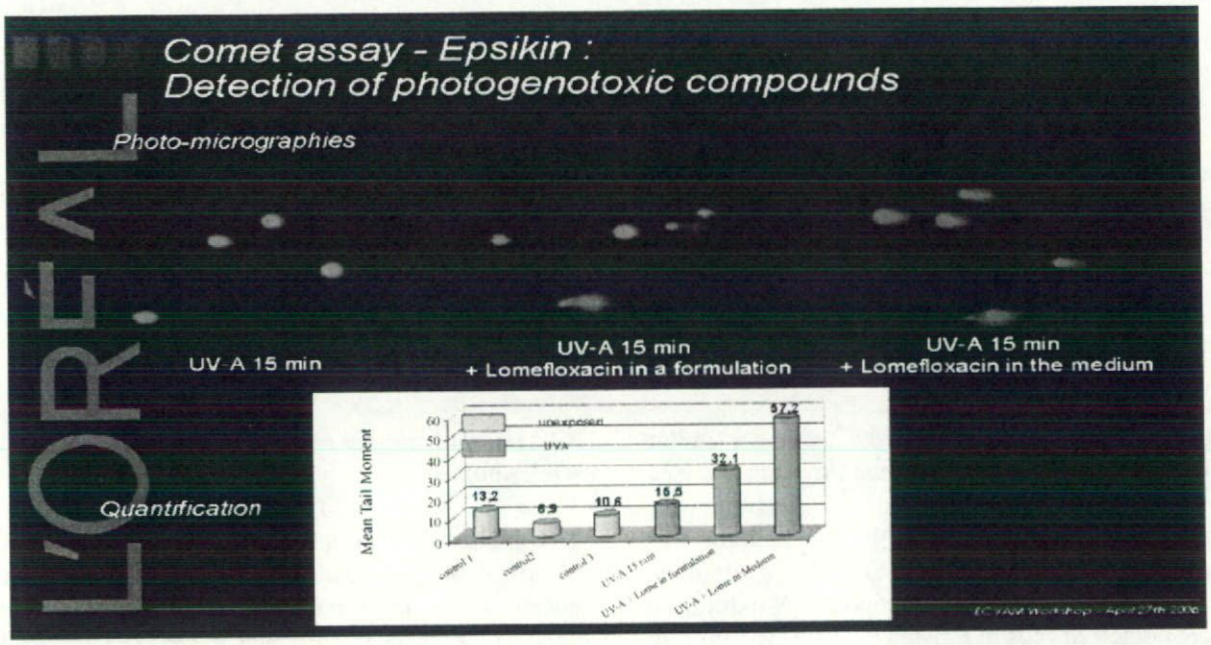

Fig. 17. Induction of comets by lomefloxacin + UV in Episkin. For the photosensitization experiments with lomefloxacin, the Episkin® epidermises were treated either by adding a 50 μm lomefloxacin solution in ethanol to the Episkin® culture medium "underneath the skin" (mimicking the systemic flow) or by topical application of 1 mg/cm^2 of a 4% lomefloxacin cream (kindly provided by L'Oreal Applied Research Laboratories at Chevilly LaRue, France). In those cases the Episkin® was treated for 1 h with lomefloxacin, rinsed three times with PBS and then irradiated 15 min with UVA light. The comet assay was performed immediately after the irradiation. All the unexposed controls (topically treated or culture-media treated epidermises, untreated epidermis) were included in the experiment as described in the figure.

- Check for degradation products in culture medium.
- Investigate possible confounding effects of apoptosis (e.g. by using CTLL-2 *bcl2* cells or Annexin-V analysis) and necrosis.
- Check for agonist and antagonist effects on kinases.
- Check for loss of cellular homeostasis (e.g. high osmolality, low pH).
- Check for metabolic poisoning and inhibition of DNA synthesis.
- Check for possible exposure to UV light.
- Check for nucleotide pool imbalances.
- Check for metabolic overload (e.g. glutathione depletion).
- Determine absence of DNA adducts under genotoxic conditions, preferably using radiolabeled chemical rather than ^{32}P-postlabeling.

In order that practising scientists and regulatory reviewers may become more familiar with the likely causes of genotoxic responses that are not relevant for humans, we encourage journals to publish data on coded compounds (from industrial in-house databases) that will help exemplify non-DNA or non-relevant mechanisms of genotoxicity. However, in such cases, as much information as possible should be provided on each test agent, and an independent review of the classification of positive and negative calls will probably be needed.

In addition to the above suggestions, it was acknowledged that 3D tissue models, such as those presented for skin, could provide valuable information on the relevance of *in vitro* positive results, and the further development of such models is to be encouraged. However, given that relatively few human or animal skin carcinogens (UV light, PAHs) exist, these models may be of limited value as replacements for whole animals.

3.2. Cell culture conditions and techniques

It was agreed that "good housekeeping" of cell cultures is a requirement for reliable and reproducible results. Practitioners should avoid working with high passage cells and should look at chromosomal content, karyotype and other characteristics such as metabolic capability and response to reference genotoxins for evidence of genetic drift. There are some suggestions that cell density and culture size can have an impact on the response of the cultures to chemical insult and this needs to be investigated. The ECVAM task force on good cell culture practice (GCCP) has previously published recommendations on GCCP [8]. These recommendations need to be reviewed in light of the current workshop to

see if they need to be up-dated and/or adapted to *in vitro* genotoxicity testing.

There was concern at the possibility of reactive oxygen species being formed by reaction between the culture medium and test chemical. More data are needed before advice can be given on whether certain cell/media systems are less likely to produce artefactual results through oxidative stress than others, and Halliwell is encouraged to continue his investigations in the hope that such recommendations can be made.

