count	0.82	0	2000	36	16.47	3.61	9	26	N/D
ML1	TK6	N	NI	Non-integer	0.42		?1000	166	4.25	1.64	0	8.3	N/D

N Number of independent cultures
NI No information
N/D Not done because data not suitable for test
?1000 Number of cells that summary statistic based on is not clear, presumed 1000
* Mean count based upon 1000 or 2000 cells
P value P value of goodness of fit test for fit to Poisson distribution

BC Specific comment about binucleated cells being scored (Y/N)
Cyto Specific comment on use of cytochalasin-B (Y/N/NI)

Table 7. Fold change that would be detected in a pair-wise test for different control incidences and sample sizes in the *in vitro* micronucleus test

	%						
Cells	0.05	0.1	0.15	0.2	0.3	0.4	0.5
1000	15.1	8.7	6.6	5.4	4.3	3.7	3.3
2000	8.8	5.5	4.3	3.7	3.0	2.7	2.5
4000	5.5	3.7	3.0	2.7	2.3	2.1	2.0
5000	4.8	3.3	2.8	2.5	2.1	2.0	1.8
10000	3.3	2.5	2.1	2.0	1.8	1.6	1.6

Table 8. Power associated with the design for detecting a fold increase or doubling over the negative control incidence in a pair-wise test for the *in vitro* micronucleus test.

	%						
Cells	0.05	0.1	0.15	0.2	0.3	0.4	0.5
1000	0.11	0.14	0.17	0.20	0.26	0.31	0.36
2000	0.14	0.20	0.26	0.31	0.41	0.50	0.57
4000	0.20	0.31	0.41	0.50	0.64	0.75	0.83
5000	0.23	0.36	0.48	0.57	0.72	0.83	0.89
10000	0.36	0.57	0.72	0.83	0.94	0.98	0.99

Table 9. Confidence intervals for the expected value (parameter) of a Poisson random variable.
(from: http://faculty.washington.edu/heagerty/Books/Biostatistics/TABLES/Poisson/index.html)

	Percent confidence					
	90		95		99	
n	Lower limit	Upper limit	Lower limit	Upper limit	Lower limit	Upper limit
0	0.00	3.00	0.00	3.69	0.00	5.30
1	0.05	4.74	0.03	5.57	0.01	7.43
2	0.36	6.30	0.24	7.22	0.10	9.27
3	0.82	7.75	0.62	8.77	0.34	10.98
4	1.37	9.15	1.09	10.24	0.67	12.59
5	1.97	10.51	1.62	11.67	1.08	14.15
6	2.61	11.84	2.20	13.06	1.54	15.66
7	3.29	13.15	2.81	14.42	2.04	17.13
8	3.98	14.43	3.45	15.76	2.57	18.58
9	4.70	15.71	4.12	17.08	3.13	20.00
10	5.43	16.96	4.80	18.39	3.72	21.40
http://faculty.washington.edu/heagerty/Books/Biostatistics/TABLES/Poisson/index.html						

Table 10. Proportions of counts of 0, 1, 2, 3 etc when each column here represents the underlying mean of the population (i.e. 1, 2, 3), For example 37% of cultures will have zero micronuclei and a further 37% will have one micronucleus when the mean level is 1 per culture.

	Mean									
N	1	2	3	4	5	6	7	8	9	10
0	0.37	0.14	0.05	0.02	0.01	0.00	0.00	0.00	0.00	0.00
1	0.37	0.27	0.15	0.07	0.03	0.01	0.01	0.00	0.00	0.00
2	0.18	0.27	0.22	0.15	0.08	0.04	0.02	0.01	0.00	0.00
3	0.06	0.18	0.22	0.20	0.14	0.09	0.05	0.03	0.01	0.01
4	0.02	0.09	0.17	0.20	0.18	0.13	0.09	0.06	0.03	0.02
5	0.00	0.04	0.10	0.16	0.18	0.16	0.13	0.09	0.06	0.04
6	0.00	0.01	0.05	0.10	0.15	0.16	0.15	0.12	0.09	0.06
7	0.00	0.00	0.02	0.06	0.10	0.14	0.15	0.14	0.12	0.09
8	0.00	0.00	0.01	0.03	0.07	0.10	0.13	0.14	0.13	0.11
9	0.00	0.00	0.00	0.01	0.04	0.07	0.10	0.12	0.13	0.13
10	0.00	0.00	0.00	0.01	0.02	0.04	0.07	0.10	0.12	0.13

Table 11. The percentage of cultures with from 0 to 15 aberrant or micronucleated cells in samples of size n cells when the 'true' mean number in cultures range from 0.5 to 5.

	Number of cells with an aberration or a micronucleus in culture															
Mean	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0.5	60.7	30.3	7.6	1.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	36.8	36.8	18.4	6.1	1.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.5	22.3	33.5	25.1	12.6	4.7	1.4	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2.0	13.5	27.1	27.1	18.0	9.0	3.6	1.2	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2.5	8.2	20.5	25.7	21.4	13.4	6.7	2.8	1.0	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
3.0	5.0	14.9	22.4	22.4	16.8	10.1	5.0	2.2	0.8	0.3	0.1	0.0	0.0	0.0	0.0	0.0
3.5	3.0	10.6	18.5	21.6	18.9	13.2	7.7	3.9	1.7	0.7	0.2	0.1	0.0	0.0	0.0	0.0
4.0	1.8	7.3	14.7	19.5	19.5	15.6	10.4	6.0	3.0	1.3	0.5	0.2	0.1	0.0	0.0	0.0
4.5	1.1	5.0	11.3	16.9	19.0	17.1	12.8	8.2	4.6	2.3	1.0	0.4	0.2	0.1	0.0	0.0
5.0	0.7	3.4	8.4	14.0	17.6	17.6	14.6	10.4	6.5	3.6	1.8	0.8	0.3	0.1	0.1	0.0

Mean is count per culture. Mean of 5 is an incidence of 1% if 500 cells are scored or 5% if 100 cells are scored. (Assumes Poisson distributions. If an individual laboratory scores are 'over-dispersed' then the number of zeros may be increased for a given mean.)
Poisson distribution approximates to a normal distribution when np >5

Table can be used to identify the number of cells to score to have a low probability of encountering a zero count. For instance, to have <1% of cultures having zero scores when the mean is 2.5/100 cells then 200 cells/culture would need to be scored.

Figure 1. Histogram of mean % incidence of chromosomal aberration in 65 combinations of cell types, S9 fractions and treatment times

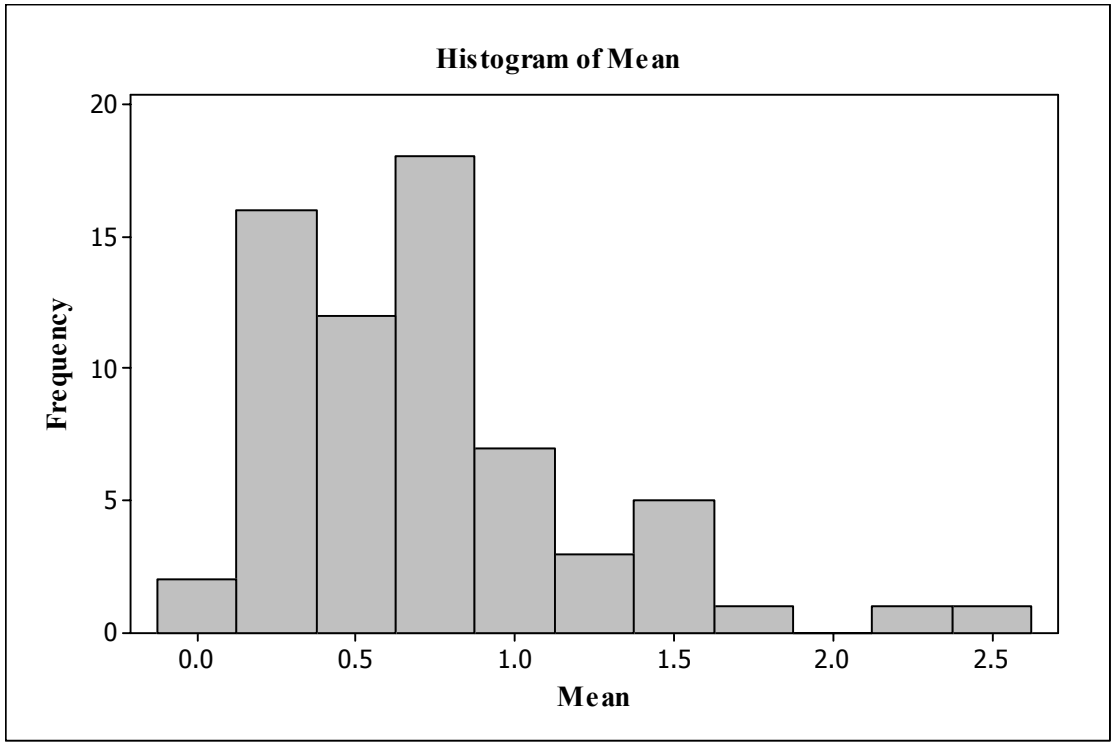


Figure 2. Means and 95% confidence intervals for the incidence of chromosomal aberrations in 65 sets of samples from 19 laboratories. Colour codes: Black, HPBL; Red, CHO; Green, CHL/IU. (Means are % cells with chromosomal aberrations) (The size of the confidence interval depends in part on the sample size: very narrow intervals are based upon large samples; wide intervals on small sample sizes)

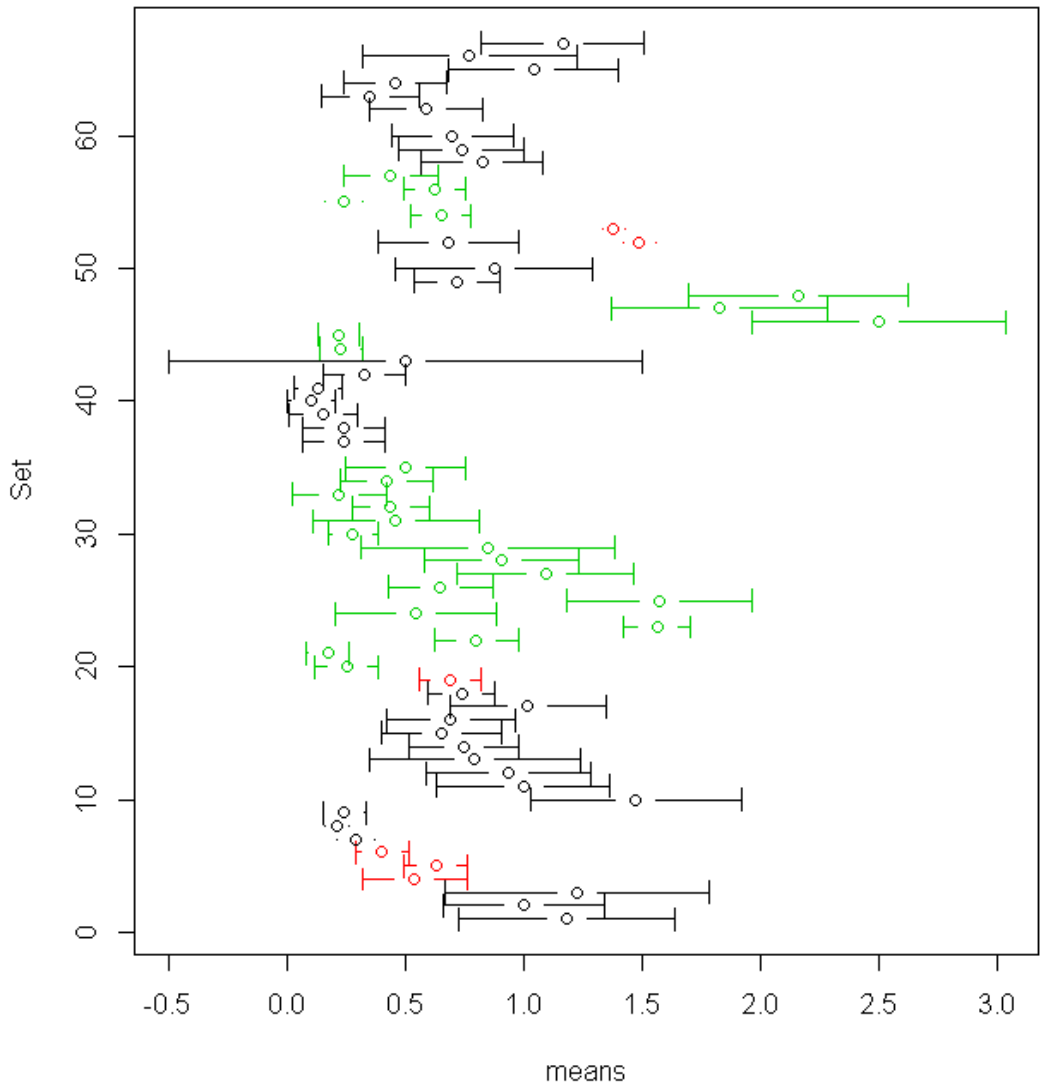


Figure 3. Means and standard deviations for the incidence of chromosomal aberrations in 65 sets of samples from 19 laboratories. Colour codes: Black, –S9 fraction; Red, +S9 fraction; Green, samples combined over S9 conditions (Means are % cells with chromosomal aberrations)

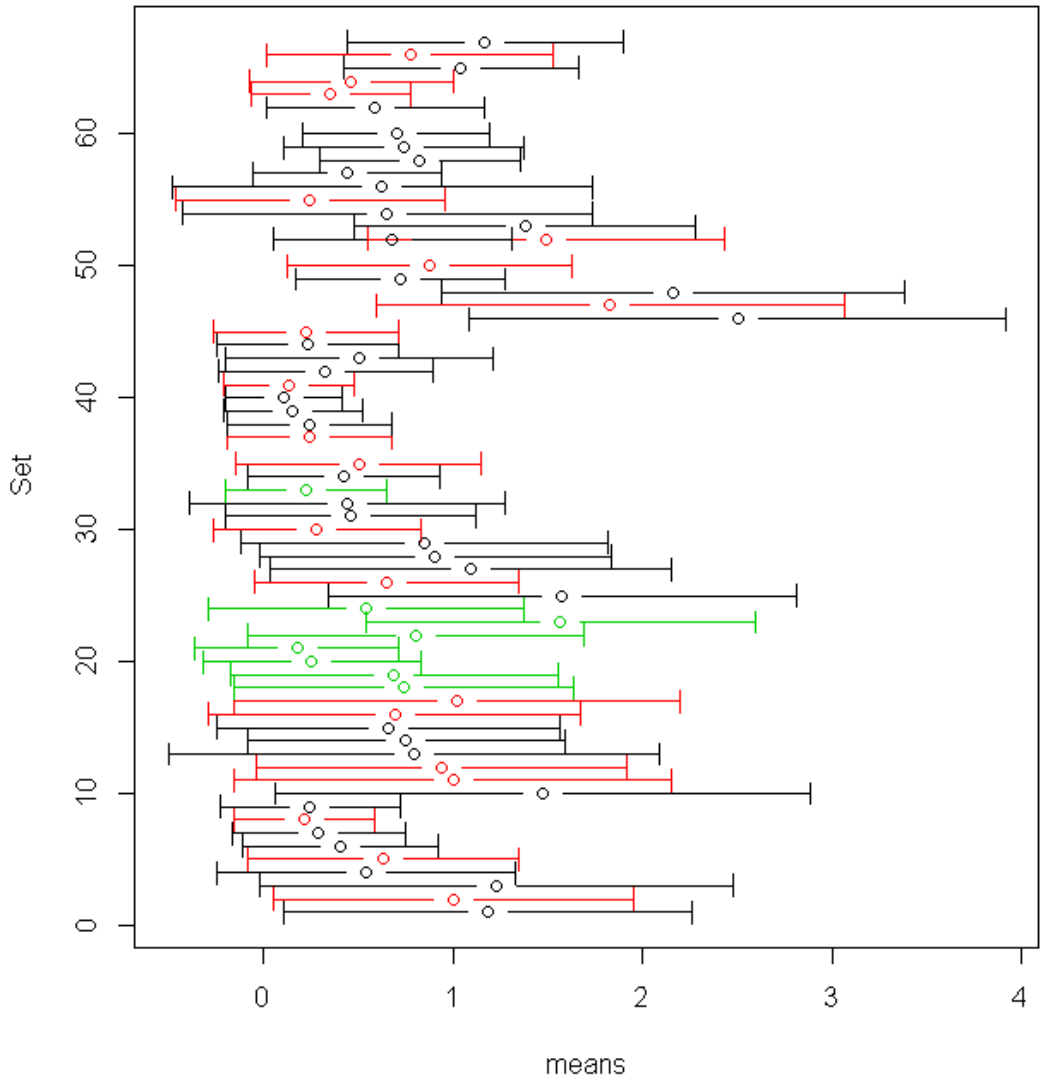
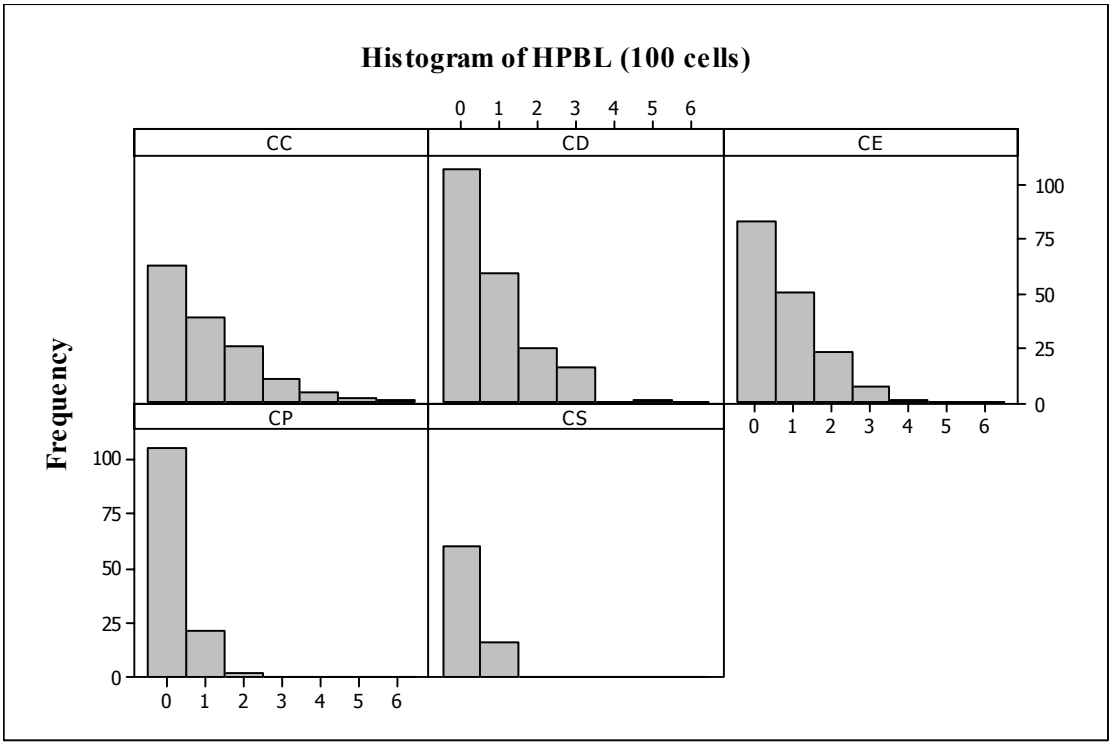
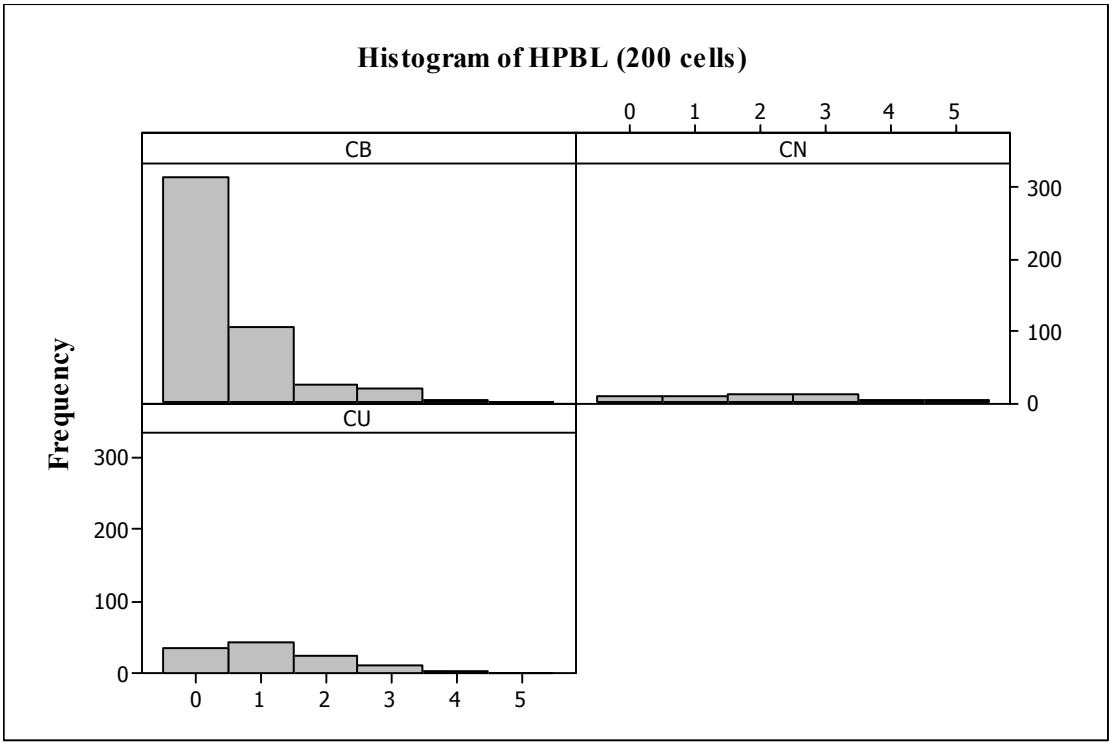


Figure 4. *In vitro* chromosome aberration test: Distribution of individual counts of numbers of chromosomal aberrations from (where possible) individual cultures for each laboratory, HPBL cells (n=100)