3.3. Biotransformation

The xenobiotic-metabolising system comprises several hundred enzymes, which are usually expressed with high selectivity in varying tissues, cell types and ontogenetic stages, and in rodents also often with high sex specificity. Some enzymes are involved in the biotransformation of many genotoxicants, others are important only for a small number of compounds, and some reaction types involve a higher risk of formation of reactive metabolites than others. Various enzymes are only present at significant levels after induction by specific endogenous or xenobiotic factors. Thus, no cell type *in vivo* reflects the full biotransformation capacity of the organism. Even hepatocytes, which are heavily involved in biotransformation, only express a limited selection of xenobiotic-metabolising enzymes. For example, there are now strong indications that the hepatocarcinogenicity of PAHs in rodents is due to bioactivation by CYP1B1 in extrahepatic tissues, and that hepatic CYP1A1 – which normally is used for "promiscuous" activation of PAHs *in vitro* – acts as a major PAH-detoxifying enzyme *in vivo*. Moreover, the expression of numerous enzymes ceases, or is drastically decreased, in cells in culture. In part, this simply reflects the propensity of the organism to avoid expression of risk-borne enzymes in proliferating cells. This is true even for hepatic cell lines (e.g. HepG2) that have retained much biotransformation activity in comparison to fibroblastoid or lymphocytic cell lines. Classical S9 is a rich source of selected CYPs, but otherwise the spectrum of enzymes present in active form is very low.

The risks resulting from the formation of reactive intermediates is reduced *in vivo* by the presence of detoxifying systems, which are often extremely efficient (but overlap with toxifying systems). Two processes of detoxification can be distinguished:

(i) Metabolic "sequestration" of a promutagen into pathways that avoid the formation of the ultimate mutagen.

(ii) Inactivation of an active metabolite after its formation.

These processes differ in the enzyme systems involved. Sequestration commonly occurs by CYPs, reductases/dehydrogenases, UGTs and SULTs. The individual members of these enzyme classes are often expressed with high selectivity in certain cell types of physiological stages. Inactivation of active (electrophilic) metabolites is normally conducted by epoxide hydrolases or GSTs. These enzymes – although not every single form – are widely expressed in many tissues and cells. Thus, microsomal epoxide hydrolase and substantial levels of GST activity towards some substrates have been detected in all mammalian cell lines studied [31]. However, high-efficiency (high V_{max}/K_m) enzymes may be most important for sequestration and inactivation, especially in systems with high-efficiency, rather than promiscuous, activation. Some proximate and ultimate genotoxicants equilibrate *in vivo*. In this case, efficient protection may even occur through enzymes located at sites different from the site of activation. However, when this equilibration is limited, the appropriate localisation of the detoxifying system may be critical. An example is aflatoxin B_1, which is a potent hepatocarcinogen in the rat but only weakly active in the mouse. Constitutive expression of Gst a5, an enzyme that efficiently inactivates aflatoxin B_1 8,9-oxide in mouse but not rat liver, appears to be an important mechanism underlying this difference. Although a rat Gst a5 is constitutively expressed in various extrahepatic tissues, this localisation appears to be inefficient for toxification. However, after hepatic induction of Gst a5 by certain chemicals, the rat becomes resistant towards the hepatocarcinogenicity of aflatoxin B_1 [32]. It is not possible to mimic such complex, varying pharmacokinetic processes in a simple *in vitro* screening model.

Detoxification may occur to the parent compound or downstream to its metabolites. The capacity of *in vitro* systems to detoxify is usually modest, often limited to the nanomolar to low micromolar concentration range over the entire exposure period. Thus, no significant competition between toxifying and detoxifying activities would occur unless the maximum concentration of parent compound tested was rigorously restricted to a very low range, and the relevant enzymes were present. The situation is very different for metabolites. Their concentrations are low, implying high-affinity (or more precisely high-efficiency) detoxification reactions can potentially occur.

The purported lack of enzymic detoxification *in vitro* is in contrast to observations that various chemicals,

when tested at high concentrations, are positive in direct genotoxicity assays, but negative in the presence of S9. Although the underlying mechanism(s) for the lack of genotoxicity in the presence of S9 have not been elucidated in most cases, it has been demonstrated for some cases that heat-inactivated S9 was also protective (H.R. Glatt, unpublished results). This suggests that physical trapping (e.g. of lipophilic compounds in microsomes) or chemical trapping (e.g. reaction of electrophiles with nucleophilic sites), rather than enzymatic activities, produced the effect. Therefore, it is difficult to assess the *in vivo* significance of this observation.

Currently the impact of metabolic differences between *in vitro* and *in vivo* test systems on the false positive rate in *in vitro* genotoxicity tests is not known. However, it is clear that variation of the metabolising system can have dramatic effects on the results of *in vitro* tests as well as for animal studies, where genetic knockout or inhibition of an individual enzyme can eliminate the ability of a carcinogen to induce tumours. In an ideal world the same metabolic modulation should have parallel consequences *in vivo* and *in vitro*.

A review of the important *in vivo* genotoxins and DNA-reactive mutagenic carcinogens is needed to determine whether metabolic differences between *in vitro* and *in vivo* test systems are, in fact, contributing to the high false positive rate, and to better define the relevant metabolic systems to include in *in vitro* tests. Perhaps we will have to include a much larger variety or different set of enzyme systems than have been traditionally used in *in vitro* tests to predict better what happens in animals. Genetic engineering may be used to help address this goal, but this approach will be very time- and cost-intensive. In addition, to achieve the proper balance of all of the relevant enzymes in a given engineered cell line to appropriately detect all *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens is likely to be an insurmountable task. Moreover, such an approach would not reflect the *in vivo* situation, where different enzymes are often compartmentalised in different cells. Alternatively, a panel of individual cell lines, each with a small number of expressed enzymes, could be used. Almost infinite possibilities for permutations of various enzymes exist with such an approach, generating ideal tools for research on mechanisms. However, the selection of a cell battery for broad scale genotoxicity screening would require some arbitrary, pragmatic decisions, and this arbitrary character, as well as large deviations from the balance of enzymes *in vivo*, would remain obvious. This is in contrast to alternative metabolising systems (S9, conventional "metabolically competent" cell lines) where the situation may be similar but less obvious and less open for remedies by systematic, hypothesis-driven research in a concrete situation.

From the point of view of genetic engineering, one may either start with a cell line that has retained some residual xenobiotic-metabolising activities (such as HepG2), or from a relatively "clean" cell line (V79 or a human equivalent, if available). The former model has the advantage that less engineering is required to achieve broad xenobiotic-metabolising capacities. The latter model is favoured when used as an analytical tool to specify critical host factors, as background activities are minimised. Further aspects that should be taken into account in the selection of the basic system(s) are:

(a) Genetic engineering intrinsically involves numerous culture passages and clonal selections; therefore, long-term stability of critical properties (obviously including biotransformation activities) is pivotal.
(b) The cell line selected should be suitable for efficient analysis of important and robust endpoints, such as gene mutations.

3.4. Top concentration for testing

Current OECD guidelines for genotoxicity testing in mammalian cells require that the top concentration with soluble and non-toxic substances should be 10 mM or 5000 μg/ml, whichever is the lower. There was some discussion about whether this may be appropriate for complex mixtures and technical grade (impure) industrial chemicals for example, where the objective is not only to test the genotoxicity of the main ingredient. However, given that mutagenic impurities are only detected in an Ames test carried out to 5 mg/plate (when spiked at a level of about 5% [33]), it was questioned whether detection of impurities is a sufficiently important role for genotoxicity testing to justify pushing to such high concentrations. However, the K_m for many biochemical reactions, whether involved in metabolic activation/inactivation, general cellular defence/balance, cellular transport or cellular turnover is less than 100 μM [34,35,36]. It is probable that low K_m reactions primarily determine the bioactivation pathway *in vivo*. These kinetic characteristics suggest that the high concentrations currently required for *in vitro* testing may not be informative for human risk assessment. The 10 mM and 5000 μg/ml requirements are seemingly based on a small number of carcinogens that needed high concentrations before giving positive responses in mammalian cell tests *in vitro*, sometimes using inappropriate metabolic conditions. It is not known whether the carcinogens that require these high concentrations for detection *in vitro*

are "important", how robust the *in vitro* mammalian cell findings are, or whether these chemicals are positive in other test systems (e.g. the Ames test) currently used in a standard battery. The fact that the published data on these chemicals are probably quite old could mean that under current chromosomal aberration and gene mutation protocols they could be detected at lower concentrations. It also has to be considered that simple detection of a carcinogen at high *in vitro* concentrations that are not relevant *in vivo* does not mean there is a mechanistic correlation between the *in vitro* genotoxicity and the *in vivo* carcinogenicity.

Further to the above considerations on the K_m of important biochemical processes, general considerations on *in vivo* exposure to (toxic) compounds have been put forward. In this context one can consider knowledge about intentional high dose or long-term exposure to pharmaceuticals as worst case examples. It is clear that even high dose pharmaceuticals such as antibiotics (e.g. penicillins, fluoroquinolones) or pain relief agents such as acetaminophen (also known as paracetamol) seldomly yield systemic or tissue levels >10 μM [37]. Thus, taking chronic intake, possible accumulation and overdosing scenarios into account, a lowering of the current maximum *in vitro* concentration, perhaps by 10-fold or more, may be justified, at least for certain types of chemicals (such as pharmaceuticals) from scientific and consumer protection viewpoints.

The participants therefore agreed that a new review of existing data is needed to determine whether such high concentrations as 10 mM or 5000 μg/ml are needed to detect *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens. The following actions are recommended:

- An expert panel should be assembled to determine which *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens need to be detected in *in vitro* mammalian cell tests. This subset of chemicals should be determined from both published and industry (confidential) data. A first suggestion of an important data set would be the IARC groups 1, 2A and 2B carcinogens, omitting those, such as hormones and immunosuppressants, which are acknowledged to be of a non-genotoxic mode of action. In view of the on-going initiative of the Health and Environmental Sciences Institute of the International Life Sciences Institute (ILSI-HESI) with regard to false positives in *in vitro* mammalian cell genotoxicity tests, it was suggested this action could be best achieved in conjunction with ILSI-HESI, and the outcome of their workshop (held in June 2006) will be reported elsewhere.
- The role of metabolism in the activity of the above *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens needs to be reviewed in order to define the appropriate metabolic systems to include in *in vitro* genotoxicity assays.
- The published and industry data should be reviewed to determine whether concentrations as high as 10 mM or 5000 μg/ml are needed to detect this important subset of chemicals, or whether a lower level could be justified.
- If high concentrations are needed in the mammalian cell tests, data from other tests such as the Ames test should be reviewed to see if the chemical(s) would be detected in other parts of the standard battery.
- If high concentrations are needed, and genotoxicity was not detected in other parts of the standard battery, an opinion should be formed as to whether a more modern protocol or modified metabolic conditions would be likely to detect genotoxic effects at lower concentrations. If necessary, new testing should be initiated.
- If high concentrations are needed (with appropriate metabolic conditions) a scientific effort should be mounted to elucidate whether the mechanism(s) of genotoxicity for these chemicals would trigger responses in other toxic endpoints.
- A thorough evaluation of various human exposure scenarios should be made to determine an upper limit of *in vitro* testing from the viewpoint of human consumer protection.

It is evident that the concentration of a pro-genotoxicant required for a positive test result can vary dramatically depending on the activating system used. General improvements in the activation systems, or in a chemical class-dependent manner (with corresponding positive control compounds), might be an important pre-requisite for a reduction in the top concentration. For the time being, the participants of the workshop concluded that it is prudent to challenge the current recommended upper concentration of *in vitro* testing (10 mM or 5000 μg/ml), and that a lower level appears to have scientific merit.

3.5. Measures and extent of cytotoxicity

Many different measures of cytotoxicity are used in mammalian cell tests, in particular the chromosomal aberration test. Reductions in cell count, confluency, mitotic index and population doubling are all widely used and equally accepted, but it is unlikely that all these measures would select the same top concentra-

tion for testing. Other indicators of toxicity such as ATP levels, mitochondrial function, LDH-leakage may also be appropriate. Greenwood et al. [11] showed that several non-DNA-reactive chemicals and metabolic poisons would not have given positive chromosomal aberration results if the 50% cytotoxic concentration had been chosen based on a reduction in population doubling rather than a reduction in cell count, and that, using this measure, no important DNA-reactive genotoxins would have been missed. These findings have not been independently verified and there are no other publications comparing different measures of cytotoxicity in relation to genotoxicity.