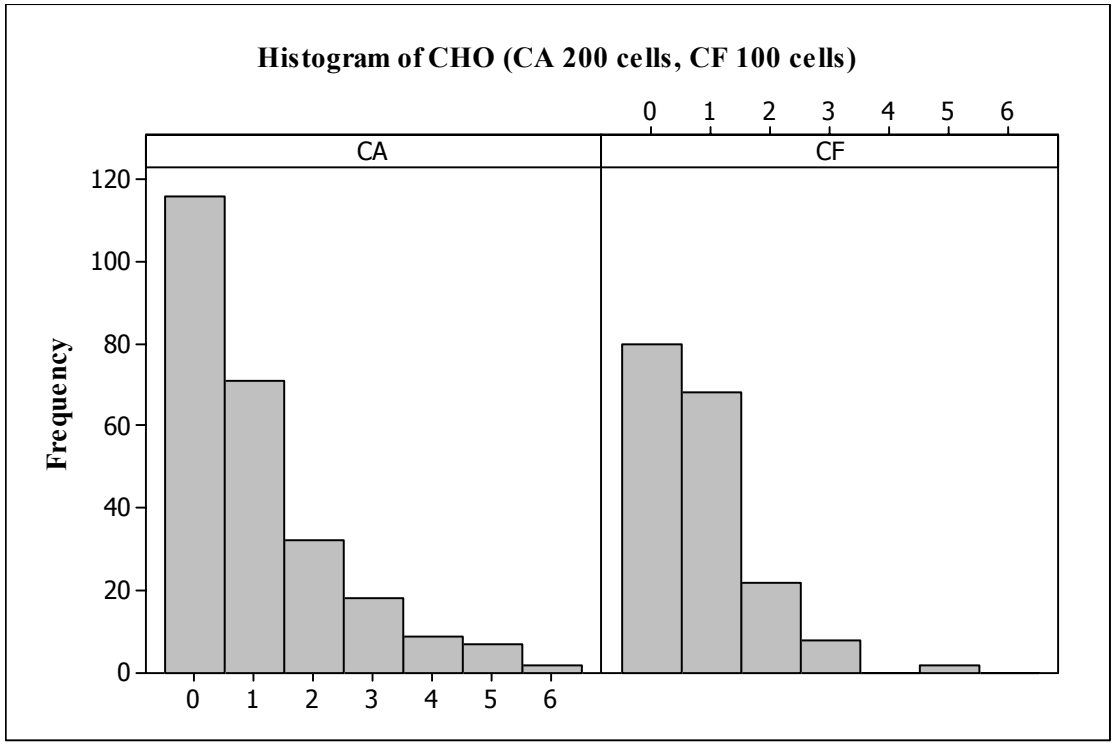
CC Counts from % Cabs based upon 100 cells
CD Counts based upon 100 cells
CE Counts based upon approx 100 cells
CP Counts based upon 100 cells except some on 50 and 200 which were mainly negative
CS Counts based upon 100 cells for single culture

Figure 5. *In vitro* chromosome aberration test: Distribution of individual counts of numbers of chromosomal aberrations from (where possible) individual cultures for each laboratory, HPBL cells (n=100) and HPBL cell (n=200)



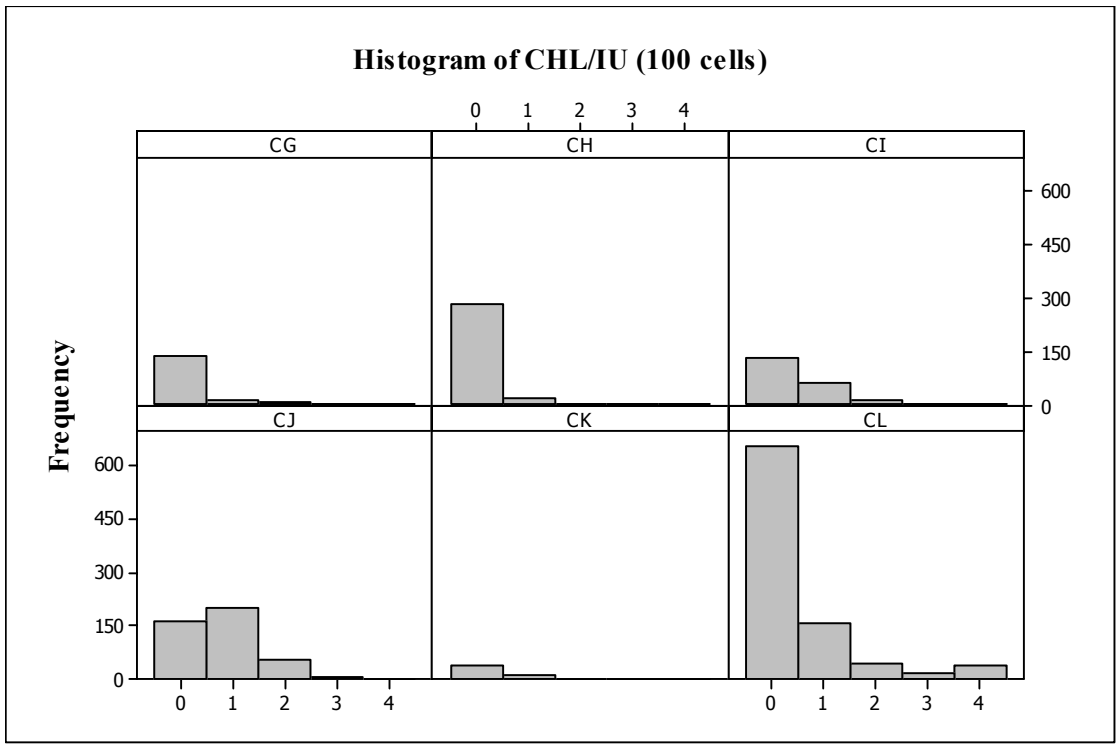
CB % abs based upon 200 cells x2 to create counts
CN % abs based upon 200 cells x2 to create counts
CU % abs based upon 200 cells x2 to create counts

Figure 6. *In vitro* chromosome aberration test: Distribution of individual counts of numbers of chromosomal aberrations from (where possible) individual cultures for each laboratory, CHO cells



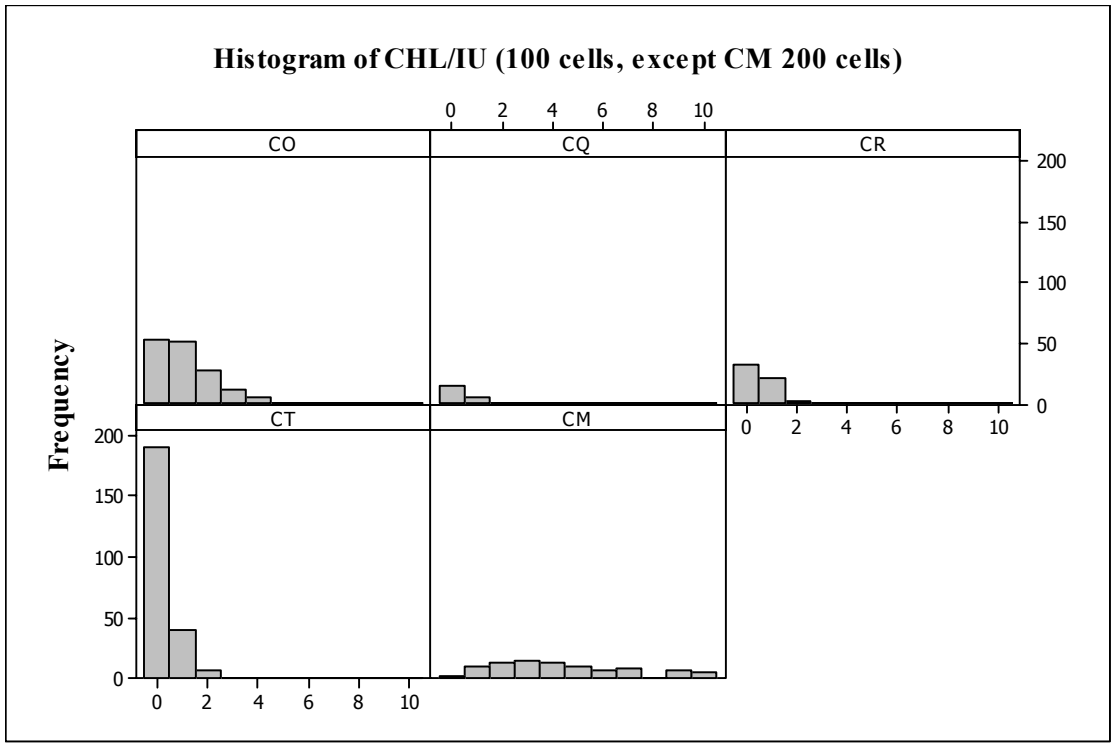
CA % abs based upon 200 cells x 2 to create counts
CF Counts based upon approx 100 cells

Figure 7. *In vitro* chromosome aberration test: Distribution of individual counts of numbers of chromosomal aberrations from (where possible) individual cultures for each laboratory, CHL/IU cells



CG Counts based upon approx 100 cells
CH Counts based upon approx 100 cells
CI Counts based upon approx 100 cells
CJ Counts based upon approx 100 cells
CK Counts based upon approx 100 cells
CL Counts based upon 100 cells except some on 50 which were mainly zero

Figure 8. *In vitro* chromosome aberration test: Distribution of individual counts of numbers of chromosomal aberrations from (where possible) individual cultures for each laboratory, CHL/IU cells



CO Counts based upon 100 cells
CQ Counts based upon 100 cells
CR Counts based upon 100 cells for single culture
CT Counts based upon 100 cells for single culture
CM Counts based upon 200 cells summed over 2 cultures

Figure 9. Power curves for pair-wise test for different number of cell (50 to 500) with an incidence of 0.5% (proportion p of 0.005) chromosome aberrations in the negative control culture

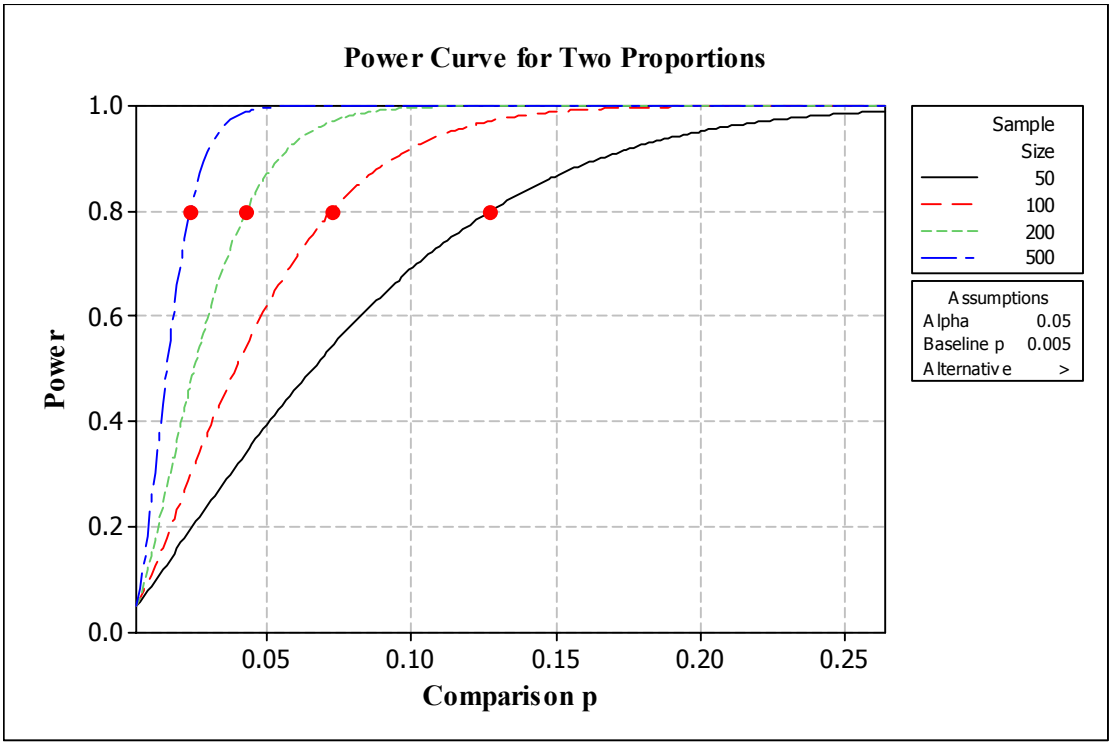


Figure 10. Power curves for pair-wise test for different number of cell (50 to 500) with an incidence of 1.0% (proportion p of 0.01) chromosome aberrations in the negative control culture

Figure 11. Power curves for pair-wise test for different number of cell (50 to 500) with an incidence of 2.0% (proportion p of 0.02) chromosome aberrations in the negative control culture

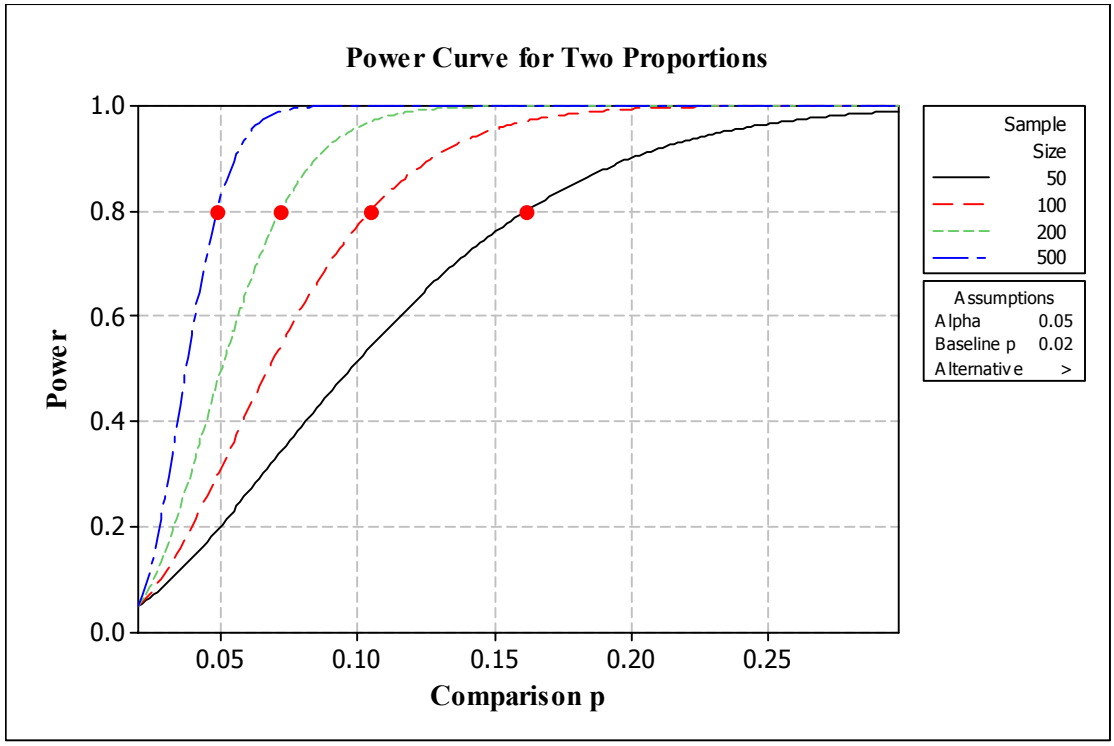


Figure 12. Histogram of mean % incidence of micronucleated cells in 53 combinations of cell types, S9 fractions and treatment times from 13 laboratories

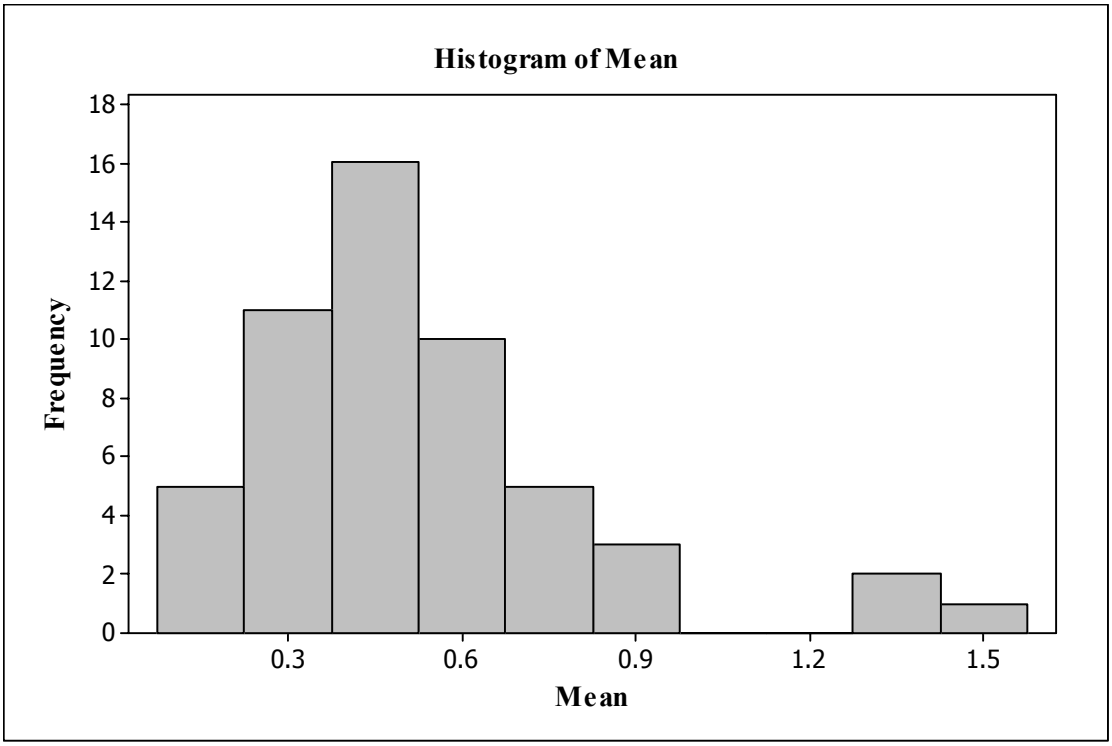


Figure 13. Means and 95% confidence intervals for the incidence of micronuclei in 53 sets of samples from 13 laboratories. Colour codes: Black, L5178Y; Red, HPBL; Green, TK6; Blue, CHL/IU (Means are cells with micronuclei/1000 cells) (The size of the confidence interval depends in part on the sample size: very narrow intervals are based upon large samples; wide intervals on small sample sizes)

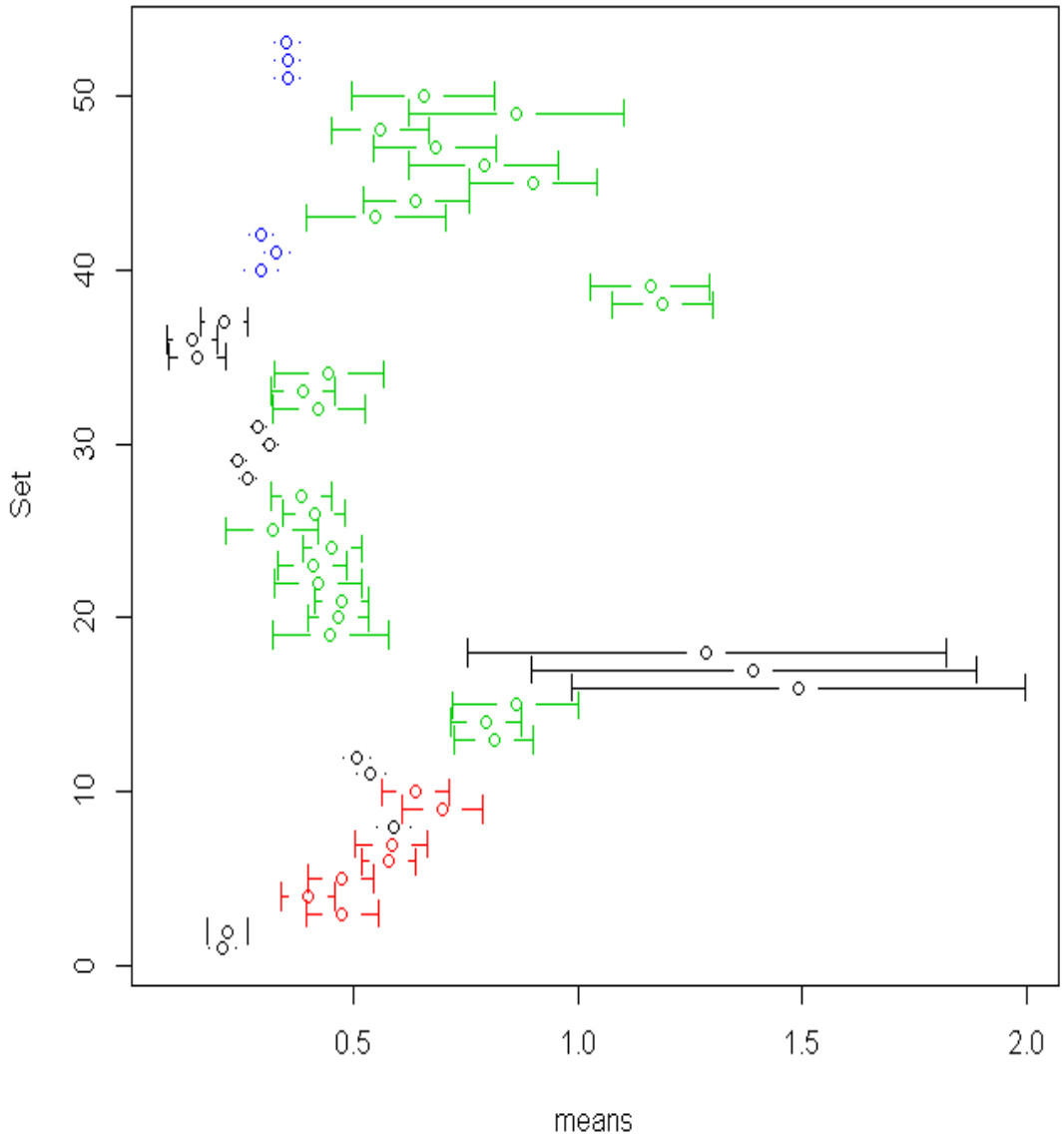


Figure 14. Means and standard deviations for the incidence of micronuclei in 53 sets of samples from 13 laboratories. Colour codes: Black, –S9 fraction; Red, +S9 fraction; Green, samples combined over S9 conditions (Means are cells with micronuclei/1000 cells)

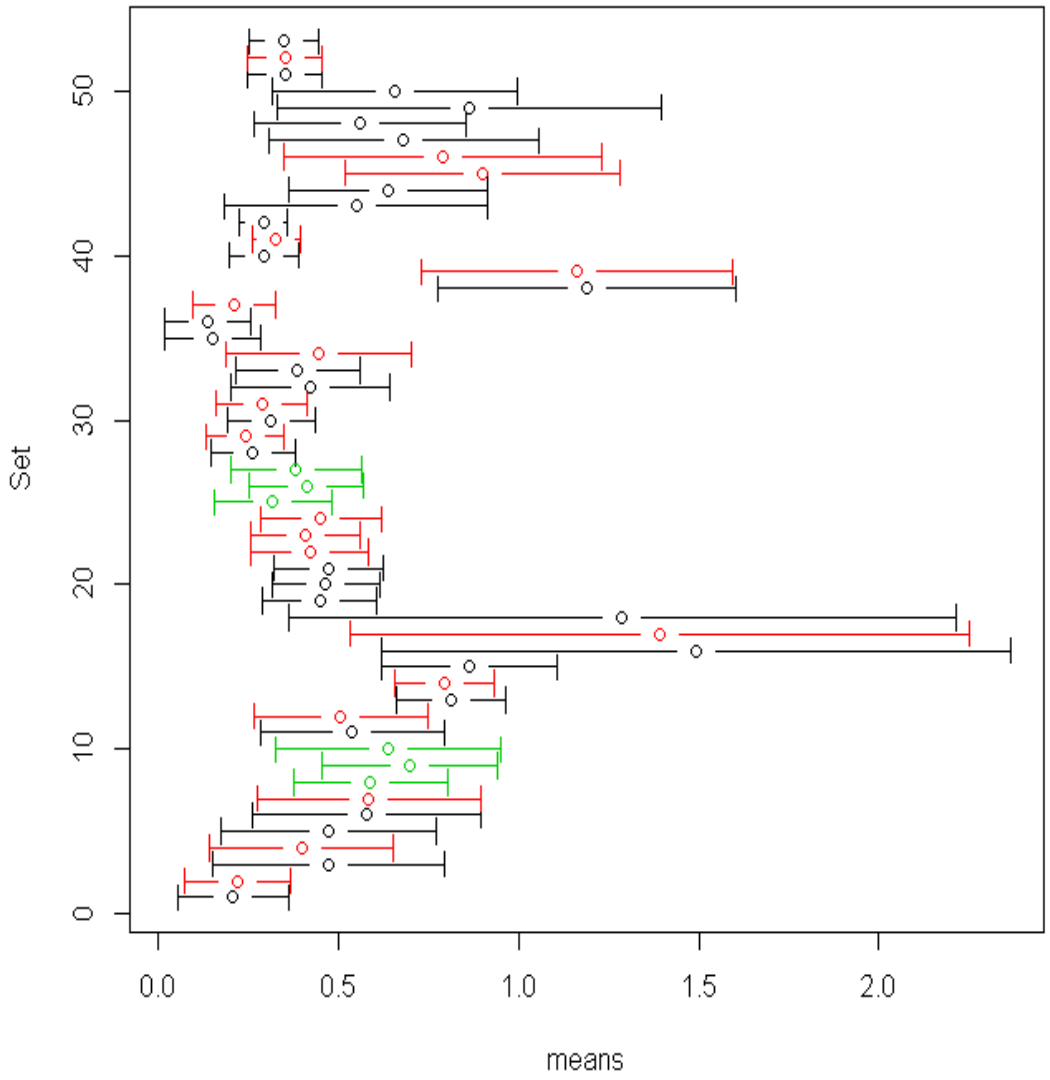
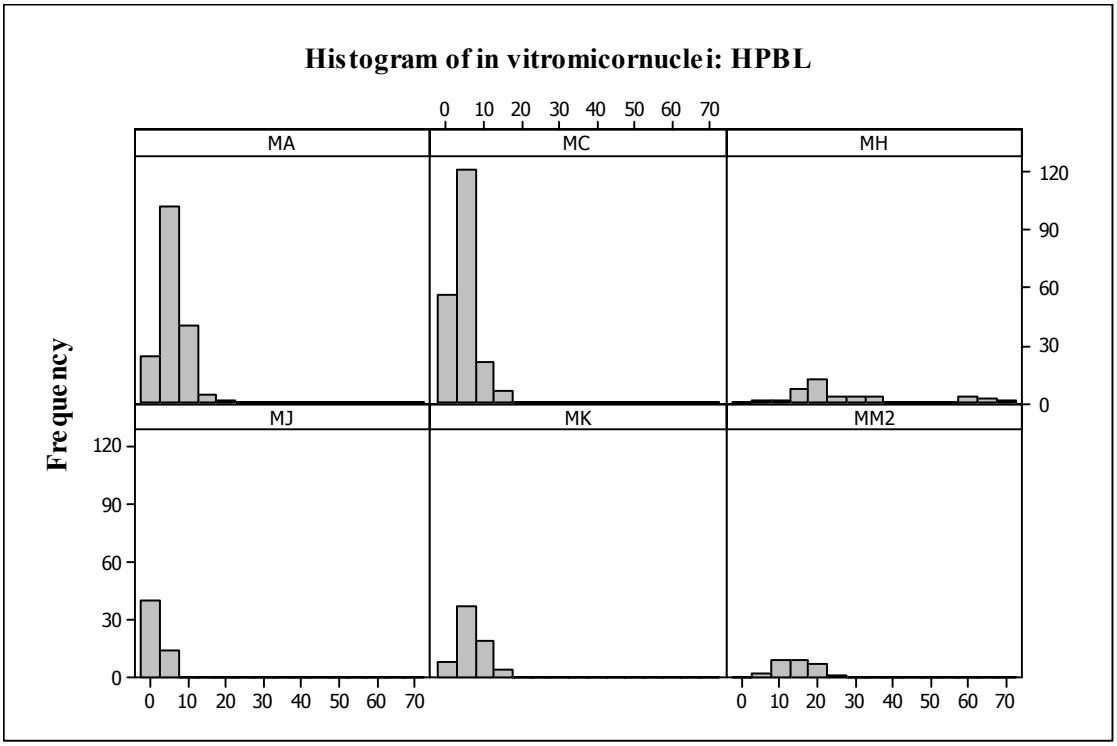
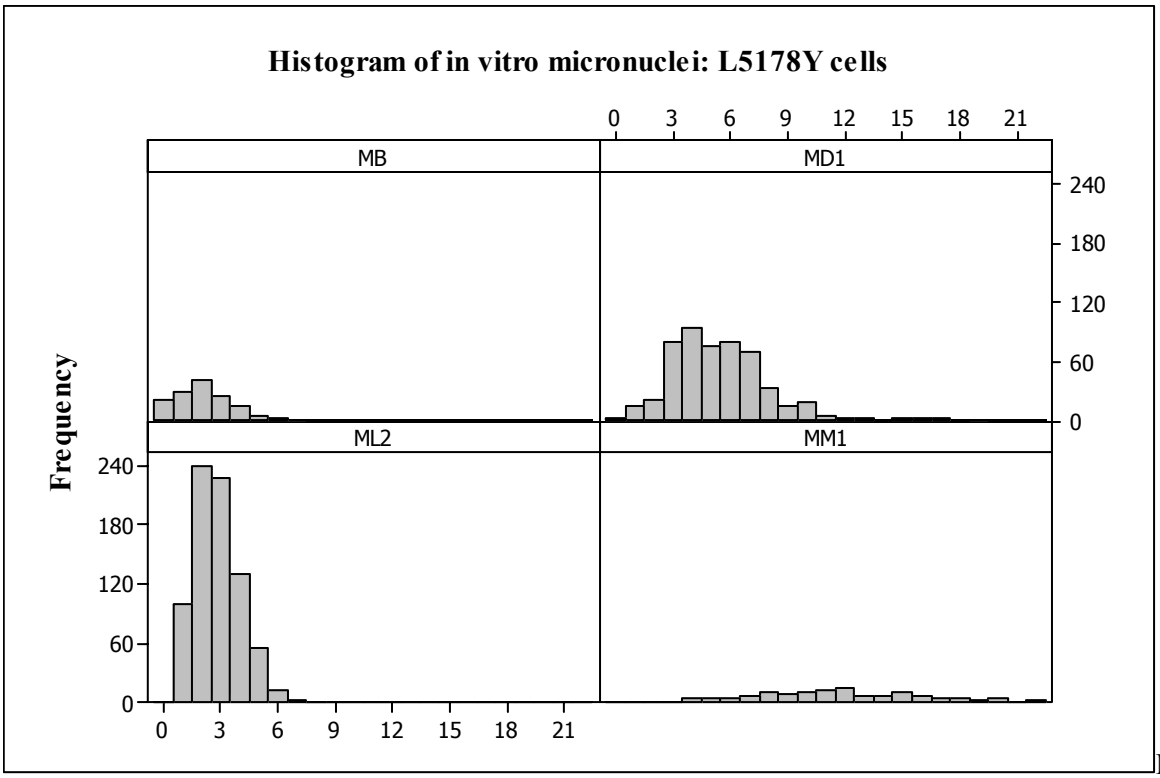


Figure 15. *In vitro* micronucleus test: Distribution of individual counts of numbers of micronuclei from (where possible) individual cultures for each laboratory: HPBL cells



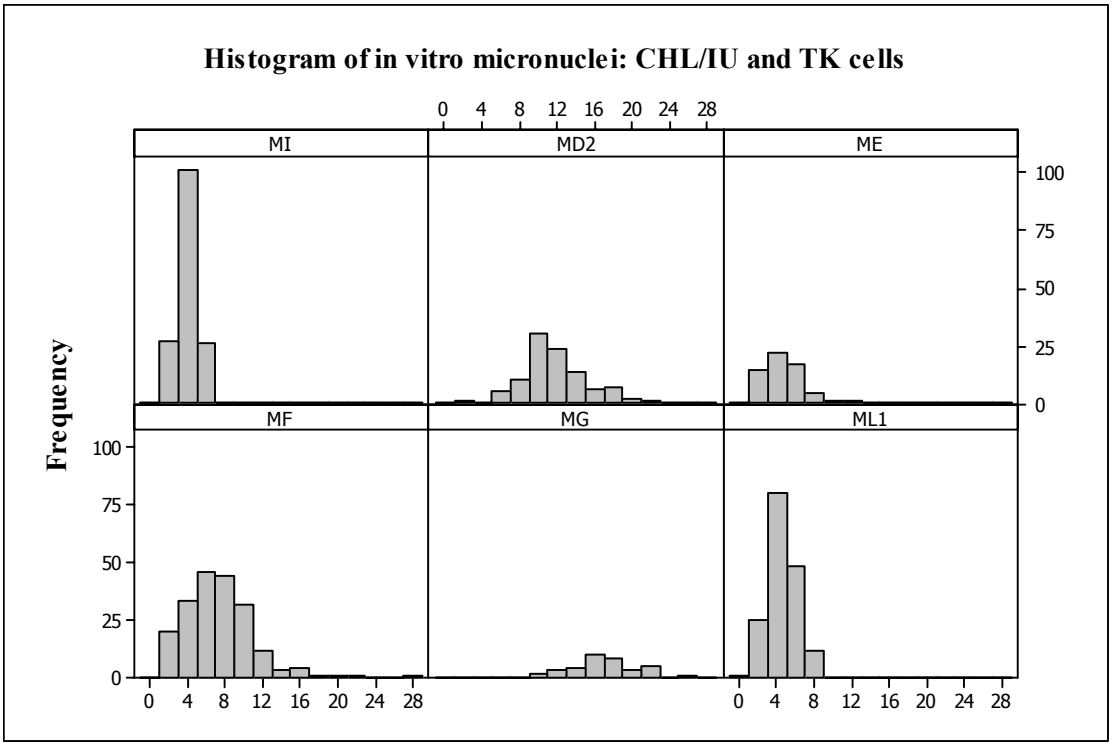
MA count from 1000 cells
MC count from 1000 cells
MH count based upon 2000 cells (data converted from %MN to count based upon 2000 cells)
MJ count from 1000 cells
MK count from 1000 and some 500 cells
MM2 count based upon 2000 cells (data converted from %MN to count based upon 2000 cells)

Figure 16. *In vitro* micronucleus test: Distribution of individual counts of numbers of micronuclei from (where possible) individual cultures for each laboratory: L5178Y cells



MB count from 1000 cells
MD1 count from 1000 cells (slight adjustments)
ML2 %MN x 10 (not possible to recreate counts)
MM1 count based upon 2000 cells (data converted from %MN to count based upon 2000 cells)

Figure 17. *In vitro* micronucleus test: Distribution of individual counts of numbers of micronuclei from (where possible) individual cultures for each laboratory: CHL/IU and TK6 cells



CHL/IU cells
MI %MN x 10 from 1000 cells
TK6 cells
MD2 count from 1000 cells (slight adjustments)
ME count from 1000 cells
MF 'approximate' counts from 1000 cells
MG based upon 2000 cells (data converted from %MN to count based upon 2000 cells)
ML1 %MN x10 (not possible to recreate counts)

Figure 18. Power curves for pair-wise test for different numbers of cells (1000 to 10000) with an incidence of 0.05% (proportion p of 0.0005 or 0.5/1000) micronuclei in the negative control culture

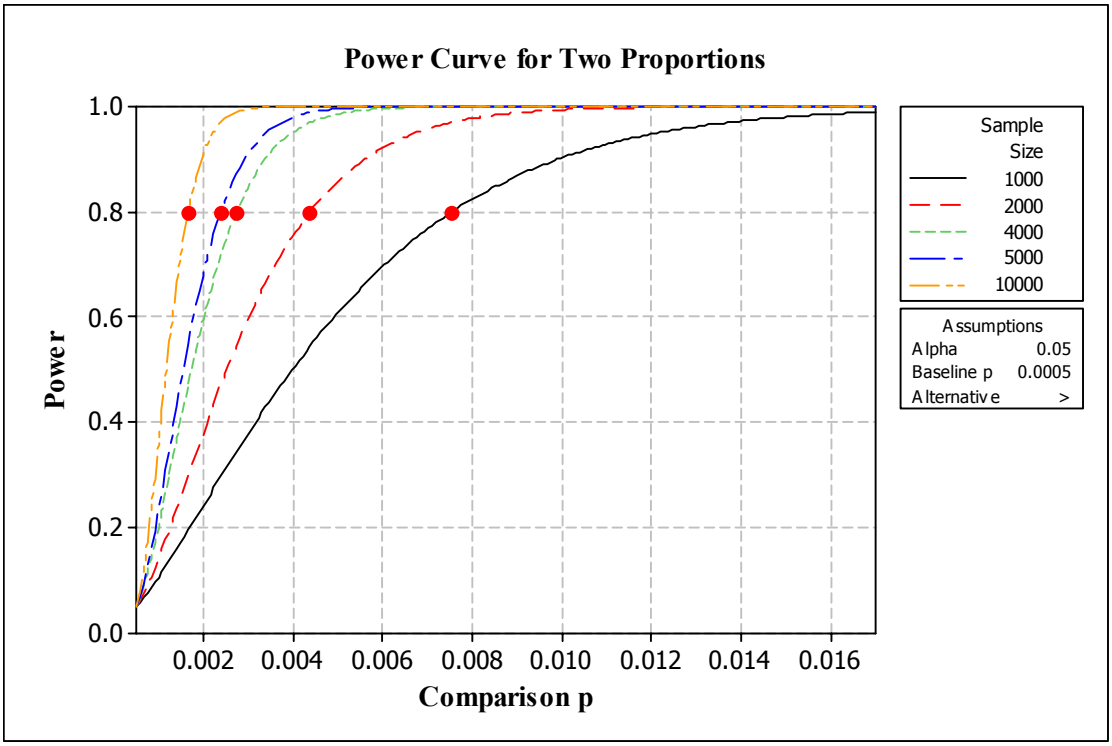


Figure 19. Power curves for pair-wise test for different numbers of cells (1000 to 10000) with an incidence of 0.1% (proportion p of 0.001 or 1/1000) micronuclei in the negative control culture

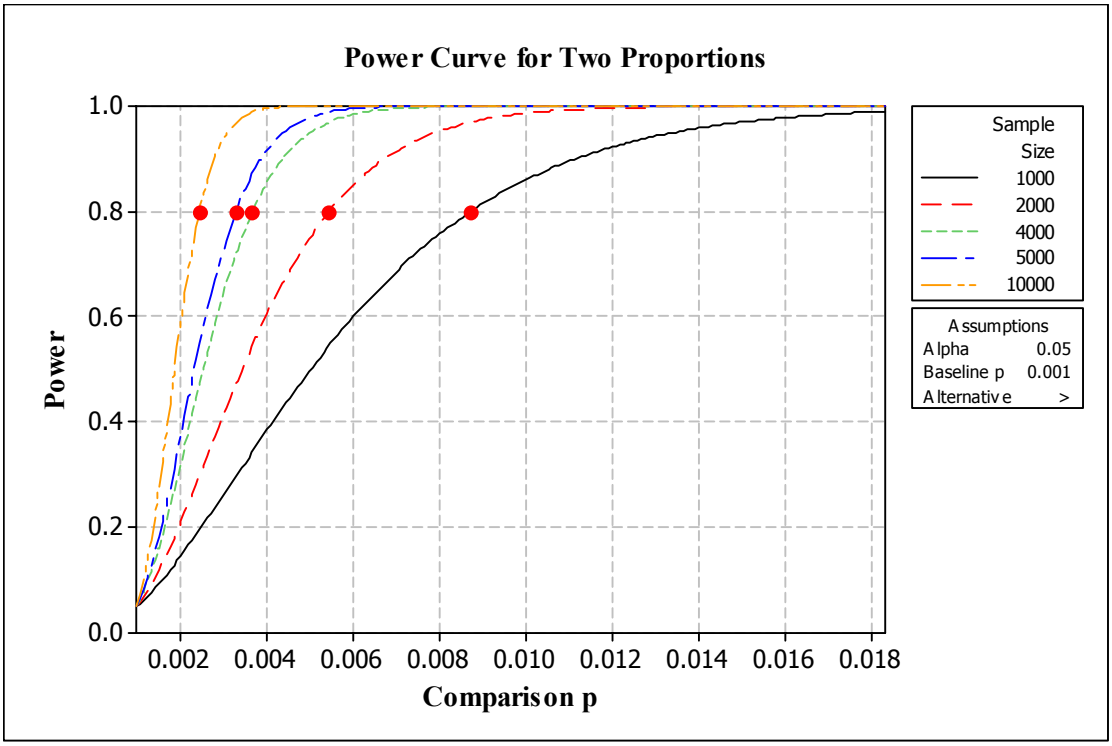


Figure 20. Power curves for pair-wise test for different numbers of cells (1000 to 10000) with an incidence of 0.2% (proportion p of 0.002 or 2/1000) micronuclei in the negative control culture

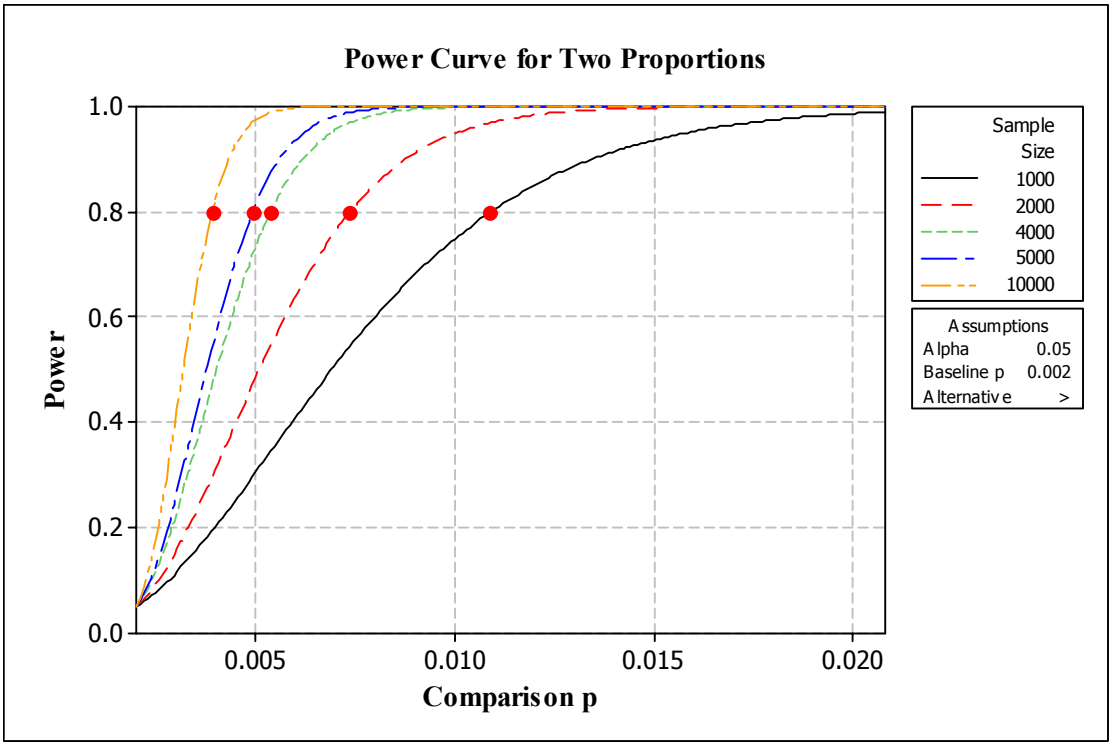


Figure 21. Power curves for pair-wise test for different numbers of cells (1000 to 10000) with an incidence of 0.5% (proportion p of 0.005 or 5/1000) micronuclei in the negative control culture

Figure 22. Plot of the percentage of zero counts for Poisson distributed data with increasing population means

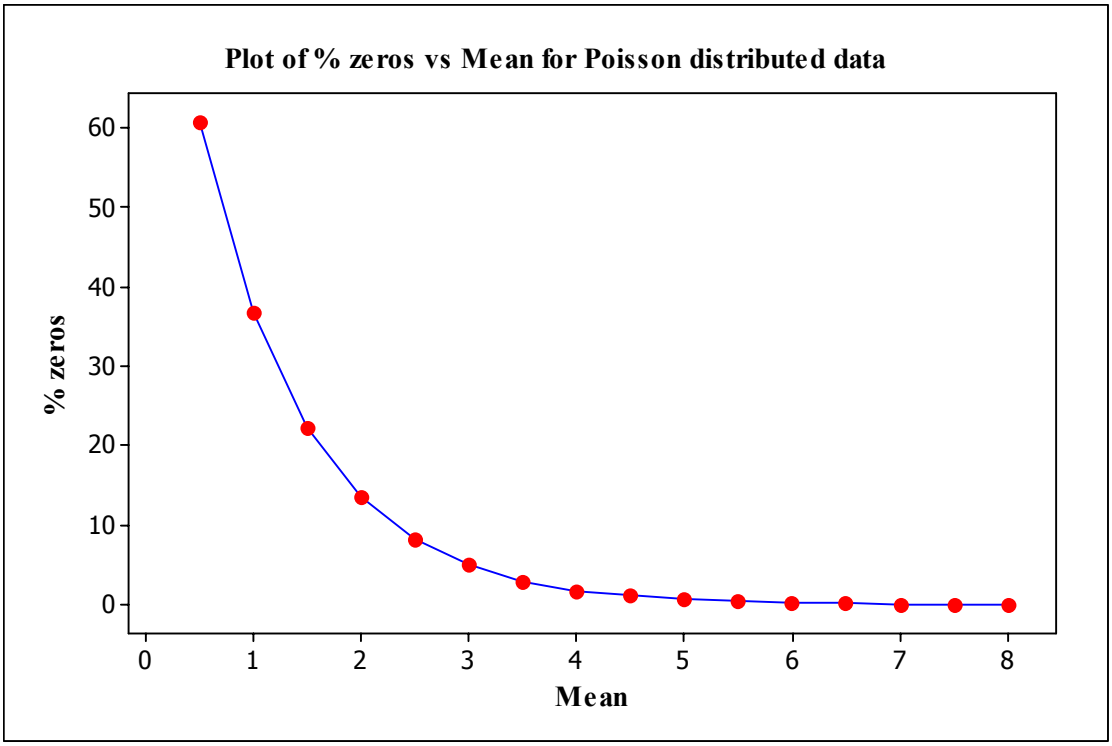


Figure 23. Poisson distribution for mean counts (N) from 1 to 10. Figure shows a graphical representation of the data in Table 10.