The participants therefore agreed there is a need for a thorough comparison of different measures of cytotoxicity in case some measures may select concentrations for testing that allow the detection of all important *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens but lowers the risk of false positives. In particular:

- A collaborative trial is needed on a selected set of chemicals. This could be the same subset of chemicals selected for evaluation of top concentration (above). Additional sets of chemicals, such as the non-DNA-reactive chemicals of the Greenwood et al. [11] publication, and agreed non-genotoxins also need to be included. If possible, this trial should be co-ordinated with any initiatives coming from the ILSI-HESI workshop (June 2006) which will be reported elsewhere.
- Multiple endpoints of toxicity need to be compared at the same time in each participating laboratory.
- Dose–response relationships for cytotoxicity and genotoxicity should be determined, and include the current required levels (i.e. at least 50% toxicity for chromosomal aberrations, at least 60% toxicity for the micronucleus assay and at least 80% toxicity in the mouse lymphoma assay).
- Human cells such as lymphocytes should be included as well as the rodent (and any other, e.g. TK6, HepG2) cell lines, and measures of cytotoxicity for lymphocytes other than mitotic index need to be identified.
- Robust endpoints of genotoxicity that are not sensitive to interference by cytotoxicity (e.g. gene mutations and DNA adducts) should be identified and included in the trials.

Although the only indication came from the data of Elhajouji, there was a belief amongst many participants that a majority of false positive results in chromosomal aberration and mouse lymphoma tests probably occur in the prolonged, continuous treatments in the absence of exogenous metabolic activation. Several factors such as extent of exposure, prolonged cytotoxicity, and lack of detoxification by S9 may be involved, but the exact reasons are not known. However, it was noted that the 50% and 80% toxicity requirements for the chromosomal aberration and mouse lymphoma assays were originally based on short (e.g. 3–6 h) treatments, and the need for these levels of toxicity (or even the appropriateness of the current measures) has not been independently justified. Therefore, this collaborative trial must include chemicals that are only positive after prolonged (e.g. 20–24 h) treatments.

3.6. Criteria for and evaluation of new mammalian cell test systems

Certain characteristics of the commonly used rodent cell lines (CHO, CHL, V79, L5178Y, etc.) such as their *p53* status, karyotypic instability, DNA repair deficiencies, etc. are recognised as possibly contributing to the high rate of false positives. The need for exogenous metabolism with the cell systems is also expected to contribute to the false positive rate. If these cell types are to be replaced in the future, any new systems should ideally:

- Be early passage.
- Be karyotypically stable and, if possible, normal.
- Be *p53* proficient.
- Be DNA repair proficient.
- Preferably consist of human cells.
- Be metabolically competent (at least Phase 1 and Phase 2 capacity should be defined), if necessary through genetic engineering. Expert advice is needed on what are the essential Phase 1 and Phase 2 enzymes that should be functional in any new system for the biotransformation of *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens. Several different cell lines with different enzyme profiles may cover a reasonable fraction of the complex biotransformation machinery present *in vivo*. However, the selective expression of a limited number of enzymes in a cell line may direct the biotransformation into pathways that are different from the major pathways occurring in animals and humans, where many different enzymes compete and interact with each other.
- Be able to detect the majority of genetic endpoints relevant to human somatic and inherited disease.
- Show improved specificity without reducing the ability to detect *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens.

The participants were aware that this wish list is far too ambitious to be realised in a single test system, at least within a reasonable time. Therefore, several systems may be required, which in combination will facilitate the distinction between correct and false positive results.

Some of the data presented at the workshop indicated that the human lymphocyte cell system might produce a lower level of false positives than the common rodent cell lines. If the use of the clastogenicity endpoint *in vitro* is to be continued, further work is needed to establish whether human lymphocytes do offer a lower false positive rate, and also to find alternative methods (other than mitotic index) for measuring cytotoxicity.

Although many of the new systems presented at the Workshop show promise, and fulfil some of the criteria given above, none fulfils all of the criteria. In many cases (e.g. MCL-5, HepG2, transgenic cell lines) it is the lack of data on specificity that is the problem. In the case of the *GADD45a-GFP* assay it is mainly the lack of data with compounds requiring metabolic activation. The new 3D skin models are also at an early stage of development. It was therefore agreed that a collaborative research program is needed to evaluate new mammalian cell-based methods and models for genotoxicity. In addition to the cell systems discussed at the workshop, other cells that have retained some xeno-metabolic activities (e.g. HepaRG [38] and AR42J-B13 rat pancreatic stem cells [39,40]) or cell lines genetically engineered to express appropriate Phase 1 and Phase 2 metabolism should be considered.

This collaborative research program will be a major exercise, but if it is not undertaken we will still be faced with an unacceptable level of false positives and the consequential unnecessary follow-up *in vivo* testing, for decades to come. A steering committee, consisting of genotoxicity, metabolism and chemistry experts from academia, industry and the regulatory agencies should be established to draft a plan for this program. The participants did not reach any conclusion as to how this trial should be organised, but it is hoped that ECVAM may be able to contribute, to liase with other related activities from organisations such as ILSI-HESI and to begin to identify sources of funding and support.

4. Conclusions

The workshop participants agreed that some of the commonly used cells for genotoxicity testing (in particular some of the rodent cell lines) have produced an unacceptably high level of false positive results when compared with known *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens. In most of the rodent cell lines used, deficiencies in metabolism, *p53* function and DNA repair capability almost certainly contribute to this high false positive rate. Better guidance on the likely mechanisms resulting in positive results that are not relevant for humans, and on how to obtain evidence for those mechanisms, is needed both for practitioners and regulatory reviewers.

Testing up to high concentrations and high levels of cytotoxicity as is currently required in mammalian cell genotoxicity tests are also likely to contribute to the high frequency of false positive results, and may not be justified. A thorough review of published and industry data to determine whether such levels are required for the detection of *in vivo* genotoxins and DNA-reactive, mutagenic carcinogens is urgently needed. Suggestions to lower the current upper limit (perhaps by 10-fold or more) may be justified in terms of metabolic and cellular processes, and human tissue exposures. This needs urgent but careful evaluation.

Various measures of cytotoxicity are currently allowable under OECD guidelines, but there is little comparative data on whether different measures would select different concentrations for test. A detailed comparison of multiple measures of cytotoxicity, in relation to endpoints such as clastogenicity, is needed. Also, genotoxicity endpoints that are not intrinsically linked with processes leading to cytotoxicity need to be developed.