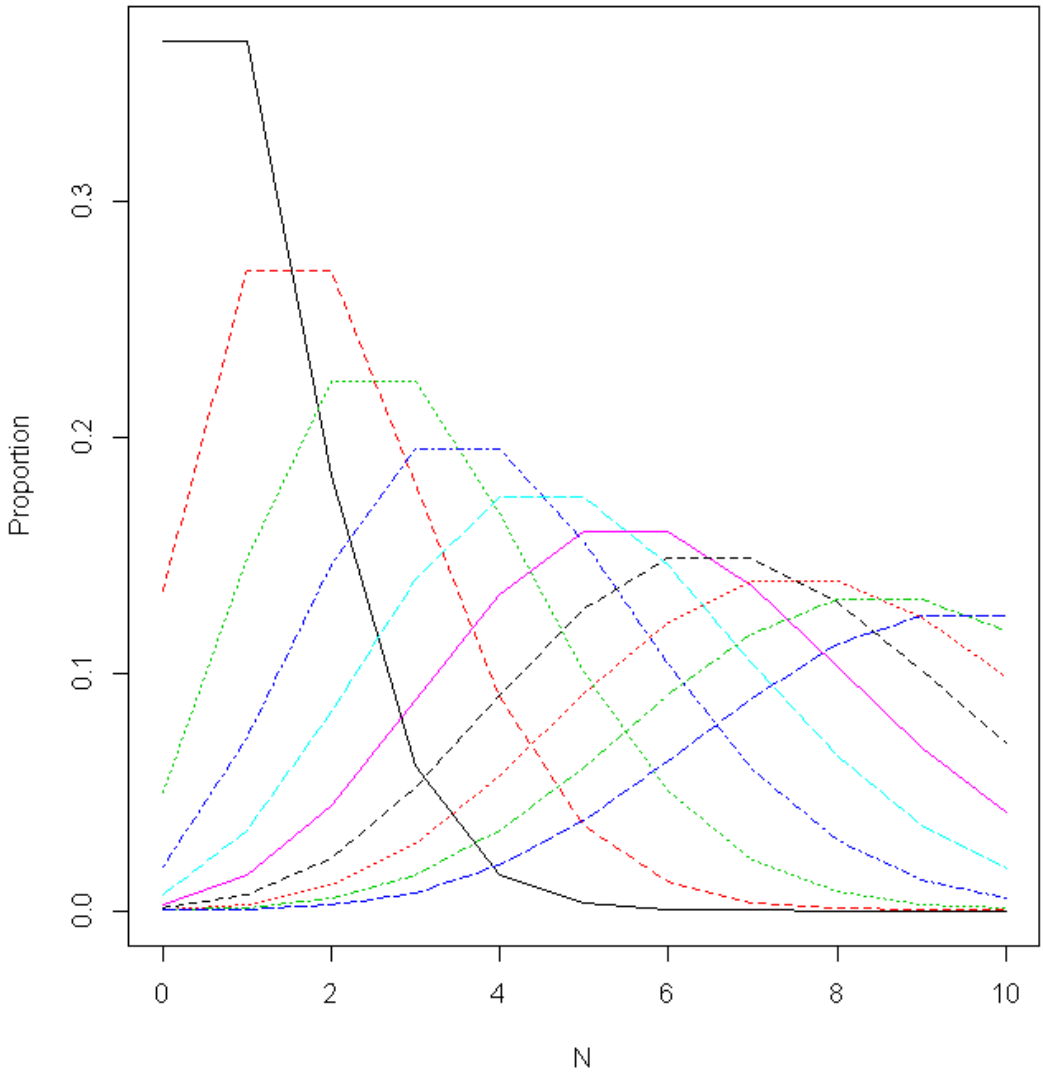


Figure 24. An illustration of how the Poisson distributions approximates to a normal distribution as the mean count (N) increases from 1 to 25.

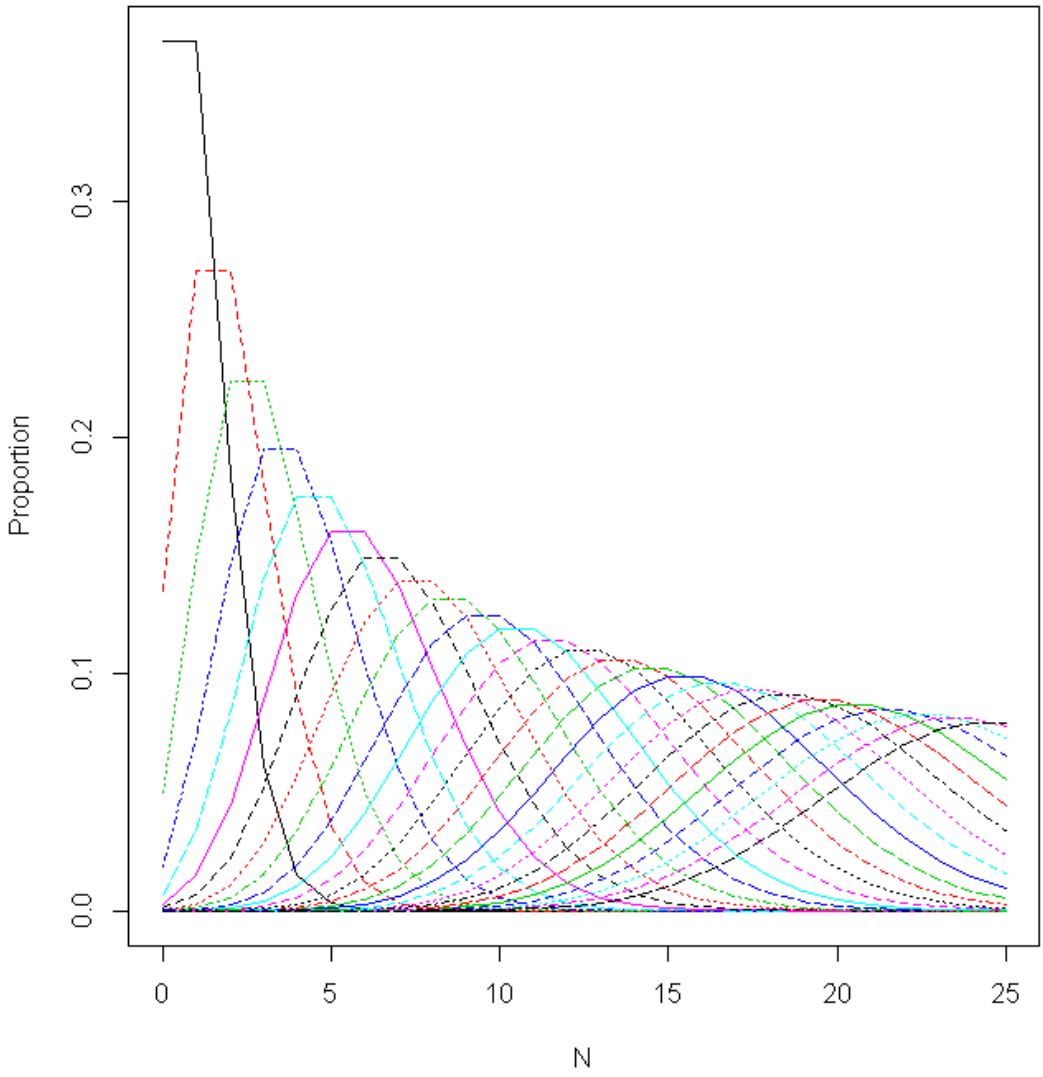
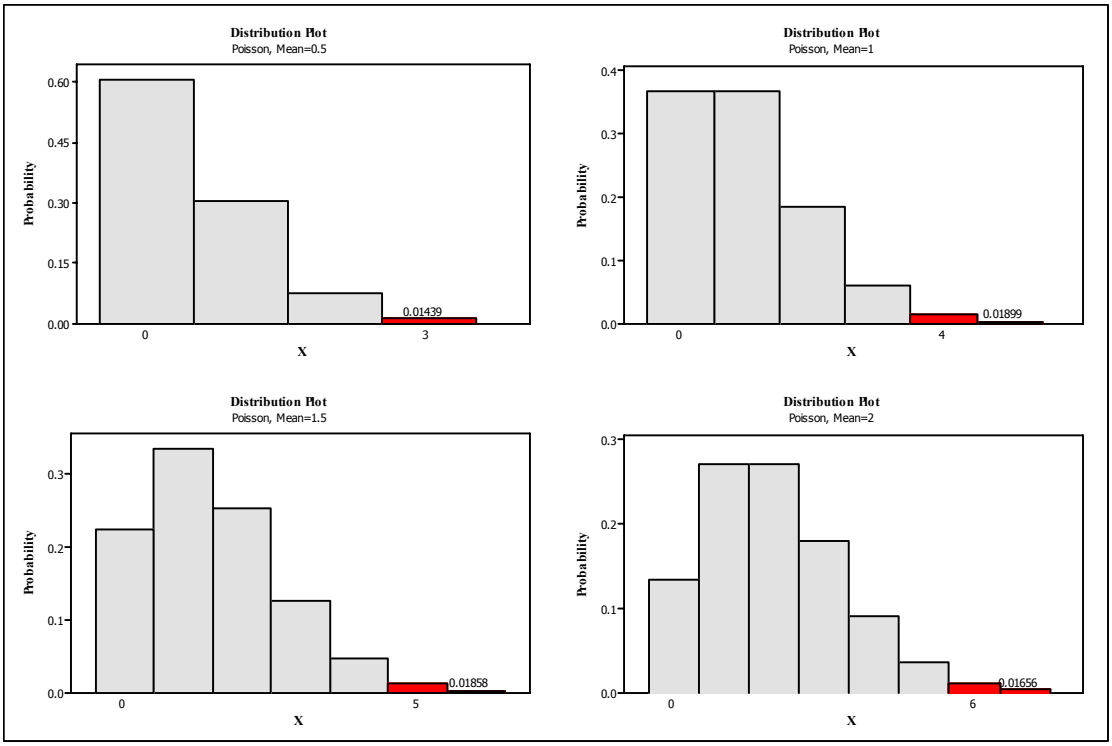


Figure 25. Counts beyond the upper 5% (0.05) of the distribution (in red) for different means based upon Poisson distributions with population means from 0.5 to 5.0



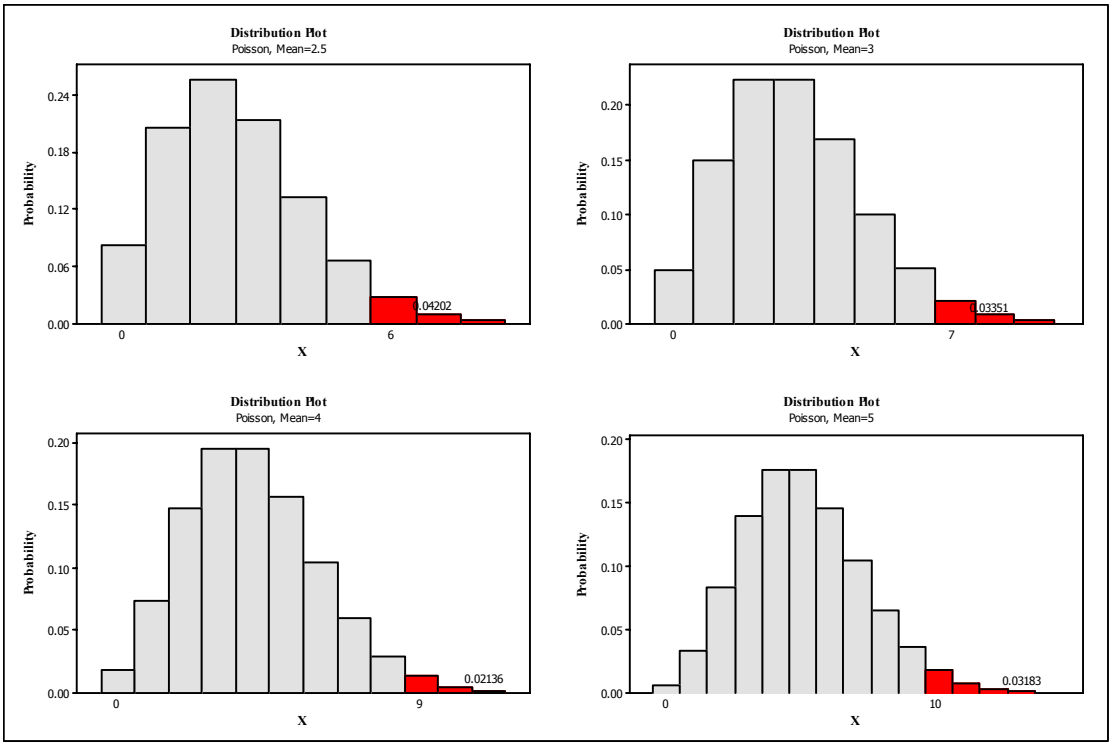
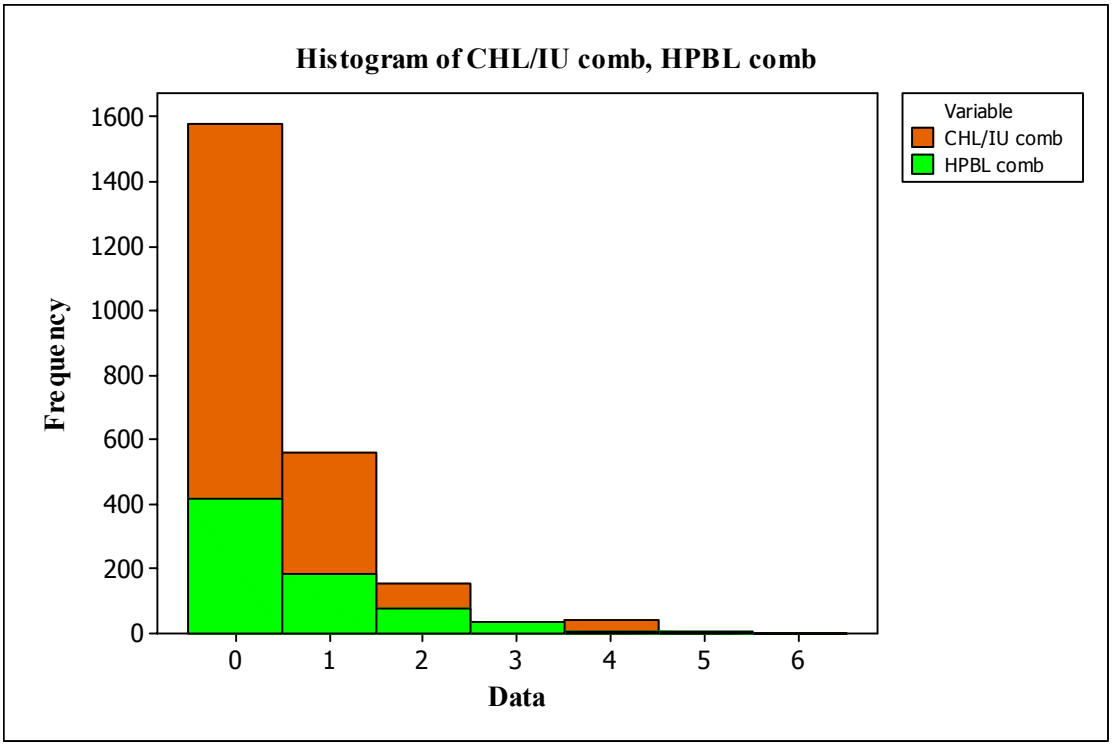
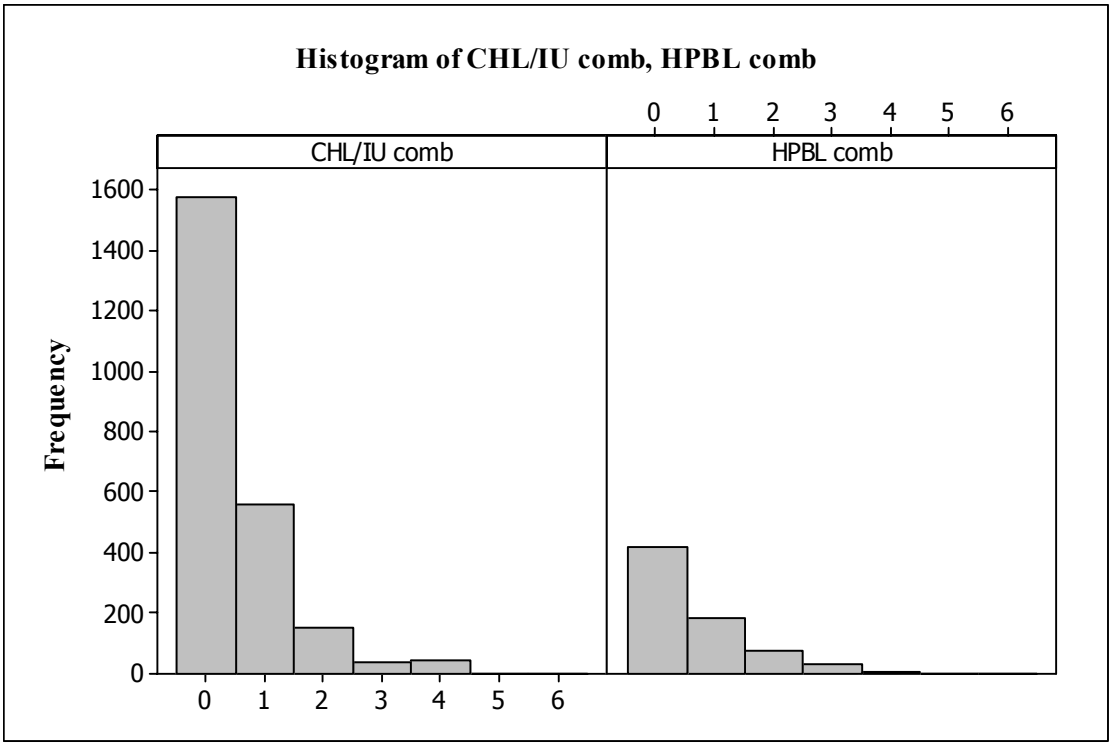


Figure 26. *In vitro* chromosomal aberration test: combining counts of HPBL and CHL/IU across laboratories (100 cells)





Poisson mean for CHL/IU comb = 0.485 Chi-sq (3df) 440.9 P<0.001
Poisson mean for HPBL comb = 0.665 Chi-Sq(3df) 44.4 P<0.001
(Mean and Standard deviation can be used to calculate GEF)

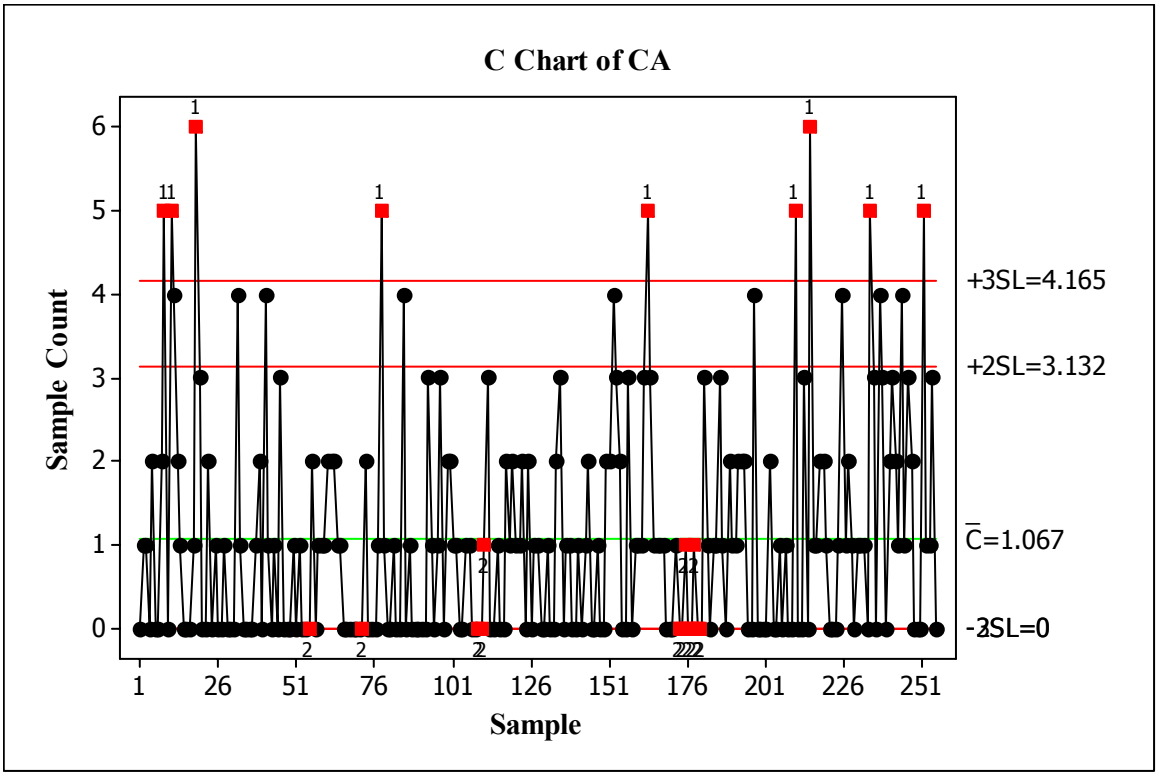
Variable	N	Mean	StDev	Min	Max	GEF
CHL/IU comb	2373	0.48	0.83	0	4	1.31
HPBL comb	722	0.66	0.96	0	6	1.62

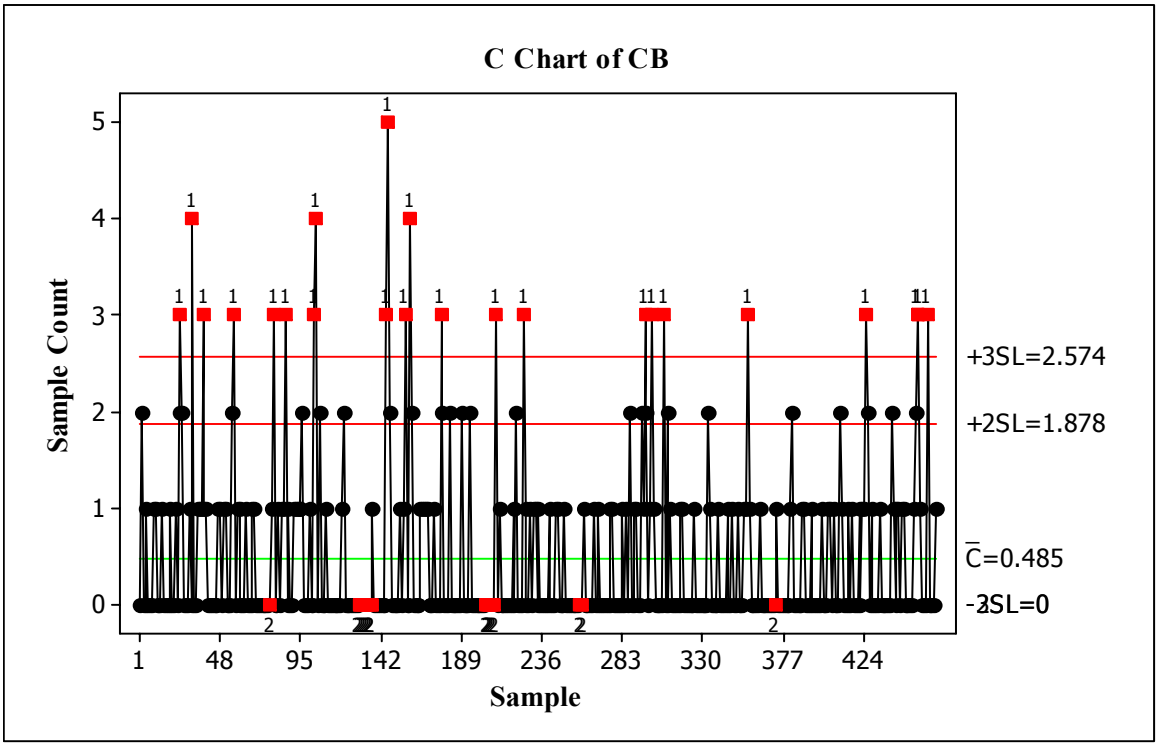
Appendix 1: Quality control C charts of chromosomal aberration counts across all cultures from laboratories where data can be expressed as counts.

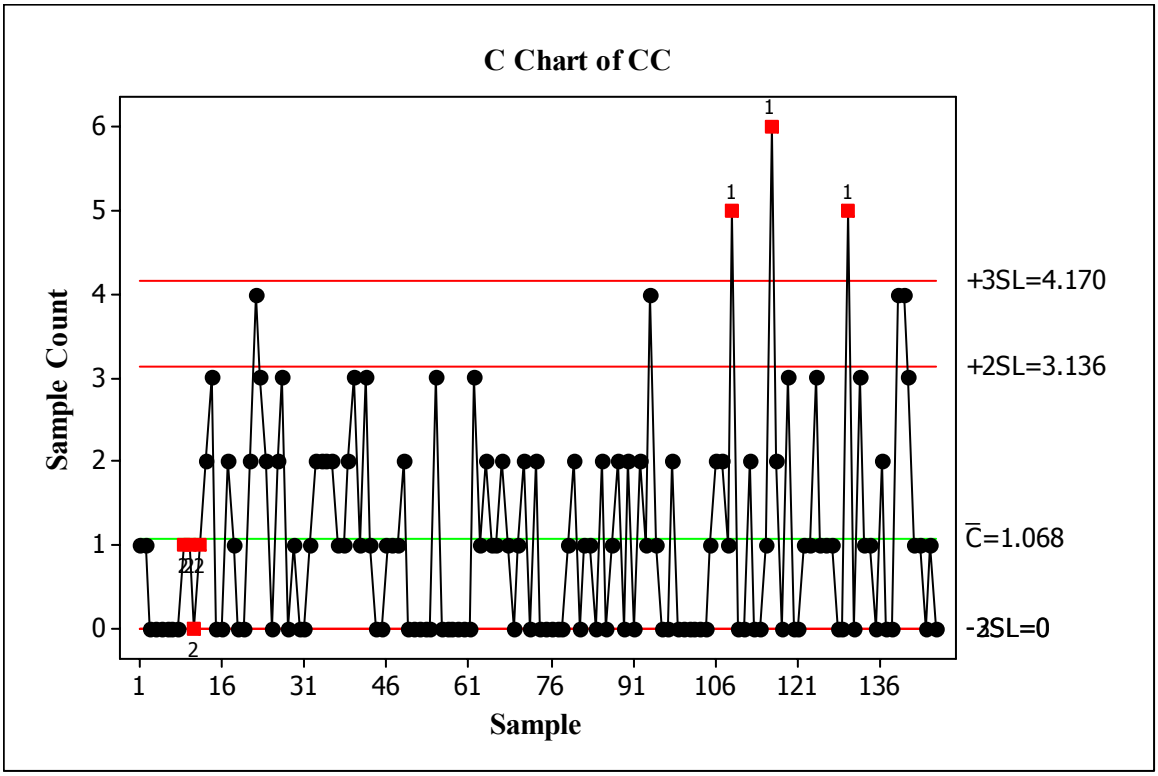
The upper and lower control limits are based upon 2 and 3 SD based upon the Poisson distribution.

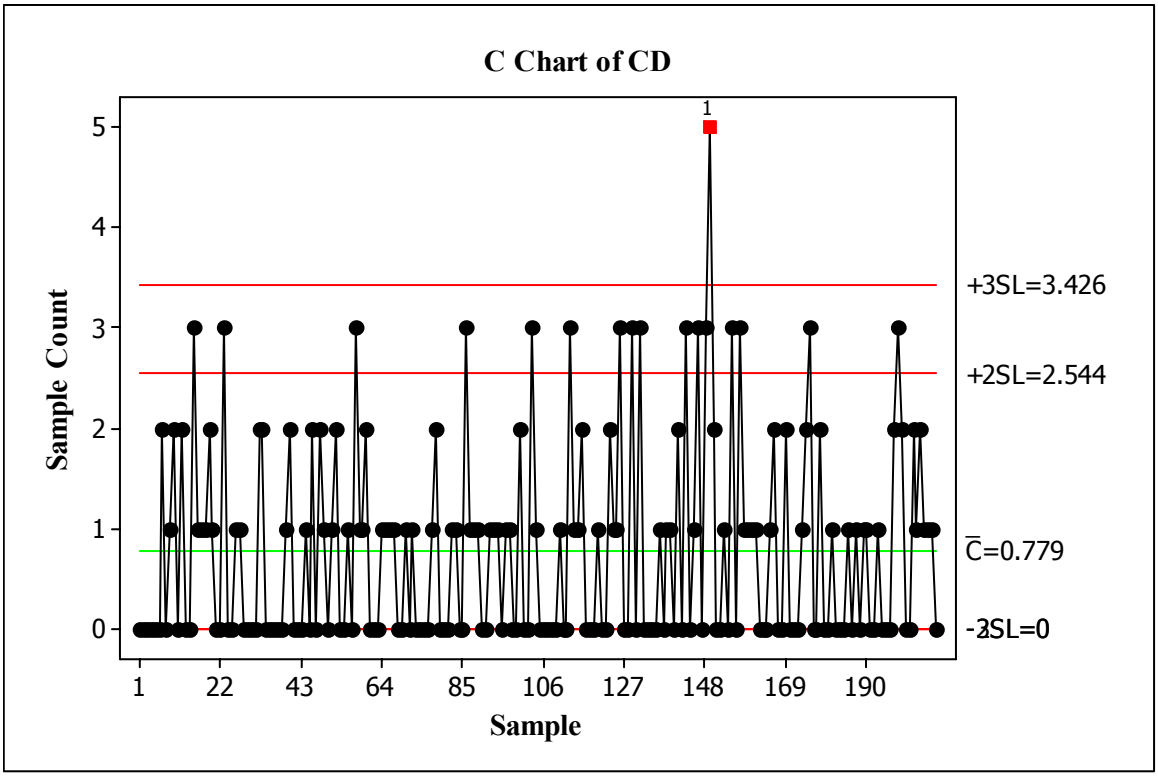
The red points on graphs represent cases 'flagged' as outside controls based upon rules 1-4 defined as:

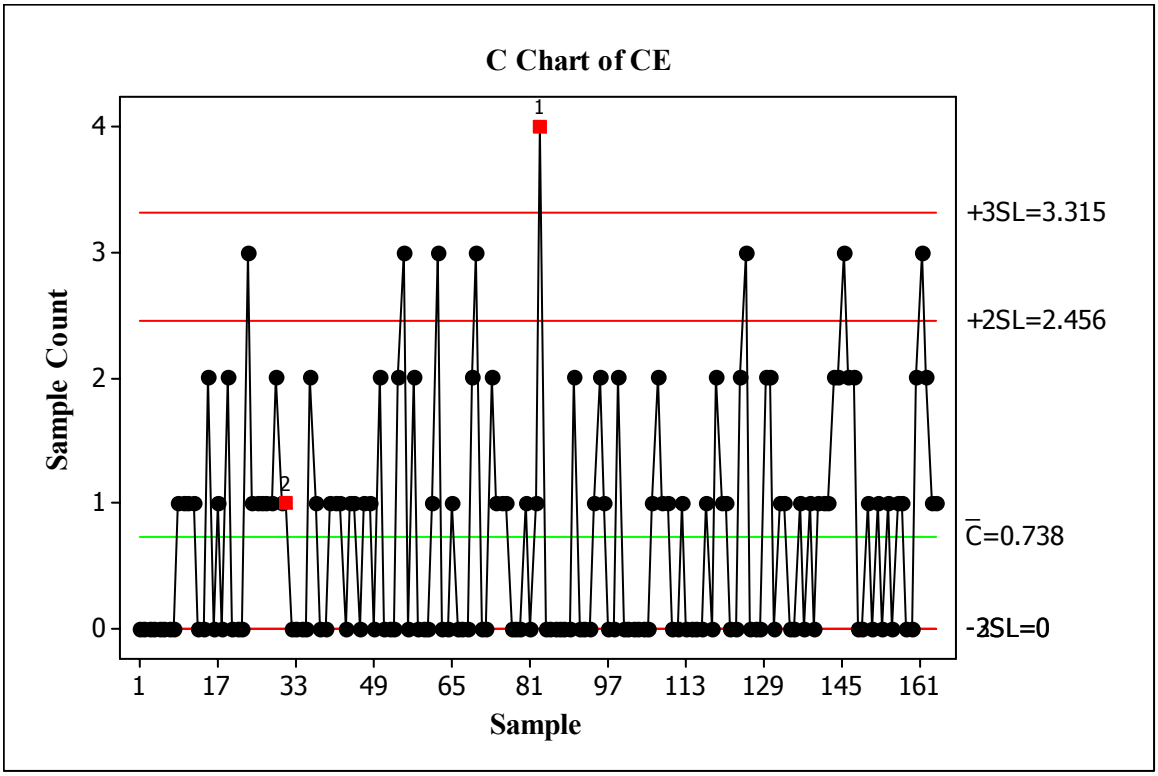
1) 1 point more than 3 standard deviations from the centre line
2) 9 points in a row on same side of centre line
3) 6 points in a row, all increasing or all decreasing
4) 14 points in a row, alternating up and down

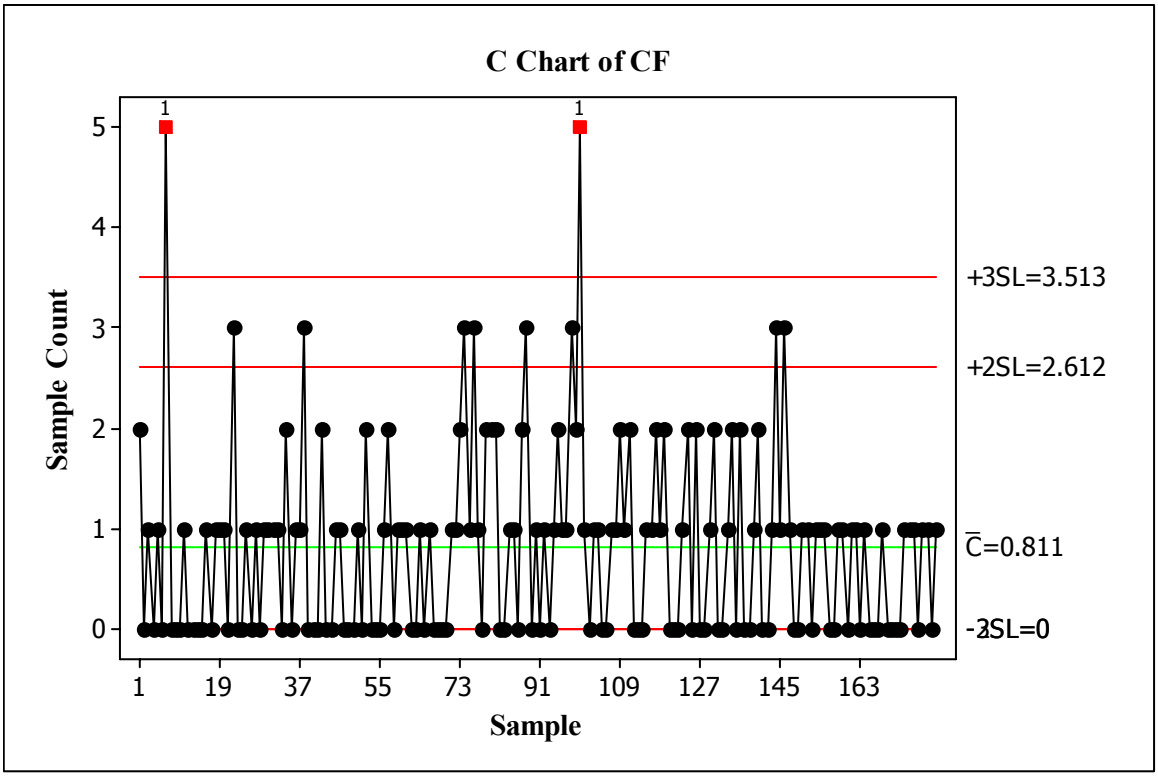

C Chart of CA
Sample Count
Sample
+3SL=4.165
+2SL=3.132
C=1.067
-3SL=0

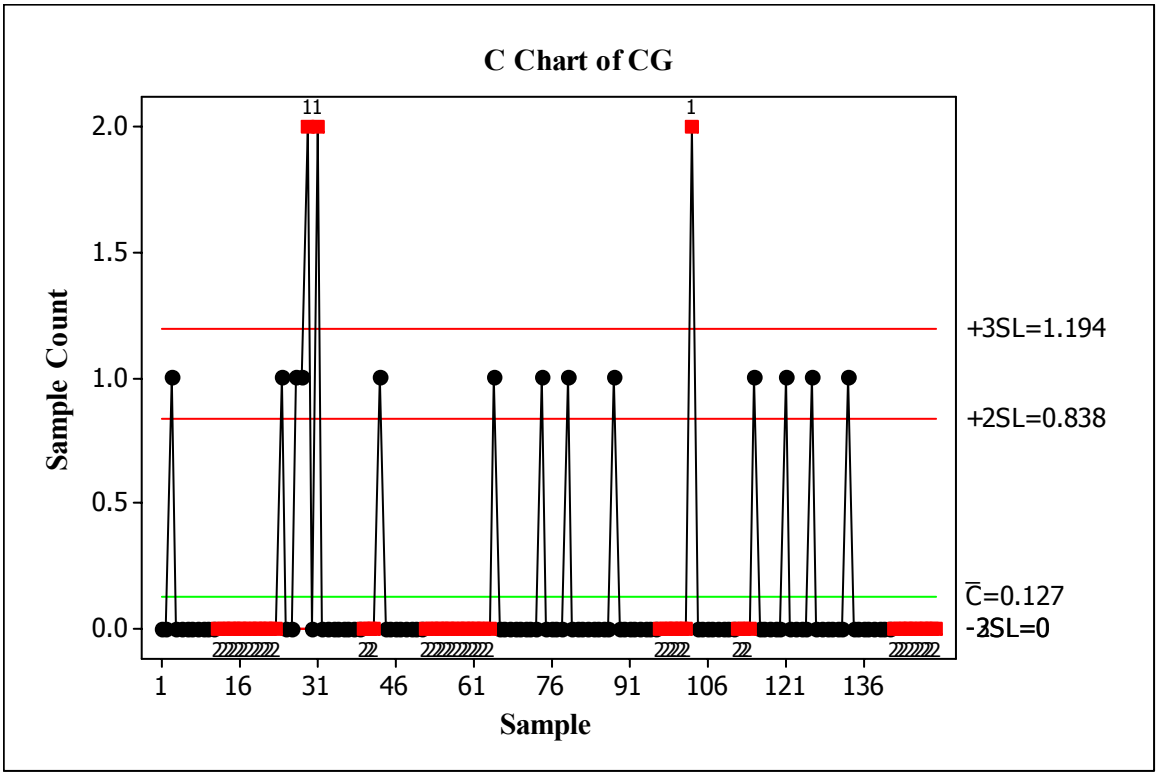

C Chart of CB
Sample Count
Sample
+3SL=2.574
+2SL=1.878
C=0.485
-3SL=0

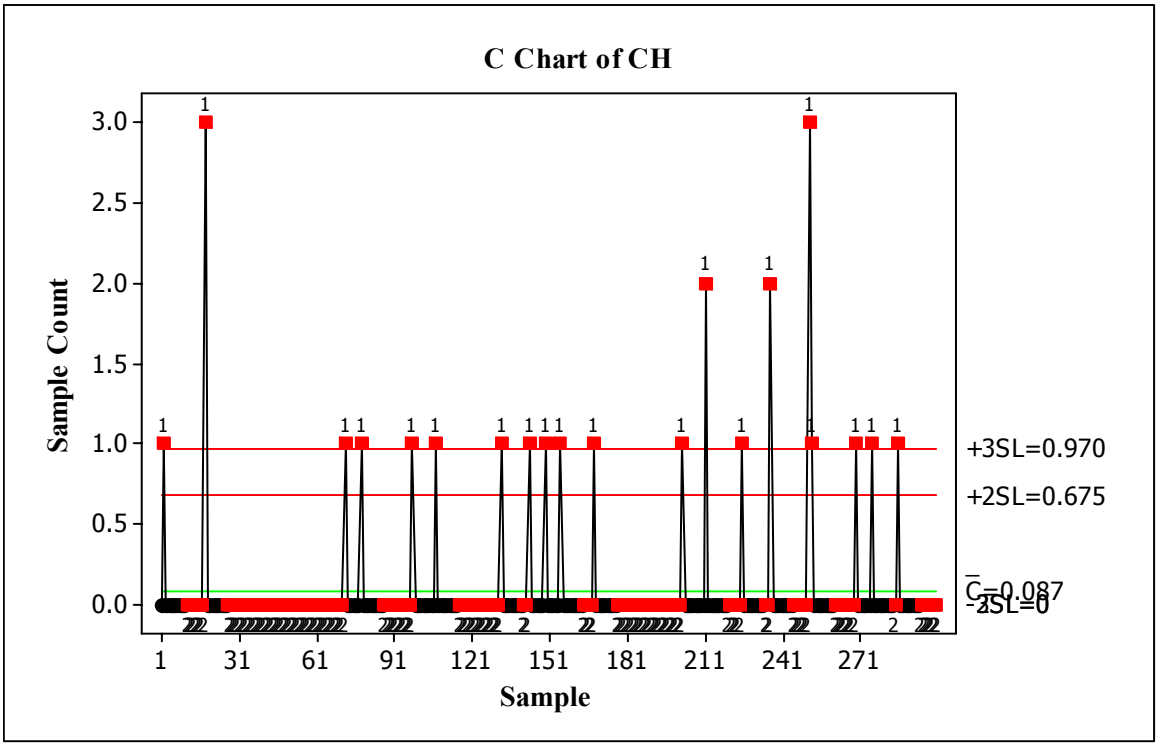

C Chart of CC
Sample Count
Sample
+3SL=4.170
+2SL=3.136
C=1.068
-2SL=0
-3SL=0

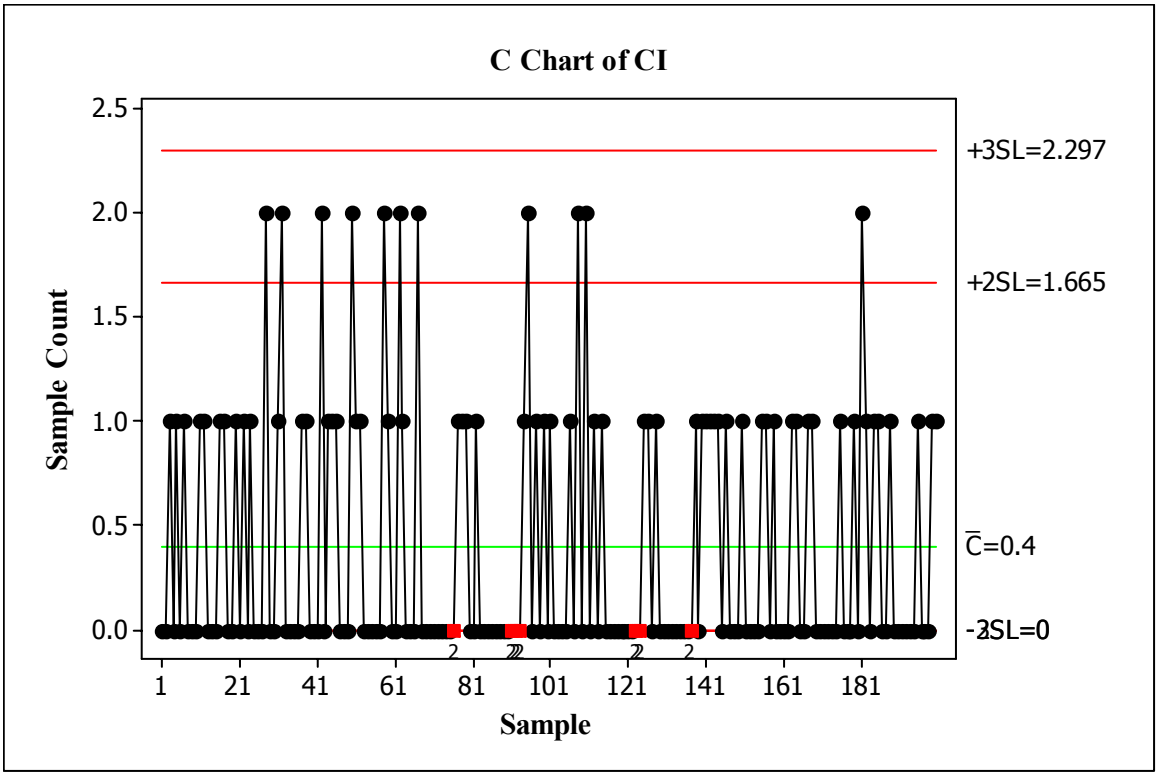

C Chart of CD
Sample Count
Sample
+3SL=3.426
+2SL=2.544
C=0.779
-2SL=0
-3SL=0

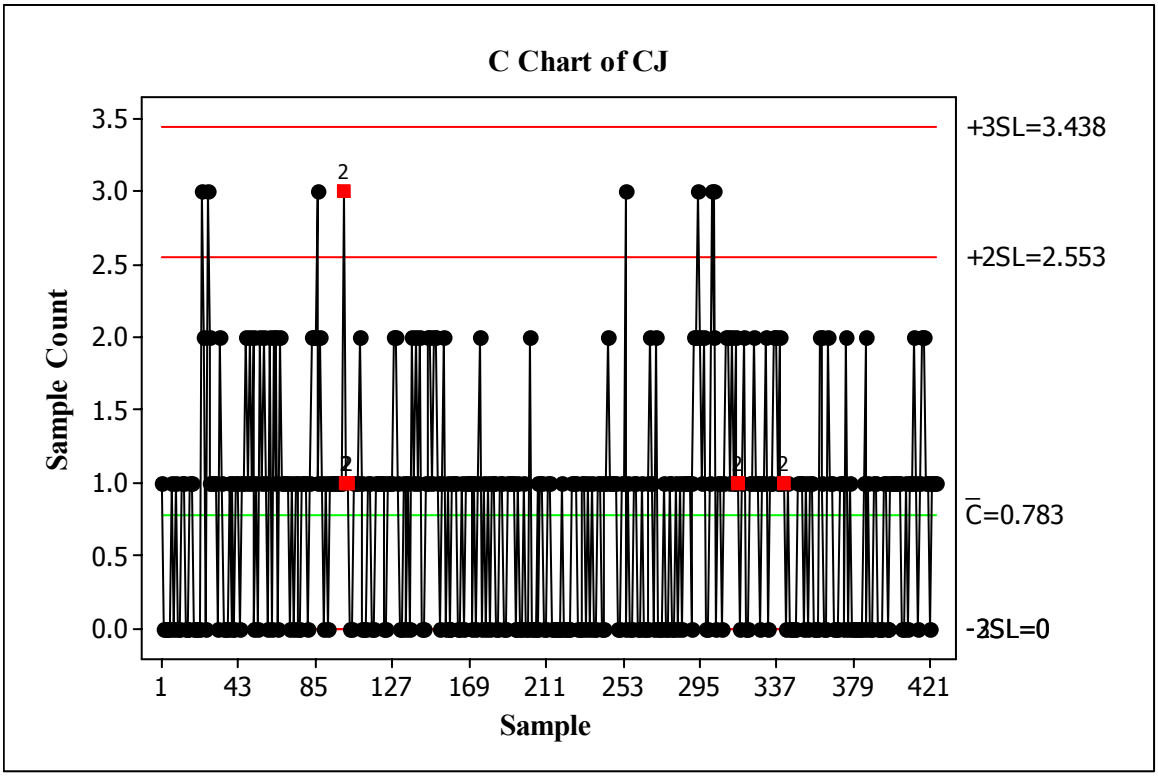

C Chart of CE
Sample Count
Sample
+3SL=3.315
+2SL=2.456
C=0.738
-2SL=0
1
2
0
1
2
3
4
1
17
33
49
65
81
97
113
129
145
161