There was agreement amongst the workshop participants that cell systems preferably of human origin, which are *p53* and DNA-repair proficient, and have defined Phase1 and Phase 2 metabolism, covering a broad set of enzyme forms, and used within the context of appropriately set limits of concentration and cytotoxicity, offer the best hope for reduced false positives in the future. Whilst there is some evidence that human lymphocytes are less susceptible to false positives than the current rodent cell lines, other cell systems based on HepG2, TK6 and MCL-5 cells, and 3D skin models based on primary human keratinocytes also show some promise. However, much effort will be required to introduce a broader spectrum of metabolic capabilities into these or other target cells. Other human cell lines such as HepaRG have not been used for genotoxicity investigations and should be studied. A collaborative research programme is needed to identify and evaluate new cell systems with appropriate sensitivity but improved specificity.

Perhaps most importantly, the participants in this workshop felt that it is time for the scientific community to give careful consideration to the carcinogens and *in vivo* genotoxins we expect any new assay, or modified existing assay, to detect. Rodent bioassays have flaws in terms of detecting human carcinogens, and this is com-

pounded by trying to use genotoxicity assays to detect as many rodent carcinogens as possible. This has led to the continued expansion of genotoxicity test protocols (e.g. addition of new treatment and sampling regimens, new strains of bacteria, etc.), and the combination of these tests in batteries. The consensus of the group was that there is a need to refocus on the detection on human carcinogens and *in vivo* genotoxins that are DNA reactive. Without this change of focus, the development and validation of any new methods or assays will likely not lead to an overall improvement.

References

[1] D. Kirkland, M. Aardema, L. Henderson, L. Müller, Evaluation of the ability of a battery of 3 *in vitro* genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. I. Sensitivity, specificity and relative predictivity, Mutat. Res. 584 (2005) 1–256.

[2] E.J. Matthews, N.L. Kruhlak, M.C. Cimino, R.D. Benz, J.F. Contrera, An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data. I. Identification of carcinogens using surrogate endpoints, Regul. Toxicol. Pharmacol. 44 (2006) 83–96.

[3] B. Halliwell, Oxidative stress in cell culture: an under-appreciated problem? FEBS Lett. 540 (2003) 3–6.

[4] L.M. Wee, L.H. Long, M. Whiteman, B. Halliwell, Factors affecting the ascorbate- and phenolic-dependent generation of hydrogen peroxide in Dulbecco's modified Eagles medium, Free Radic. Res. 37 (2003) 1123–1130.

[5] L.H. Long, D. Kirkland, B. Halliwell, Different cytotoxicities of epigallocatechin gallate or ascorbate in various cell culture media due to variable rates of oxidation in the culture medium, Mutat. Res., submitted for publication.

[6] A. Santoro, M.B. Lioi, J. Monfregola, S. Salzano, R. Barbierri, M.V. Ursini, L-Carnitine protects mammalian cells from chromosome aberrations but not from inhibition of cell proliferation induced by hydrogen peroxide, Mutat. Res. 587 (2005) 16–25.

[7] D.J. Tweats, D.G. Gatehouse, Further debate of testing strategies, Mutagenesis 3 (1988) 95–102.

[8] S. Coecke, M. Balls, G. Bowe, J. Davis, G. Gstraunthaler, T. Hartung, R. Hay, O.-W. Merten, A. Price, L. Schechtman, G. Stacey, W. Stokes, ECVAM good cell culture practice task force report 2, ATLA 33 (2005) 261–287.

[9] OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring. No. 14 on the Application of the Principles of GLP to *in vitro* studies, 2004, http://appli1.oecd.org/olis/2004doc.nsf/43bb6130e5e86e5fc12569fa005d0044c/58d5c8b13297a995c1256f5c006008a7/$FILE/JT00174939.DOC.

[10] T. Sofuni, A. Matsuoka, M. Sawada, M. Ishidate Jr., E. Zeiger, M.D. Shelby, A comparison of chromosome aberration induction by 25 compounds tested by two Chinese hamster cell (CHL and CHO) systems in culture, Mutat. Res. 241 (1990) 175–213.

[11] S.K. Greenwood, R.B. Hill, J.T. Sun, M.J. Armstrong, T.E. Johnson, J.P. Gara, S.M. Galloway, Population doubling: a simple and more accurate estimation of cell growth suppression in the *in vitro* assay for chromosomal aberrations that reduces irrelevant positive results, Environ. Mol. Mutagen. 43 (2004) 36–44.

[12] S. Meintières, A. Biola, M. Pallardy, D. Marzin, Using CTLL-2 and CTLL-2 *bcl2* cells to avoid interference by apoptosis in the *in vitro* micronucleus test, Environ. Mol. Mutagen. 41 (2003) 14–27.

[13] F.P. Guengerich, G.A. Dannan, S.T. Wright, M.V. Martin, L.S. Kaminsky, Purification and characterization of liver microsomal cytochromes P-450: electrophoretic, spectral, catalytic and immunochemical properties and inducibility of eight isozymes isolated from rats treated with phenobarbital or beta-naphthoflavone, Biochemistry 21 (1982) 6019–6030.

[14] M.O. Bradley, B. Bhuyan, M.C. Francis, R. Langenbach, A. Peterson, E. Huberman, Mutagenesis by selected agents in V79 Chinese hamster cells: a review and analysis of the literature—a report of the Gene-Tox Program, Mutat. Res. 87 (1981) 81–142.

[15] H.R. Glatt, W. Meinl, Sulfotransferases and acetyltransferases in mutagenicity testing: technical apsects, Meth. Enzymol. 400 (2005) 230–249.

[16] H.R. Glatt, Activation and inactivation of carcinogens by human sulfotransferases, in: G.M. Pacific, M.W.H. Coughtrie (Eds.), Human Sulphotransferases, Taylor & Francis, London, 2005, pp. 281–306.

[17] S. Wilkening, F. Stahl, A. Bader, Comparison of primary human hepatocytes and hepatoma cell line HepG2 with regard to their biotransformation properties, Drug. Metab. Dispos. 31 (2003) 1035–1042.