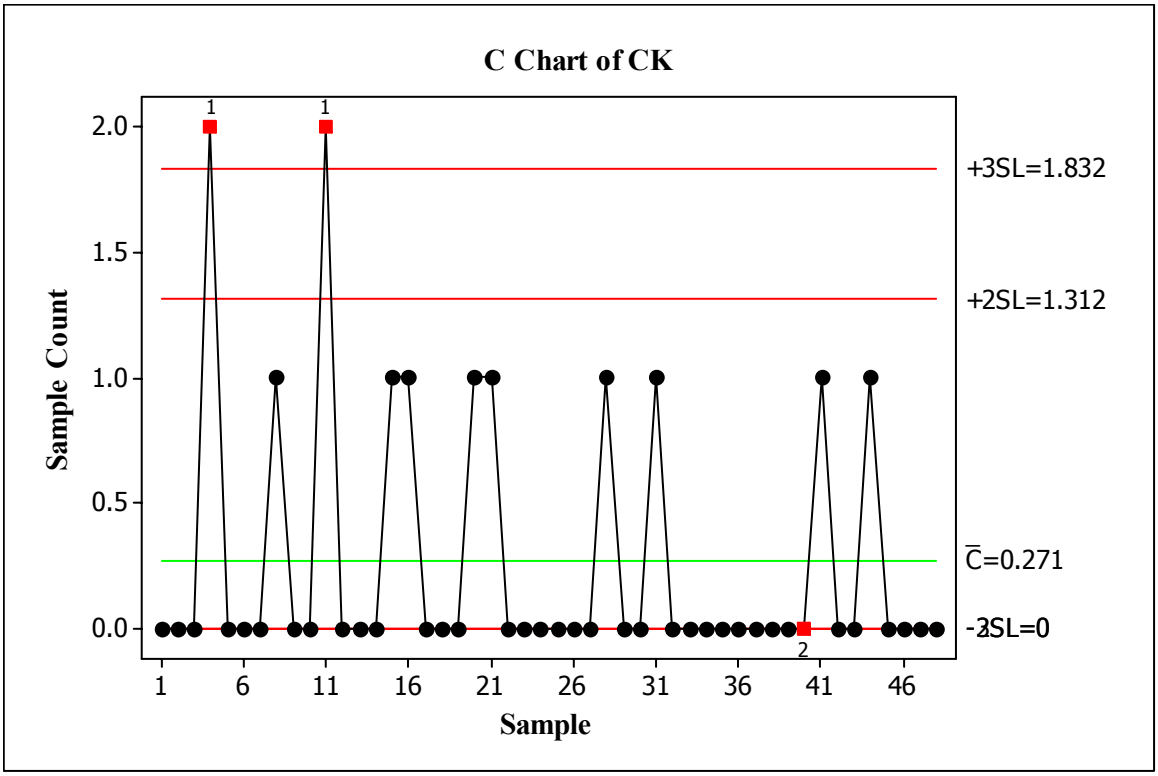

C Chart of CF
Sample Count
Sample
+3SL=3.513
+2SL=2.612
C=0.811
-2SL=0
1
1
0
1
2
3
4
5
1
19
37
55
73
91
109
127
145
163

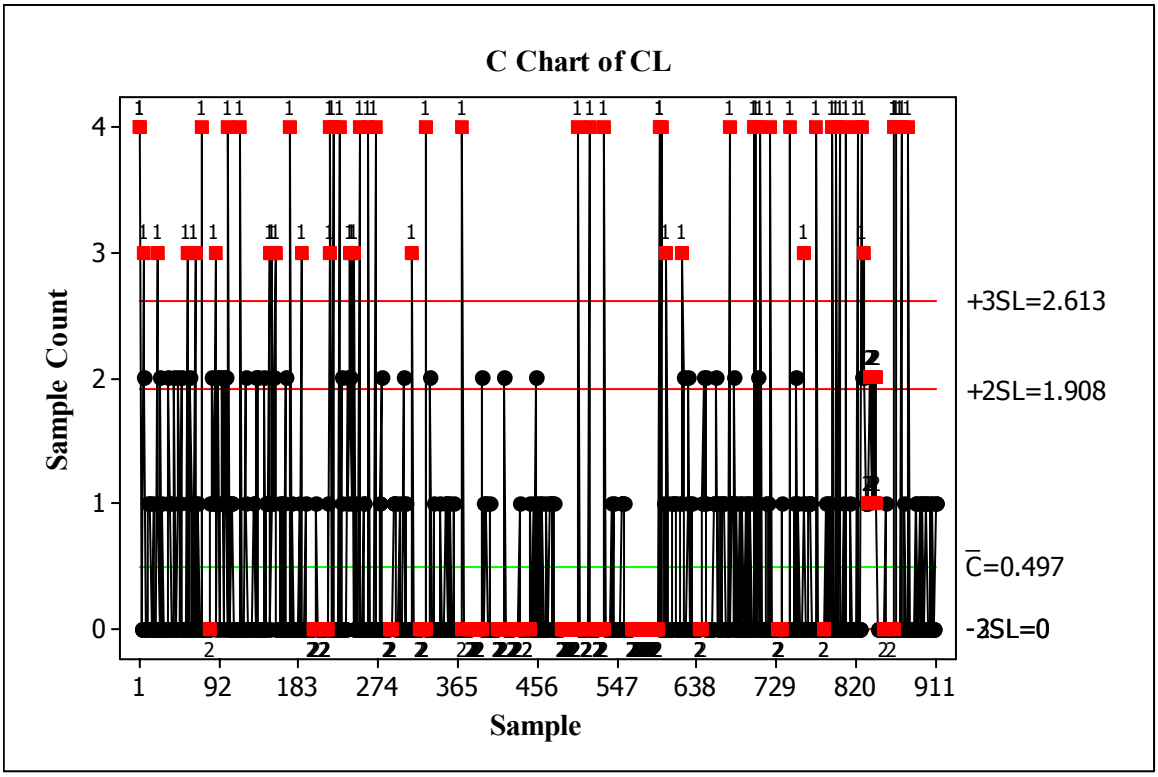

C Chart of CG
Sample Count
Sample
+3SL=1.194
+2SL=0.838
C=0.127
-2SL=0

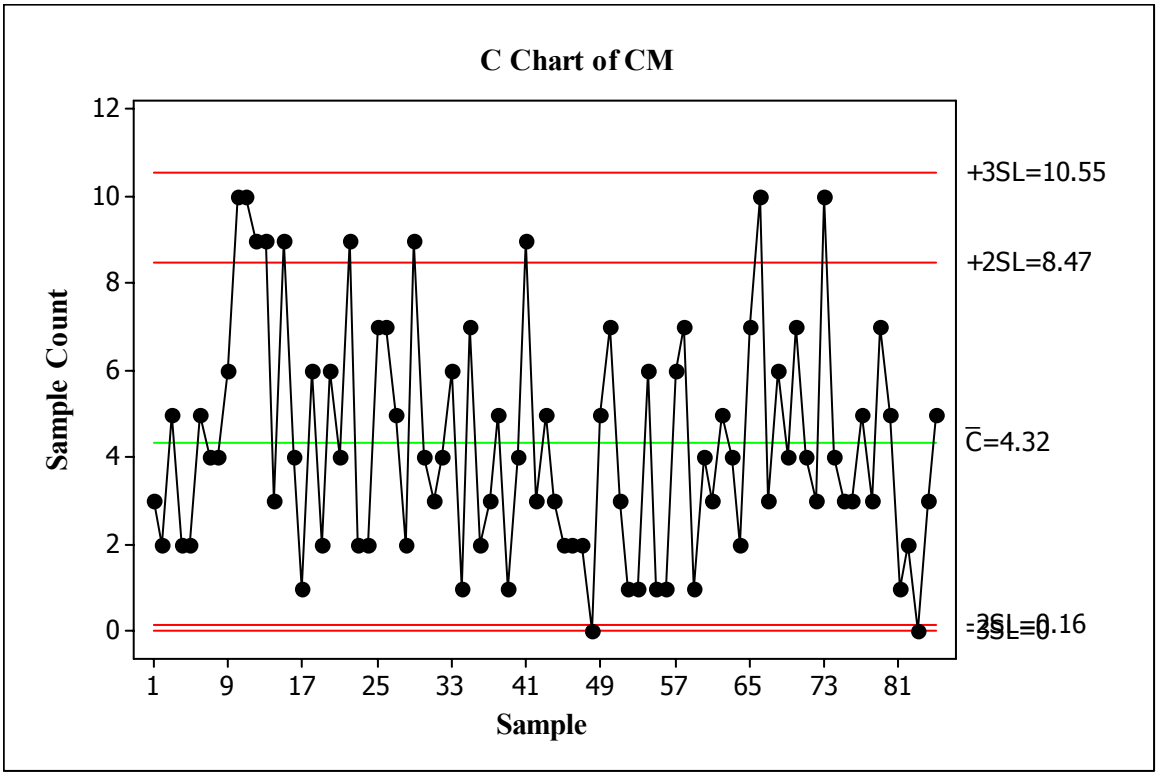

C Chart of CH
Sample Count
Sample
+3SL=0.970
+2SL=0.675
C=0.087
-2SL=0

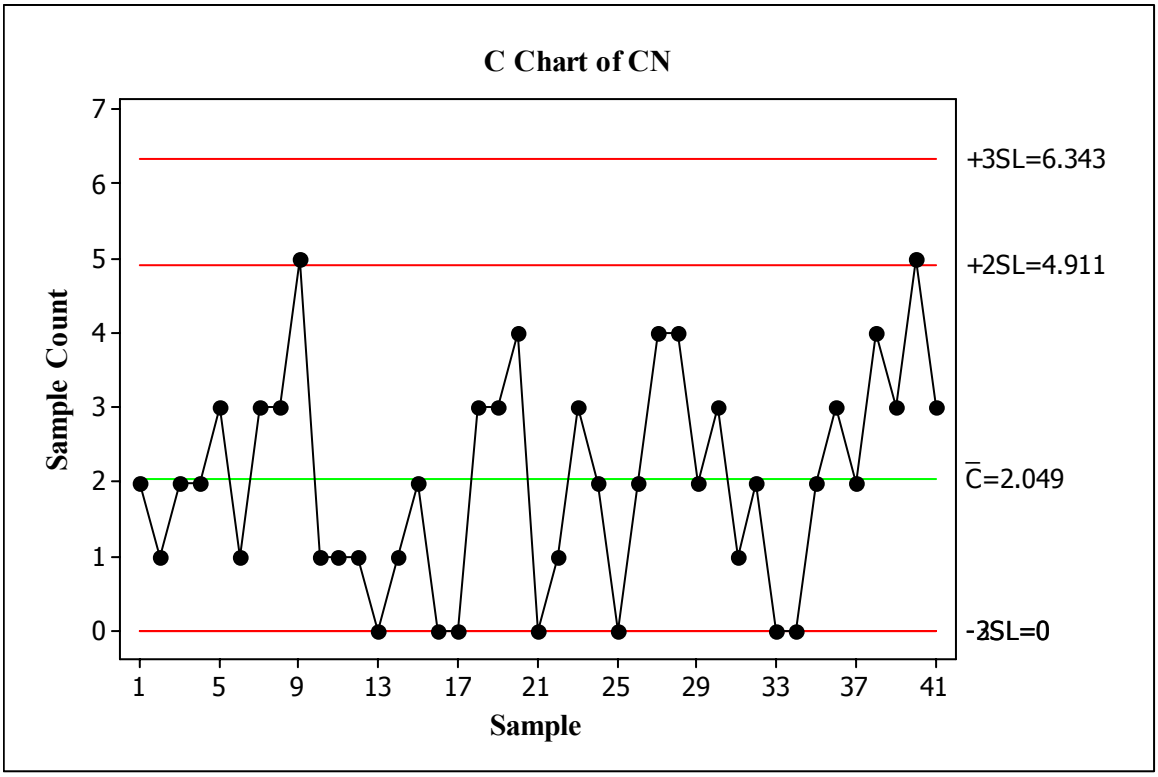

C Chart of CI
Sample Count
Sample
+3SL=2.297
+2SL=1.665
C=0.4
-3SL=0
2.5
2.0
1.5
1.0
0.5
0.0
1
21
41
61
81
101
121
141
161
181

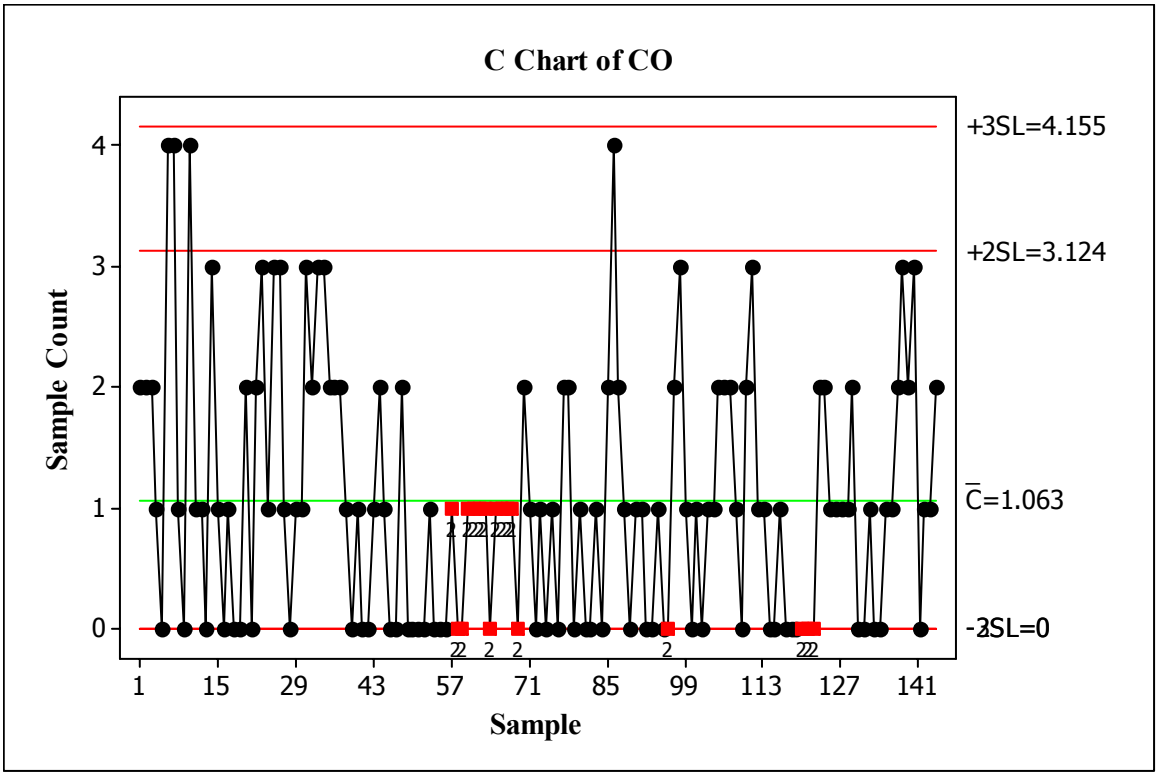

C Chart of CJ
Sample Count
Sample
+3SL=3.438
+2SL=2.553
C=0.783
-3SL=0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
43
85
127
169
211
253
295
337
379
421

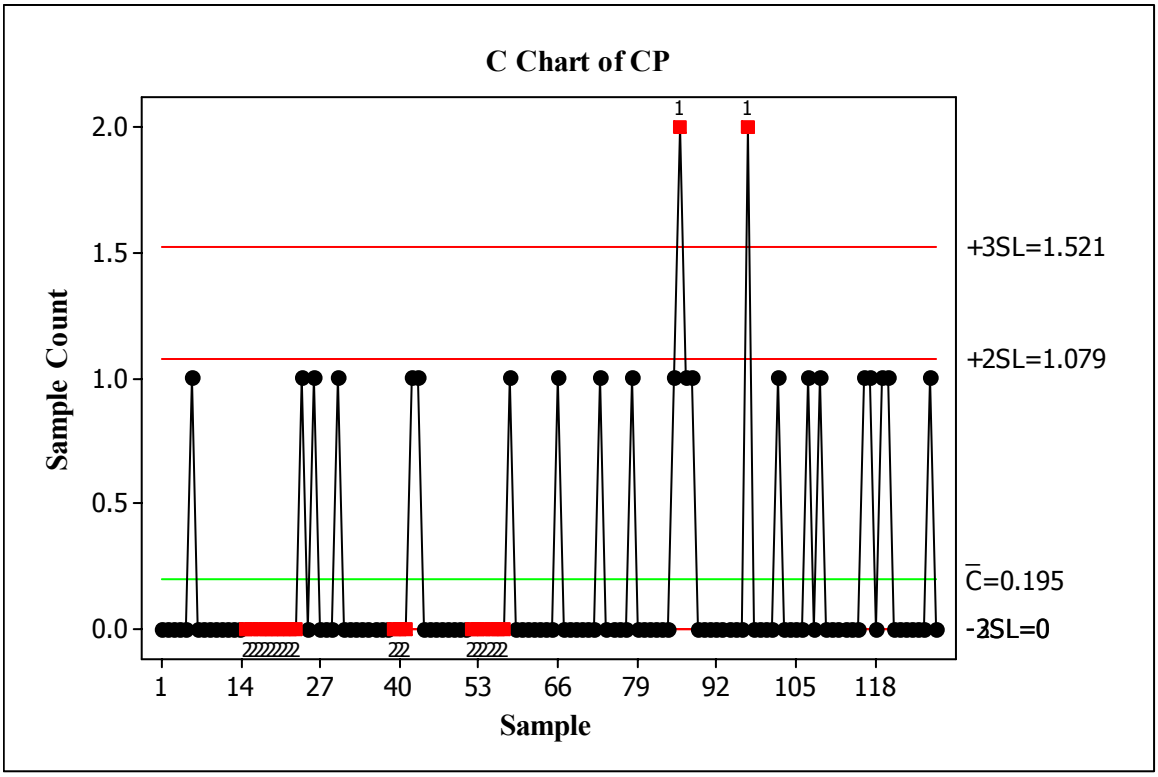

C Chart of CK
Sample Count
Sample
+3SL=1.832
+2SL=1.312
C=0.271
-2SL=0
2.0
1.5
1.0
0.5
0.0
1
6
11
16
21
26
31
36
41
46

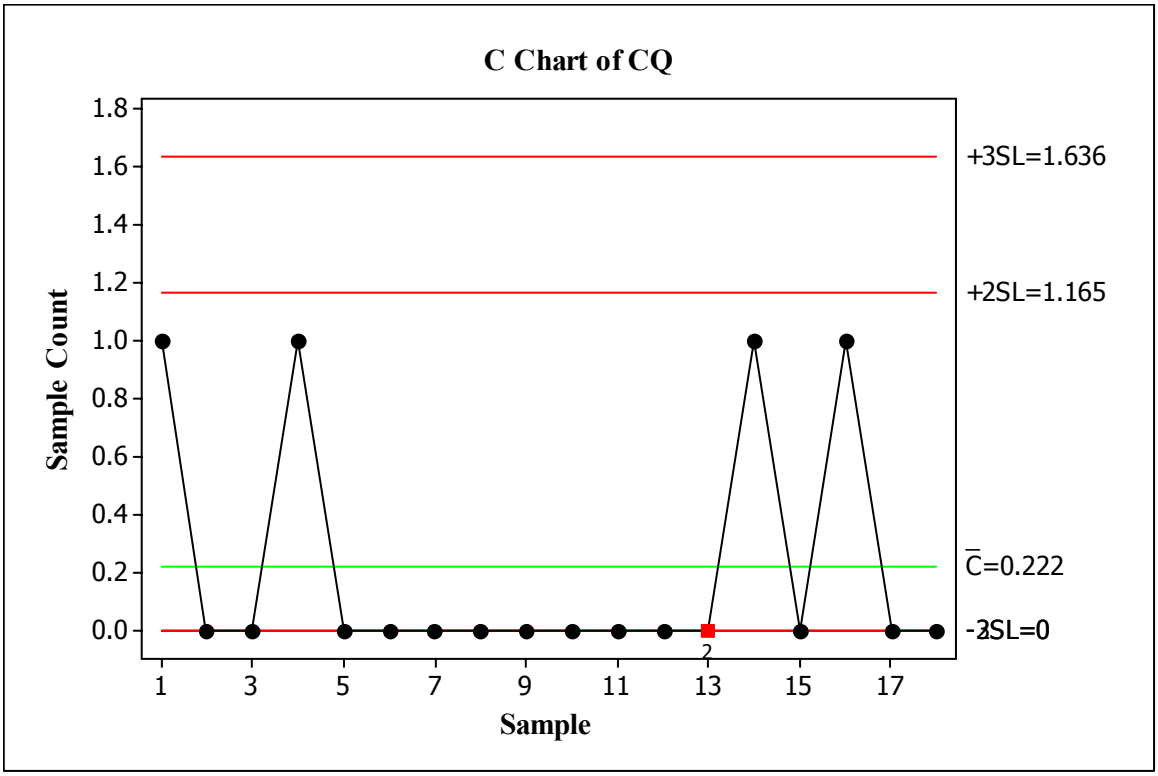

C Chart of CL
Sample Count
Sample
+3SL=2.613
+2SL=1.908
C=0.497
-2SL=0
4
3
2
1
0
1
92
183
274
365
456
547
638
729
820
911