[18] S. Knasmüller, C. Cavin, A. Chakraborty, F. Darroudi, B.J. Majer, W.W. Huber, V.A. Ehrlich, Structurally related mycotoxins ochratoxin A, ochratoxin B, and citrinin differ in their genotoxic activities and in their mode of action in human-derived liver (HepG2) cells: implications for risk assessment, Nutr. Cancer 50 (2004) 190–197.

[19] Y.C. Staal, M.H. van Herwijnen, F.J. van Schooten, J.H. van Delft, Modulation of gene expression and DNA adduct formation in HepG2 cells by polycyclic aromatic hydrocarbons with different carcinogenic properties, Carcinogenesis 27 (2006) 646–655.

[20] S. Knasmüller, V. Mersch-Sundermann, S. Kevekordes, F. Darroudi, W.W. Huber, C. Hoelzl, J. Bichler, B.J. Majer, Use of human-derived liver cell lines for the detection of environmental and dietary genotoxicants; current state of knowledge, Toxicology 198 (2004) 315–328.

[21] V. Mersch-Sundermann, S. Knasmüller, X.J. Wu, F. Darroudi, F. Kassie, Use of a human-derived liver cell line for the detection of cytoprotective, antigenotoxic and cogenotoxic agents, Toxicology 198 (2004) 329–340.

[22] J.H. van Delft, E. van Agen, S.G. van Breda, M.H. Herwijnen, Y.C. Staal, J.C. Kleinjans, Discrimination of genotoxic from non-genotoxic carcinogens by gene expression profiling, Carcinogenesis 25 (2004) 1265–1276.

[23] C.L. Crespi, W.G. Thilly, Assay for gene mutation in a human lymphoblastoid line, AHH-1, competent for xenobiotic metabolism, Mutat. Res. 128 (1984) 221–230.

[24] C.L. Crespi, F.J. Gonzalez, D.T. Steimel, T.R. Turner, H.V. Gelboin, B.W. Penman, R. Langenbach, A metabolically competent human cell line expressing 5 cDNAs encoding procarcinogen-activating enzymes: application to mutagenicity testing, Chem. Res. Toxicol. 4 (1991) 566–572.

[25] P.A. White, G.D. Douglas, J. Gingerich, C. Parfett, P. Shwed, V. Seligy, L. Soper, L. Berndt, J. Bayley, S. Wagner, K. Pound, D. Blakey, Development and characterization of a stable epithelial cell line from Muta™Mouse lung, Environ. Mol. Mutagen. 42 (2003) 166–184.

[26] J. Gossen, J. Vijg, A selective system for *LacZ*$^-$ phage using a galactose-sensitive *E. coli* host, Biotechniques 14 (1993) 326–330.

[27] P.W. Hastwell, L.-L. Chai, K.J. Roberts, T.W. Webster, J.S. Harvey, R.W. Rees, R.M. Walmsley, High specificity and high sensitivity genotoxicity assessment in a human cell line: validation of the GreenScreen HC *GADD45a-GFP* genotoxicity assay, Mutat. Res. 607 (2006) 160–175.

[28] ICH Topic S2B, Genotoxicity: a standard battery for genotoxicity testing of pharmaceuticals, in: Proceedings of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, Step 4 Guideline, July 16, 1997.

[29] C.A. Hilliard, M.J. Armstrong, C.I. Bradt, R.B. Hill, S.K. Greenwood, S.M. Galloway, Chromosome aberrations *in vitro* related to cytotoxicity of non-mutagenic chemicals and metabolic poisons, Environ. Mol. Mutagen. 31 (1998) 316–326.

[30] R.D. Curren, G.C. Mun, D.P. Gibson, M.J. Aardema, Development of a method for assessing micronucleus induction in a 3D human skin model (EpiDerm^{TM}), Mutat. Res. 607 (2006) 192–204.

[31] H.R. Glatt, I. Gemperlein, F. Setiabudi, K.-L. Platt, F. Oesch, Expression of xenobiotic metabolizing enzymes in propagatable cell cultures and induction of micronuclei by 13 compounds, Mutagenesis 5 (1990) 241–249.

[32] M.S. Yates, M.K. Kwak, P.A. Egner, J.D. Groopman, S. Bodreddigari, T.R. Sutter, K.J. Baumgartner, B.D. Roebuck, K.T. Liby, M.M. Yore, T. Honda, G.W. Gribble, M.B. Sporn, T.W. Kensler, Potent protection against aflatoxin-induced tumorigenesis through induction of Nrf2-regulated pathways by the triterpenoid 1-[2-cyano-2,12-dioxooleana-1,9(11)-dien-2,8-oyl]imidazole, Cancer Res. 66 (2006) 2488–2494.

[33] L. Müller, R.J. Mauthe, C.M. Riley, M.M. Andino, D. De Antonis, C. Beels, J. DeGeorge, A.G.M. De Knaep, D. Ellison, J.A. Fagerland, R. Frank, B. Fritschel, S. Galloway, E. Harpur, C.D.N. Humfrey, A.S. Jacks, N. Jagota, J. Mackinnon, G. Mohan, D.K. Ness, M.R. O'Donovan, M.D. Smith, G. Vudathala, L. Yotti, A rationale for determining, testing, and controlling specific impurities in pharmaceuticals that possess the potential for genotoxicity, Regul. Toxicol. Pharmacol. 44 (2006) 198–211.

[34] L.-Q. Wang, M.O. James, Inhibition of sulfotransferases by xenobiotics, Curr. Drug Metab. 7 (2006) 83–104.

[35] H.-Z. Bu, A literature review of enzyme kinetic parameters for CYP3A4-mediated metabolic reactions of 113 drugs in human liver microsomes: structure–kinetics relationship assessment, Curr. Drug Metab. 7 (2006) 231–249.

[36] F.P. Guengerich, Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity, Chem. Res. Toxicol. 14 (2001) 611–650.

[37] K. Bergman, L. Müller, S. Weberg Teigen, The genotoxicity and carcinogenicity of paracetamol: a regulatory (re)view, Mutat. Res. 349 (1996) 263–288.

[38] C. Aninat, A. Piton, D. Glaise, T. Le Charpentier, S. Langouet, F. Morel, C. Guguen-Guillouzo, A. Guillouzo, Expression of cytochromes P450, conjugating enzymes and nuclear receptors in human hepatoma HepaRG cells, Drug Metab. Dispos. 34 (2006) 75–83.

[39] C.J. Marek, G.A. Cameron, L.J. Elrick, G.M. Hawksworth, M.C. Wright, Generation of hepatocytes expressing functional cytochromes P450 from a pancreatic progenitor cell line *in vitro*, Biochem. J. 370 (2003) 763–769.

[40] Z.D. Burke, C.N. Shen, K.L. Ralphs, D. Tosh, Characterization of liver function in transdifferentiated hepatocytes, J. Cell. Physiol. 206 (2006) 147–159.

Available online at www.sciencedirect.com
ScienceDirect
Mutation Research 634 (2007) 235–240
Genetic Toxicology and Environmental Mutagenesis
www.elsevier.com/locate/gentox
Community address: www.elsevier.com/locate/mutres

Short communication

Sensitivity of the erythrocyte micronucleus assay: Dependence on number of cells scored and inter-animal variability

Grace E. Kissling [a,*], Stephen D. Dertinger [b], Makoto Hayashi [c], James T. MacGregor [d]

[a] *National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, United States*
[b] *Litron Laboratories, Rochester, NY 14623, United States*
[c] *National Institute of Health Sciences, Tokyo 158-8501, Japan*
[d] *Toxicology Consulting Services, Arnold, MD 21012, United States*

Received 21 May 2007; received in revised form 27 July 2007; accepted 30 July 2007
Available online 6 August 2007

Abstract

Until recently, the *in vivo* erythrocyte micronucleus assay has been scored using microscopy. Because the frequency of micronucleated cells is typically low, cell counts are subject to substantial binomial counting error. Counting error, along with inter-animal variability, limit the sensitivity of this assay. Recently, flow cytometric methods have been developed for scoring micronucleated erythrocytes and these methods enable many more cells to be evaluated than is possible with microscopic scoring. Using typical spontaneous micronucleus frequencies reported in mice, rats, and dogs we calculate the counting error associated with the frequency of micronucleated reticulocytes as a function of the number of reticulocytes scored. We compare this counting error with the inter-animal variability determined by flow cytometric scoring of sufficient numbers of cells to assure that the counting error is less than the inter-animal variability, and calculate the minimum increases in micronucleus frequency that can be detected as a function of the number of cells scored. The data show that current regulatory guidelines allow low power of the test when spontaneous frequencies are low (e.g., $\leq$0.1%). Tables and formulas are presented that provide the necessary numbers of cells that must be scored to meet the recommendation of the International Working Group on Genotoxicity Testing that sufficient cells be scored to reduce counting error to less than the inter-animal variability, thereby maintaining a more uniform power of detection of increased micronucleus frequencies across laboratories and species.

Keywords: Erythrocyte micronucleus assay; Flow cytometry; Binomial counting error; Inter-animal variability; Power calculation

1. Introduction

The ability of the *in vivo* erythrocyte micronucleus assay to detect small increases in the spontaneous background frequency of micronucleated cells in a group of animals (or study subjects) is limited by either the binomial counting error, when the number of cells scored gives small numbers of scored events (micronucleated cells) [1–4], or by inter-animal variation, when that variation is so large that it obscures a small, but real, increase. Furthermore, the sensitivity to detect small increases in the micronucleus frequency in an individual animal is limited at small micronucleus counts by the binomial counting error and at higher counts by the spontaneous variability in that individual animal over the time span in

* Corresponding author. Tel.: +1 919 541 1756; fax: +1 919 541 4311.
E-mail address: kissling@niehs.nih.gov (G.E. Kissling).

1383-5718/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.mrgentox.2007.07.010

which measurements are made. Recognizing these facts, the working group on the *in vivo* micronucleus assay organized by the International Workshops on Genotoxicity Testing (IWGT) has recommended that, whenever possible, sufficient cells should be scored to reduce the counting error to less than the variability in MN frequency between individual animals (for comparison of values in different treated groups) [5].

Prior to the development of flow cytometric scoring methods, the number of cells scored was generally limited by the practical consideration of the number of cells that could be scored in a reasonable period of time by a microscopist, and therefore the minimum number of cells recommended to be scored in current regulatory guidelines is generally less than that required to discern differences between individual animals. Flow cytometric methodologies now make it practical to reduce the counting error to very small values [6–8], allowing, for the first time, reliable determination of the inter- and intra-animal variation in the spontaneous micronucleus frequency.

We summarize below experimentally determined mean and variability among animals in the spontaneous frequency of micronucleated reticulocytes (MN-RETs) in peripheral blood reticulocytes in the Sprague–Dawley rat, CD-1 mouse, and beagle dog, and in bone marrow reticulocytes in the Sprague–Dawley rat, and compare this inter-animal variability with the microscopic counting error associated with the current regulatory recommendations for scoring bone marrow or peripheral blood reticulocytes in these species. From these values, we determine the minimum increase in group mean frequencies of MN-RETs that can be detected in these species, tabulate the minimum increases that can be detected as a function of the number of RETs scored, and identify the numbers of cells that need to be scored to meet the IWGT recommendation that sufficient cells should be scored such that the error in individual animal MN-RET frequencies is less than the inter-animal variability.

2. Inter-animal variability

The inter-animal variability of the percentage of MN-RETs among RETs (no. of MN-RET/no. of RETs scored × 100) in the peripheral blood of Sprague–Dawley rats, CD-1 mice, and purpose-bred beagle dogs, and also in the bone marrow of Sprague–Dawley rats after removal of nucleated cells on a cellulose mini-column as described by Romagna [9] and Weiner et al. [10], was estimated by scoring 20,000 reticulocytes using the flow cytometric method described by Dertinger et al. [11–13]. These data are summarized in Table 1. The data are taken from

Table 1
Mean and inter-animal variation of the micronucleated reticulocyte frequency in the peripheral blood (PB) and bone marrow (BM) of Sprague–Dawley rats, Swiss mice, and Beagle dogs

Species	Strain/breed	Tissue	Mean %MN-RET	S.D. of %MN-RET	Inter-animal %CV[a] (%)	No. of animals	No. of experiments.	References
Rat	SD	PB	0.11	0.045	41	15	3	[14]
Rat	SD	BM	0.23[b]	0.059[b]	26	190	38	[21] Fiedler, personal communication
Mouse	CD-1	PB	0.20	0.070	35	79	9	[15]
Dog	Beagle	PB	0.31	0.092	30	22	4	Manuscript in preparation

[a] %CV = S.D./mean × 100%.