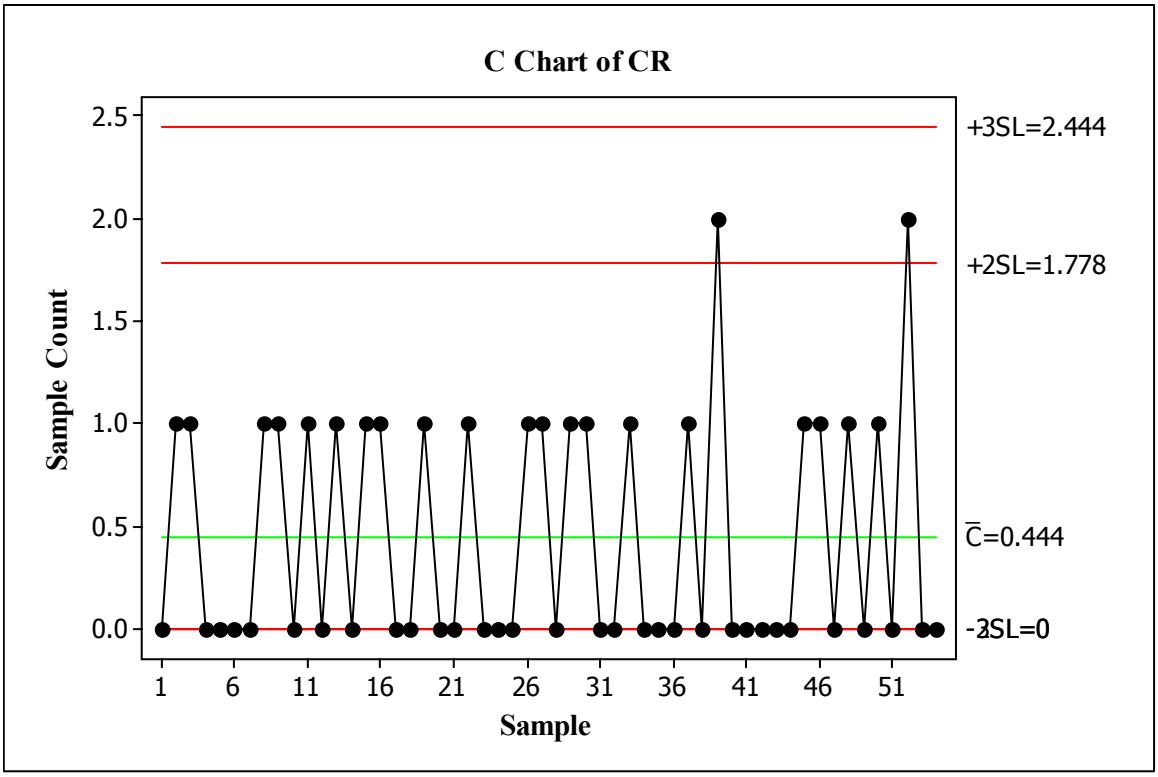

C Chart of CM
Sample Count
Sample
+3SL=10.55
+2SL=8.47
C=4.32
-2SL=0.16
-3SL=0
12
10
8
6
4
2
0
1
9
17
25
33
41
49
57
65
73
81

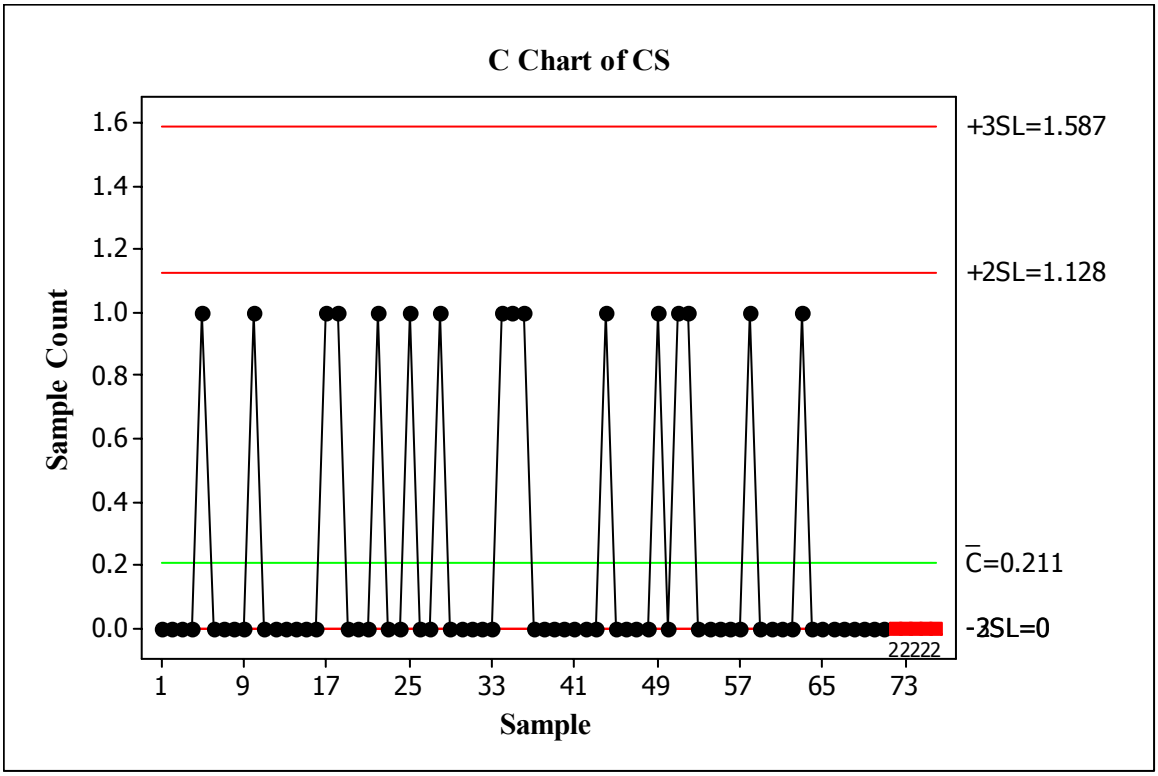

C Chart of CN
Sample Count
Sample
+3SL=6.343
+2SL=4.911
C=2.049
-2SL=0
-3SL=0
7
6
5
4
3
2
1
0
1
5
9
13
17
21
25
29
33
37
41

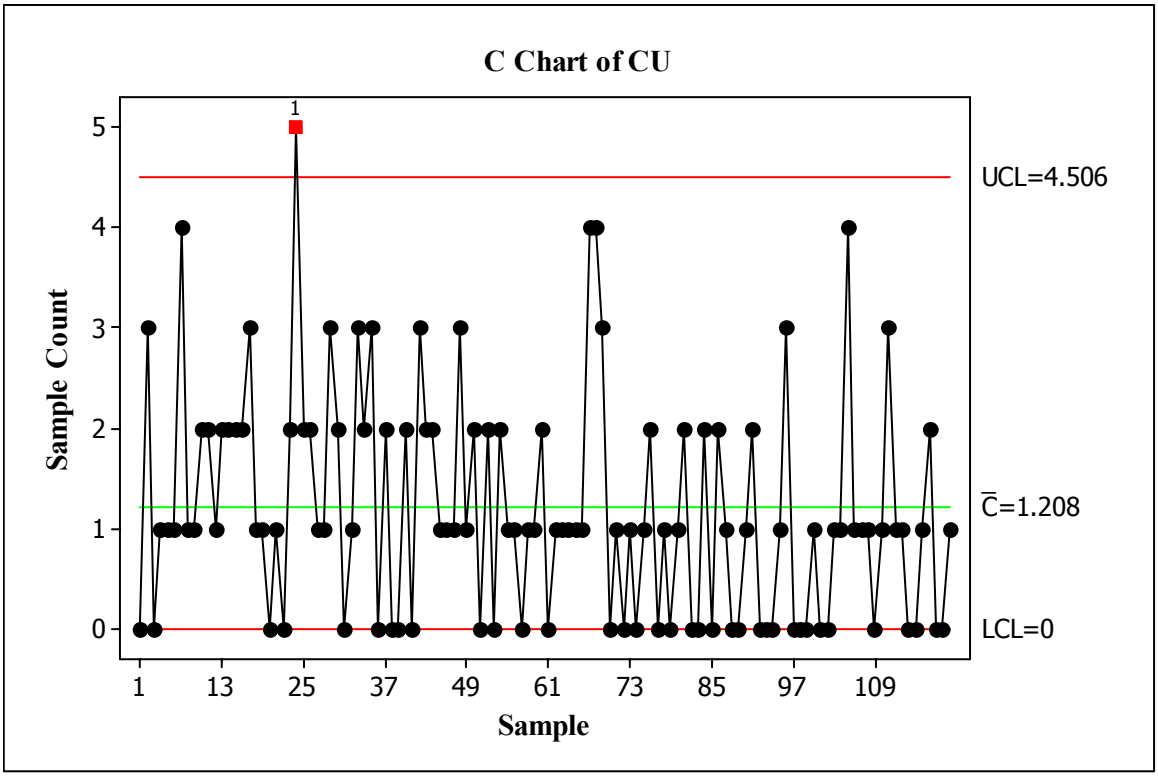

C Chart of CO
Sample Count
Sample
+3SL=4.155
+2SL=3.124
C=1.063
-2SL=0
-3SL=0
0
1
2
3
4
1
15
29
43
57
71
85
99
113
127
141


C Chart of CP
Sample Count
Sample
+3SL=1.521
+2SL=1.079
C=0.195
-2SL=0
-3SL=0
0.0
0.5
1.0
1.5
2.0
1
14
27
40
53
66
79
92
105
118


C Chart of CQ
Sample Count
Sample
+3SL=1.636
+2SL=1.165
C=0.222
-3SL=0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1
3
5
7
9
11
13
15
17


C Chart of CR
Sample Count
Sample
+3SL=2.444
+2SL=1.778
C=0.444
-3SL=0
2.5
2.0
1.5
1.0
0.5
0.0
1
6
11
16
21
26
31
36
41
46
51


C Chart of CS
Sample Count
Sample
+3SL=1.587
+2SL=1.128
C=0.211
-2SL=0
-3SL=0
22222
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1
9
17
25
33
41
49
57
65
73

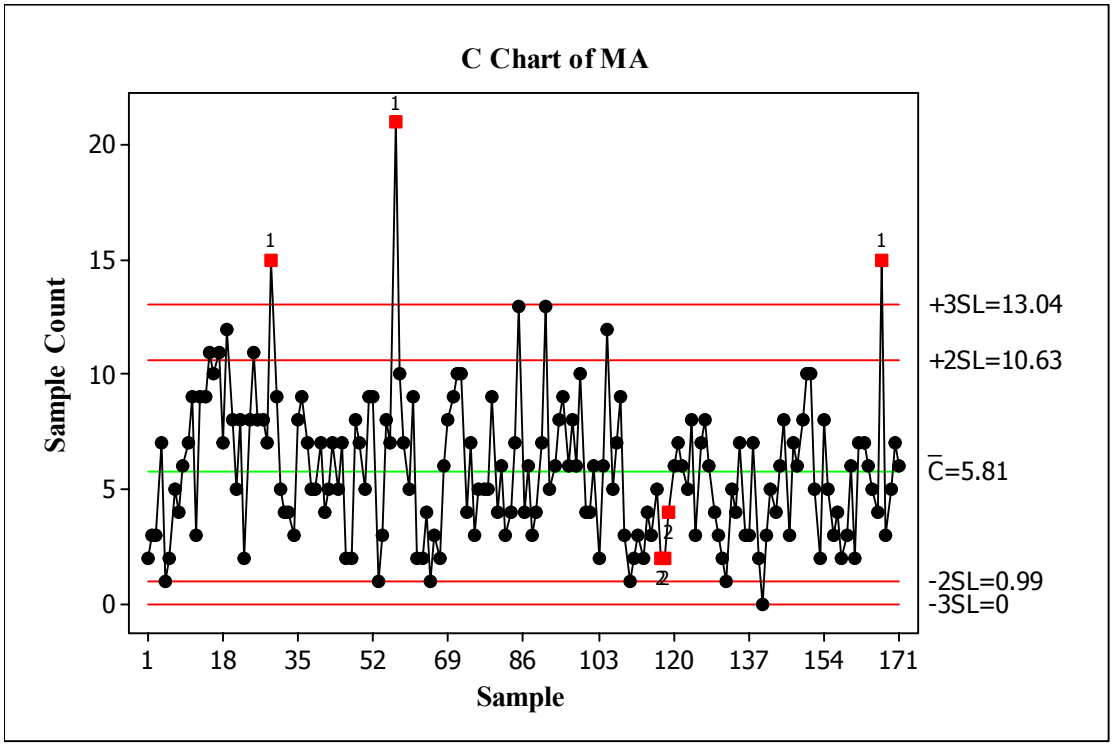

C Chart of CU
Sample Count
Sample
UCL=4.506
C=1.208
LCL=0
1
5
4
3
2
1
0
1
13
25
37
49
61
73
85
97
109

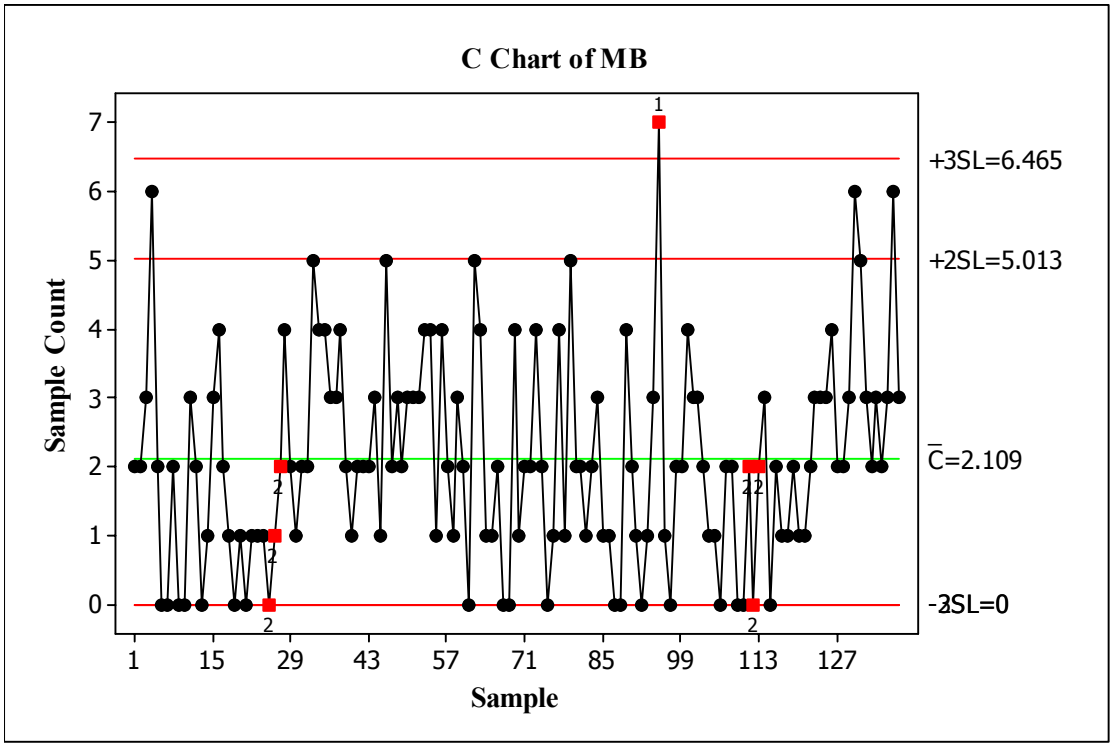
Appendix 2: Quality control 'C' charts of micronuclei counts across all cultures from laboratories where data can be expressed as counts.

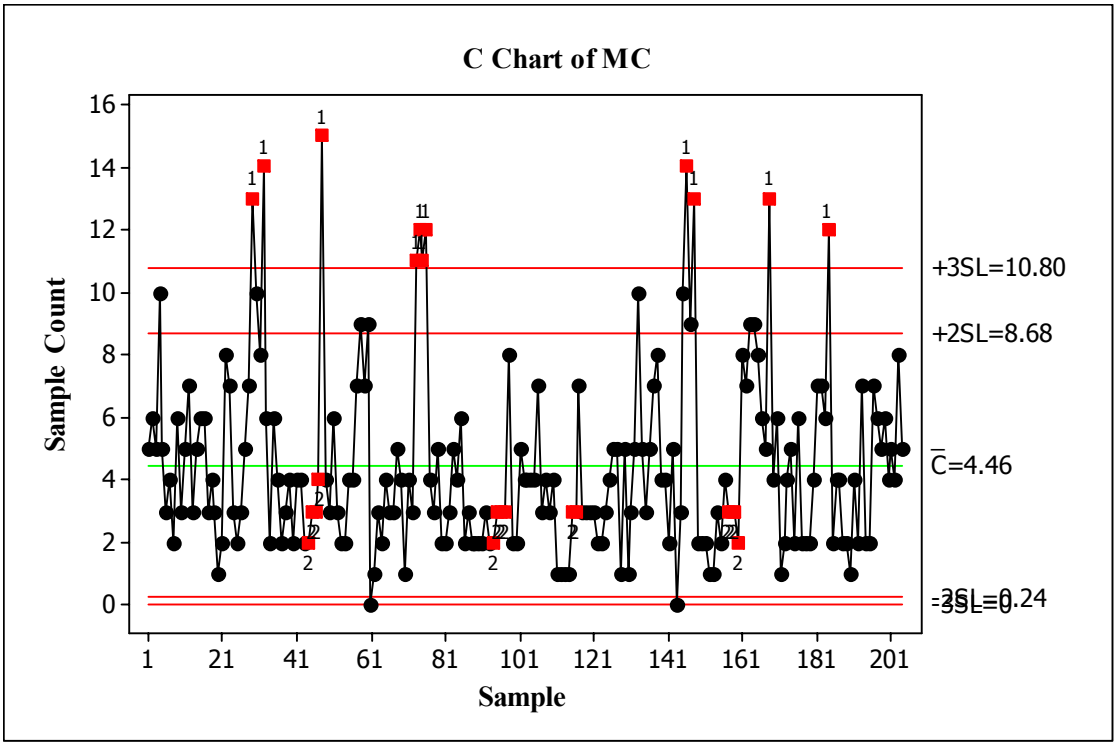
The upper and lower control limits are based upon 2 and 3 SD based upon the Poisson distribution.

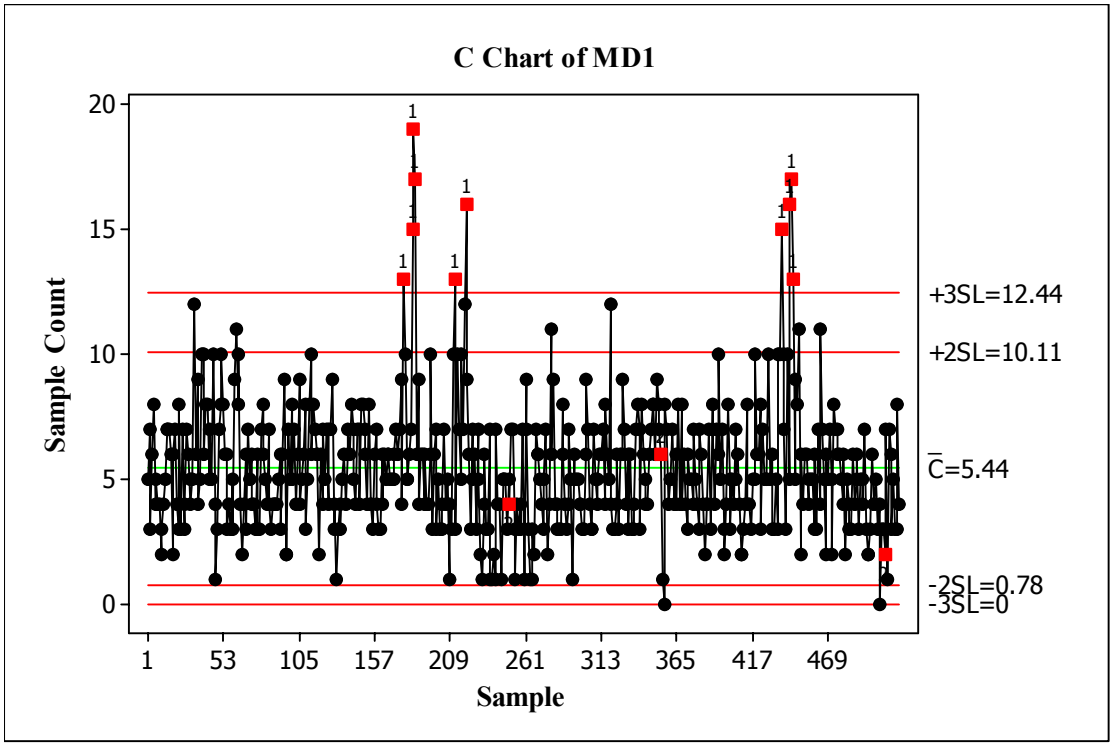
The red points on graphs represent cases 'flagged' as outside control based upon rules 1-4 defined as:

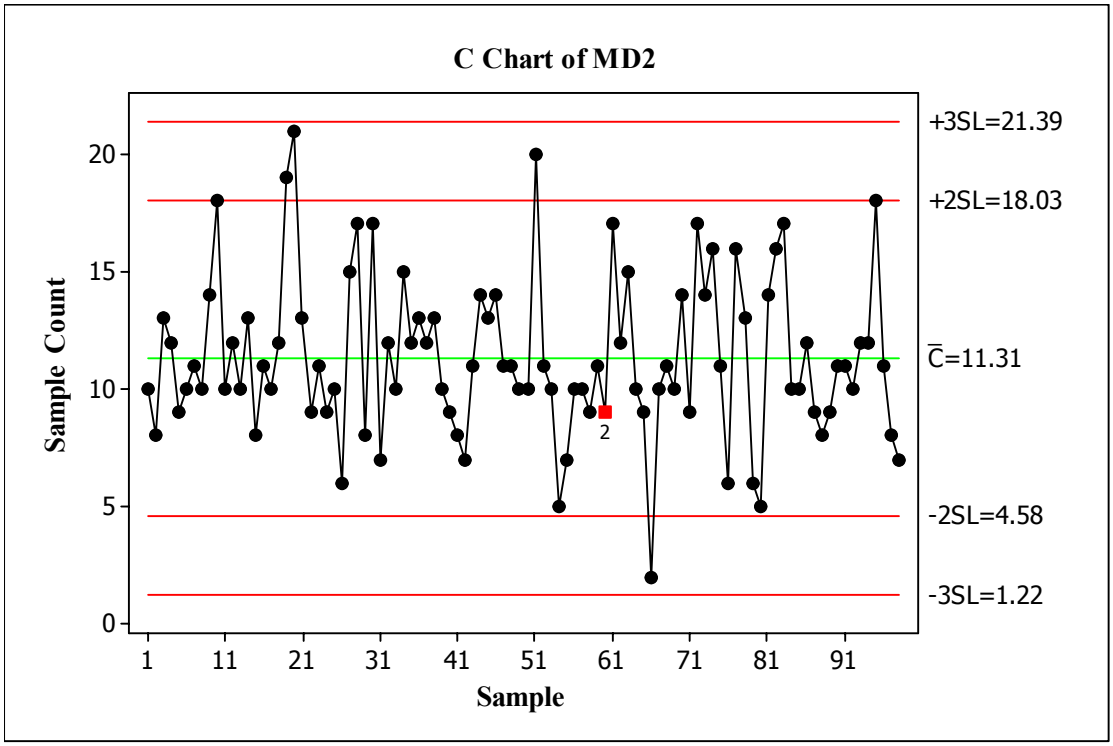
1) 1 point more than 3 standard deviations from the centre line
2) 9 points in a row on same side of centre line
3) 6 points in a row, all increasing or all decreasing
4) 14 points in a row, alternating up and down

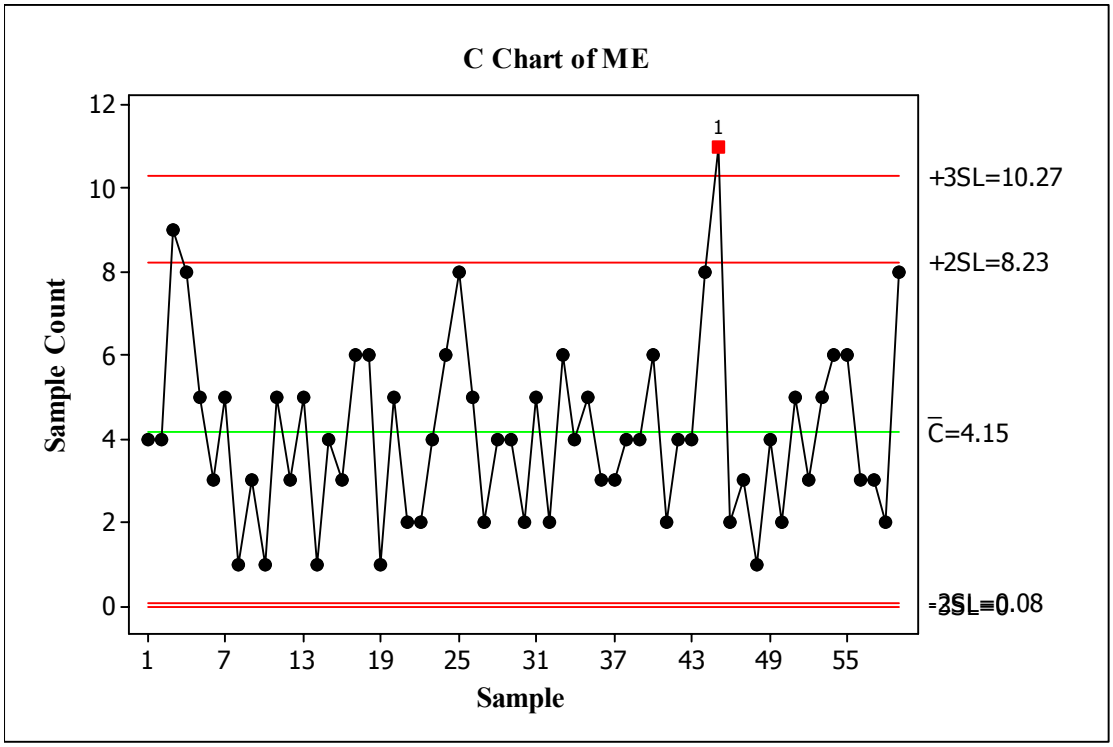

C Chart of MA
Sample Count
Sample
+3SL=13.04
+2SL=10.63
C=5.81
-2SL=0.99
-3SL=0

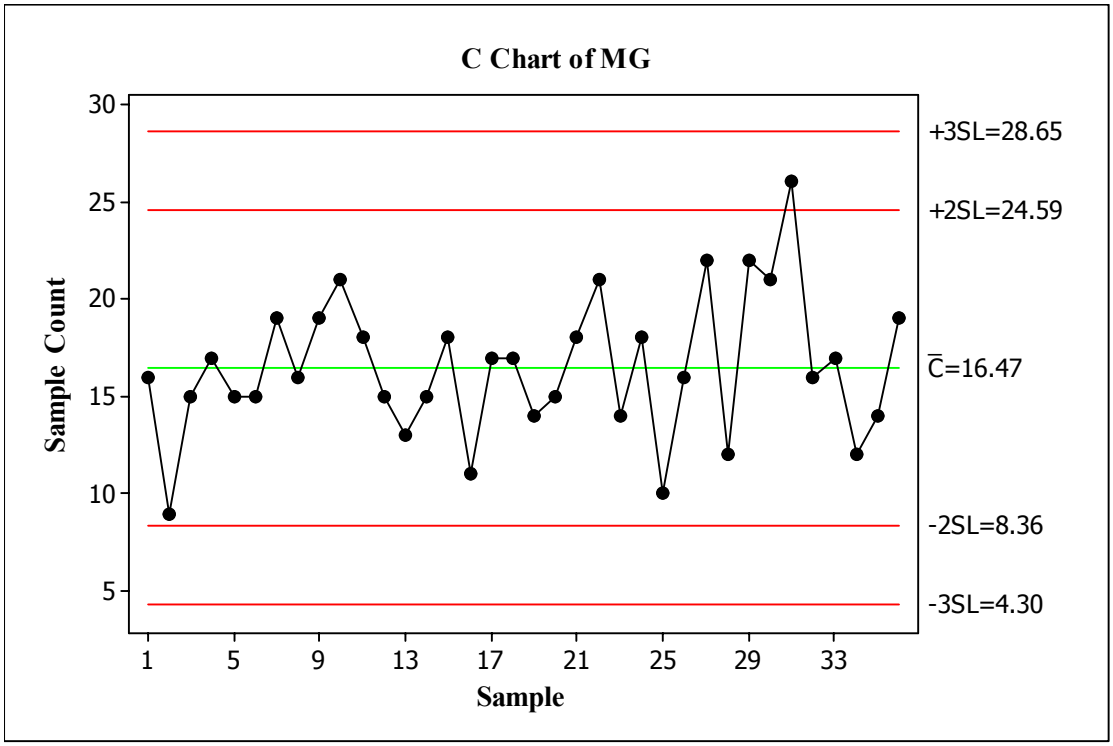

C Chart of MB
Sample Count
Sample
+3SL=6.465
+2SL=5.013
C=2.109
-3SL=0

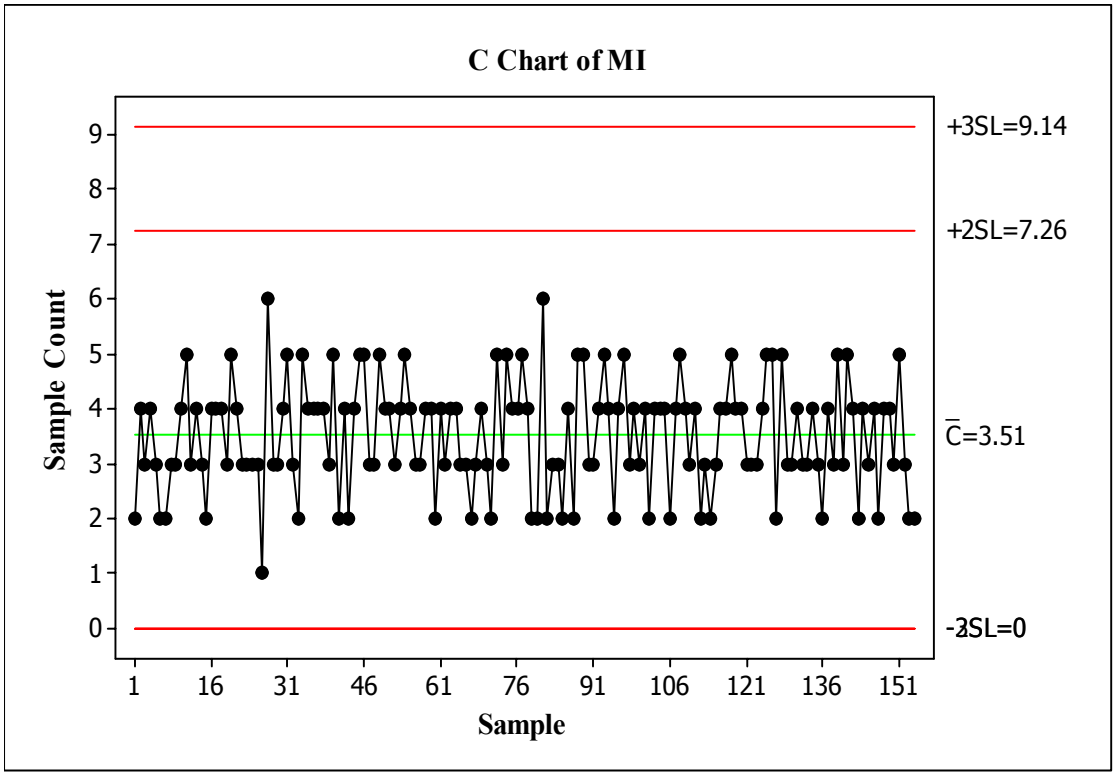

C Chart of MC
Sample Count
Sample
+3SL=10.80
+2SL=8.68
C=4.46
-2SL=0.24
-3SL=0

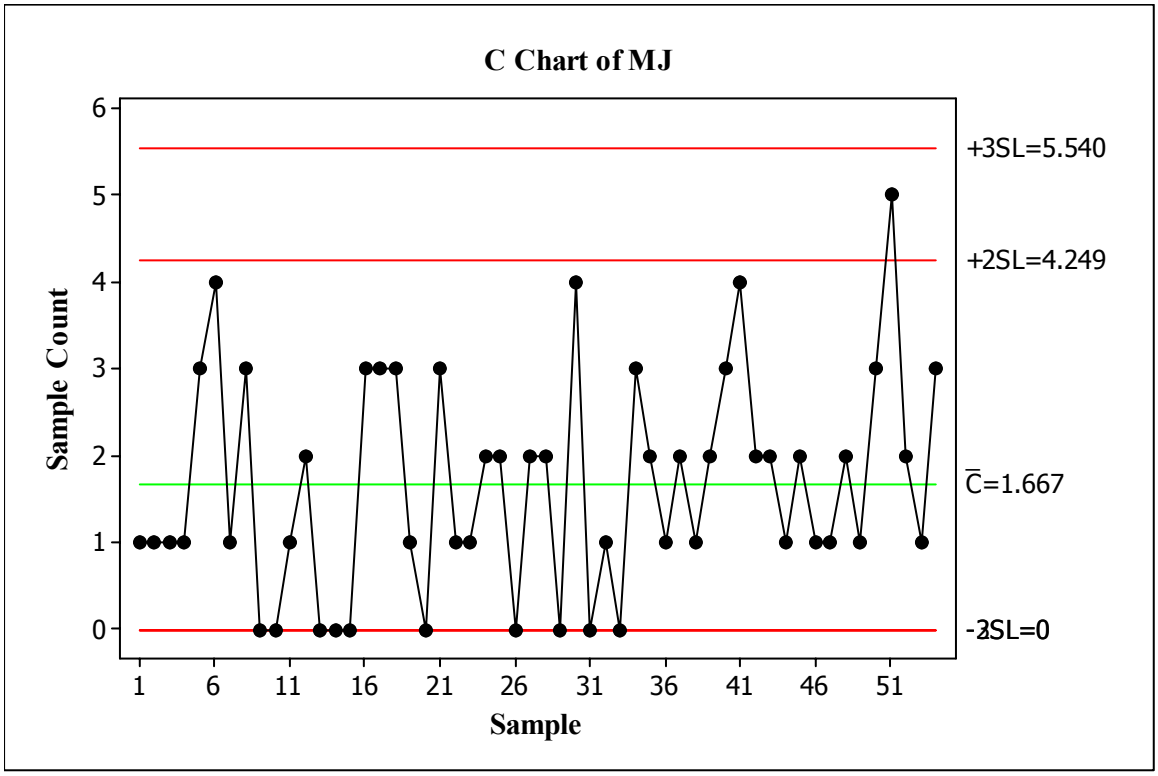

C Chart of MD1
Sample Count
Sample
+3SL=12.44
+2SL=10.11
C=5.44
-2SL=0.78
-3SL=0

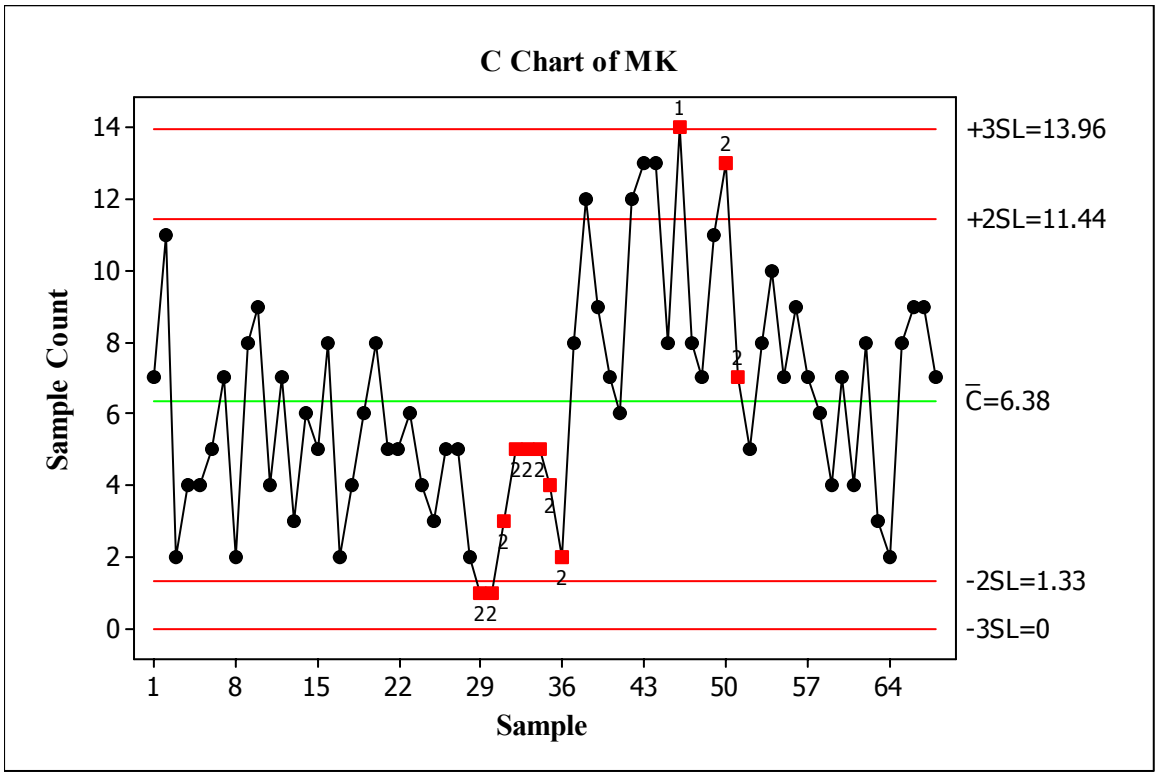

C Chart of MD2
+3SL=21.39
+2SL=18.03
C̄=11.31
-2SL=4.58
-3SL=1.22
Sample Count
Sample
0
5
10
15
20
1
11
21
31
41
51
61
71
81
91
2


C Chart of ME
+3SL=10.27
+2SL=8.23
C̄=4.15
-2SL=0.08
-3SL=0
Sample Count
Sample
0
2
4
6
8
10
12
1
7
13
19
25
31
37
43
49
55
1


C Chart of MG
Sample Count
Sample
+3SL=28.65
+2SL=24.59
C=16.47
-2SL=8.36
-3SL=4.30
30
25
20
15
10
5
1
5
9
13
17
21
25
29
33


C Chart of MI
Sample Count
Sample
+3SL=9.14
+2SL=7.26
C=3.51
-2SL=0
9
8
7
6
5
4
3
2
1
0
1
16
31
46
61
76
91
106
121
136
151


C Chart of MJ
Sample Count
Sample
+3SL=5.540
+2SL=4.249
C=1.667
-2SL=0
-3SL=0


C Chart of MK
Sample Count
Sample
+3SL=13.96
+2SL=11.44
C=6.38
-2SL=1.33
-3SL=0

ANNEX

Rationale for the recommendations developed in the updated Test Guidelines

The section below has been developed during the update of the Test Guidelines on chromosomal aberration and micronucleus tests *in vitro* and *in vivo* (TG 473, TG 475, TG 487 and TG 474), to support and explain the recommendations included in the updated TGs. They are based on the above statistical report and discussions of the expert group on the review of OECD TGs on genotoxicity.

TG 473:

Statistical considerations on the number of cells to score

Scoring 300 cells per concentration was considered a compromise between achieving an optimal number for the statistical power of the assay and the practicability of scoring. It also takes into account that where the events (cells with a chromosomal aberration) have an incidence of 1% (p=0.01) and are Poisson distributed and 500 cells (n=500) are scored per cultures then cultures with no cells with aberrations will be rare. This is based upon the Poisson distribution of counts with a mean (np) of 5 where less that 0.7% of cultures are expected to be zero. The Poisson distribution also approximates to a normal distribution when np ≥5. In the case of 300 cells, assuming a Poisson distribution approximately 5% of cultures would be expected to be zero.

TG 473 and TG 487:

Statistical considerations on the use of historical data

Historical negative control data are used both for acceptability of the assays and to help with the interpretation of the results in those case where results are neither clearly negative nor clearly positive. Quality control methods (for example C-charts or X-bar charts) should be used to characterise the distribution of the negative control data base.

Concurrent negative controls should ideally be within the 95% control limits of the distribution of the laboratory's historical negative control database. Where concurrent negative control data fall outside the 95% control limit they may be acceptable for inclusion in the historical control distribution as long as these data are not extreme outliers and there is no evidence that the test system is no longer 'under control' and no evidence of technical or human failure.

These methods can be used to assess the acceptability of the assay, using the 95% control limits. In order to avoid narrowing the database, all data should be included in the historical data base, even when rejected for the evaluation of the results, unless a technical or human failure has been identified. In this case the data will not be added to the database. The choice of the 95% control limit over the more stringent 99% control limit provides flexibility to avoid unnecessarily discarding experiments and avoids narrowing the control database range.

In the interpretation of the results, the treated values should be compared with the 95% control levels for the distribution of the historical negative control data base. Results which fall within the 95% control levels for the historical control could be considered broadly consistent with historical negative control data.

TG474:

Statistical considerations on the number of cells to score

Micronucleated immature erythrocytes (MN-IEs) arise spontaneously at low frequencies, typically from 0.02 to 0.2% ([1][2]; unpublished survey), which has focused attention on the question of whether maintaining the recommendation to score 2000 IEs per animal will provide sufficient statistical power to detect treatment-induced increases in the MN-IE frequency.

Recommendations on the number of IEs to score have been made as an outcome of several past International Workshops on Genotoxicity Testing (IWGT). The 2nd IWGT expert meeting of March 25-26, 1999 recommended that, "Sufficient cells should be analyzed to assure the detection of a doubling in micronucleus frequency with 80% power at a significance level of less than 0.05" [3]. Subsequently, the 4th IWGT meeting of September 8-9, 2005 further recommended that, "...it is desirable that each laboratory should determine the minimum cell sample size required to ensure that scoring error is maintained below the level of animal-to-animal variation" [4].

The power of various experimental designs has been considered in several studies [5][6], which have indicated scoring 2000 IEs does not provide sufficient power to detect a doubling in MN-IE frequency with 80% power at a significance level of less than 0.05. In general, studies employing sample sizes of 5 animals/group and scoring 4000-5000 cells per animal will have approximately 80% power to detect about 2-3 fold increases when the spontaneous frequency is ≥0.1%, but that the power will be much smaller for lower background levels.

Consequently, scoring 4000 cells per animal would provide reasonable power to detect an increase that would meet the criteria for a biologically significant positive response and serve as a compromise allowing continued manual scoring by microscopy. This number of cells also takes into account that where Poisson distributed events (e.g., micronucleated immature erythrocytes) have a background incidence of 0.1% (p=0.001) and 4000 cells (n=4000) are scored per animal, then the probability of detecting no spontaneously arising immature erythrocytes with micronuclei in an animal will be rare (1.8% of animals). This determination is based upon the Poisson distribution of counts with a mean (np) of 4 where approximately 1.8% of animals are expected to have a zero count.

---

(1) Krishna, G., Urda, G. and Paulissen, J. (2000). Historical vehicle and positive control micronucleus data in mice and rats. Mutat Res, 453, 45-50.

(2) Salamone, M.F. and Mavournin, K.H. (1994). Bone marrow micronucleus assay: a review of the mouse stocks used and their published mean spontaneous micronucleus frequencies. Environ Mol Mutagen, 23, 239-273.

(3) Hayashi, M., MacGregor, J.T., Gatehouse, D.G., Adler, I.D., Blakey, D.H., Dertinger, S.D., Krishna, G., Morita, T., Russo, A. and Sutou, S. (2000). In vivo rodent erythrocyte micronucleus assay. II. Some aspects of protocol design including repeated treatments, integration with toxicity testing, and automated scoring. Environ Mol Mutagen, 35, 234-252.

(4) Hayashi, M., MacGregor, J.T., Gatehouse, D.G., Blakey, D.H., Dertinger, S.D., Abramsson-Zetterberg, L., Krishna, G., Morita, T., Russo, A., Asano, N., Suzuki, H., Ohyama, W., Gibson, D. and In Vivo Micronucleus Assay Working Group, IWGT. (2007). In vivo erythrocyte micronucleus assay III. Validation and regulatory acceptance of automated scoring and the use of rat peripheral blood reticulocytes, with discussion of non-hematopoietic target cells and a single dose-level limit test. Mutat Res, 627, 10-30.

(5) Kissling, G.E., Dertinger, S.D., Hayashi, M. and MacGregor, J.T. (2007). Sensitivity of the erythrocyte micronucleus assay: dependence on number of cells scored and inter-animal variability. Mutat Res, 634, 235-240.

(6) Hayes, J., Doherty, A.T., Adkins, D.J., Oldman, K. and O'Donovan, M.R. (2009). The rat bone marrow micronucleus test—study design and statistical power. Mutagenesis, 24, 419-424.

TG 475:

Statistical considerations on the number of cells to score

Cells with structural chromosomal aberrations arise spontaneously at low frequencies, typically about 1%[(7)], which has focused attention on the question of whether maintaining the recommendation to score 100 metaphases per animal will provide sufficient statistical power to detect treatment-induced increases.

Scoring 200 cells per animal would, assuming a background frequency of 1%, provide reasonable power to detect an increase that would meet the criteria for a biologically significant positive. Where Poisson distributed events (e.g., metaphases with structural chromosomal aberrations) have a background incidence of 1% (p=0.01) and 200 cells (n=200) are scored per animal, then the probability of detecting no cells with spontaneously arising structural chromosomal aberrations (excluding gaps) in an animal will be somewhat low (13.5% of animals). This determination is based upon the Poisson distribution of counts with a mean (np) of 2.

(7) Adler, I.D., Bootman, J., Favor, J., Hook, G., Schriever-Schwemmer, G., Welzl, G., Whorton, E., Yoshimura, I. and Hayashi, M. (1998). Recommendations for statistical designs of in vivo mutagenicity tests with regard to subsequent statistical analysis. Mutat Res, 417, 19-30.