[b] Bone marrow %MN-RET values were determined by separation of nucleated cells on a mini-cellulose column [9,22], with subsequent scoring of the MN-RET frequency among 20,000 RETs by the same flow cytometric procedure used for analysis of peripheral blood.

previously reported studies in these species [14,15]. Details of the experimental methodology are reported in the previous publications. As is discussed below, scoring 20,000 RETs results in a sufficient number of events (MN-RETs) that the error associated with individual animals does not exceed approximately 50% of the inter-animal variability of spontaneous MN-RET frequencies in the respective species. The inter-animal % coefficients of variation (%CV = S.D. / mean × 100%) of the MN-RET frequencies were 41% for the rat, 35% for the mouse, and 30% for the dog.

Table 2 presents the binomial error in the count of MN cells in an individual animal obtained by scoring 2000, 4000, 8000, or 20,000 RETs as a function of the spontaneous frequency of MN-RETs. It should be noted that the spontaneous frequency in rodent bone marrow or peripheral blood reported by different experienced laboratories has ranged from 0.05% in rat (see individual laboratory values in Ref. [14]) to a mean value of 0.2% in the mouse [15,16] and 0.31% in the beagle dog. Since the counting error depends on the background rate and the number of cells scored, we have tabulated values over the range of spontaneous frequencies commonly reported in rodents and recently observed in the beagle dog (manuscript in preparation). As can be seen in Table 2, when 2000 cells are scored (the recommended number in the current OECD, FDA, and EPA regulatory guidelines [17–19]) the error in the counts observed in individual animals is substantially greater than the variation between animals (Table 1). When the spontaneous frequency is 0.1%, approximately 6000 cells would need to be scored to reduce the error in the individual animal count to less than the inter-animal variability observed in the rat.

Table 2
Counting error (standard deviation (S.D.) and coefficient of variation) of individual animal values of MN-RET frequency as a function of true spontaneous frequency and number of RETs scored

True %MN-RET	No. of RETs scored per animal	S.D. of count	Counting error %CV
0.05	2,000	0.050	100
	4,000	0.035	71
	8,000	0.025	50
	20,000	0.016	32
0.10	2,000	0.071	71
	4,000	0.050	50
	8,000	0.035	35
	20,000	0.022	22
0.20	2,000	0.100	50
	4,000	0.071	35
	8,000	0.050	25
	20,000	0.032	16
0.30	2,000	0.122	41
	4,000	0.086	29
	8,000	0.061	20
	20,000	0.039	13

3. Sensitivity to increases above the spontaneous frequency

Table 3 summarizes the minimum increases above the spontaneous frequency that can be detected in groups of five animals (the minimum currently recommended in OECD, FDA, and EPA guidelines [17–19]) as a function of the number of target cells scored (in this case RETs) and observed spontaneous frequency (in this case %MN-RETs among RETs). Minimum detectable increases in MN-RET frequencies at $p \leq 0.05$ or ≤ 0.01, with 90% or 95% power were determined using Monte Carlo simulations. Specifically, to reflect inter-animal variability, five binomial probabilities were randomly selected from a normal distribution with the following mean, μ_0, and standard deviation, σ, combinations: $(\mu_0, \sigma) = (0.05\%, 0.02\%)$, $(0.10\%, 0.045\%)$, $(0.20\%, 0.070\%)$, or $(0.30\%, 0.092\%)$. For a given fold-increase, f, a second set of five binomial probabilities were randomly selected from a normal distribution with mean, $\mu_1 = \mu_0 f$, and the same σ given above. Using the five binomial probabilities from the spontaneous mean group, five MN-RETs frequencies were randomly generated from binomial distributions, with n = number of RETs scored, 2000, 4000, or 20,000. Such selection from a binomial distribution introduces the binomial counting error. Five MN-RET frequencies were similarly generated using the five binomial probabilities from the increased mean group. A one-tailed Mann–Whitney test was then performed on these 10 counts, comparing the spontaneous group to the increased group, and the p-value was noted as to whether it was 0.05 or less and/or 0.01 or less. This was repeated 3000 times and the percentages of the 3000 'samples' for which the p-value was 0.05 or less and 0.01 or less were calculated. The process was repeated over a series of increases, f, at increments of 0.1, to determine the first point at which the power exceeded 90 or 95%. We obtained very similar results (not shown) by generating the five binomial probabilities from beta distributions having the above combinations of μ_0, μ_1, and σ.

For the line labeled "∞" in Table 3, there is no counting error; rather, the variability in frequencies is due to inter-animal variation alone. If we assume that inter-animal variation is normally distributed, the minimum difference between μ_1 and μ_0, $\delta = \mu_1 - \mu_0$, detectable using five animals per group with significance level α

Table 3
Minimum detectable increases in MN-RET frequency in groups of five animals as a function of spontaneous frequency and number of RETs scored

Spontaneous frequency (%MN-RET)	No. of RETs 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

Values for the cases of infinite cell counts are calculated based on the observed inter-animal variability (standard deviation from Table 1) for the species stated, assuming no counting error; the inter-animal variability for frequency 0.05% is assumed to be 0.02%. The detectable increase depends on the relative magnitudes of both the counting error and the inter-animal variability. Although counting error can be reduced by scoring more RETs, the minimum detectable increase cannot go below a bound determined by the inter-animal variability (i.e., the value given in the infinite cell count rows). Species entries correspond to the approximate spontaneous frequency and associated inter-animal standard deviation in the species specified in Table 1.

and power $1-\beta$ is [20]:

$$\delta = (t_\alpha + t_\beta)\sigma\sqrt{\frac{2}{5}}.$$

Here, t_α and t_β are the critical values from the $5+5-2=8$ degree of freedom t-distribution having upper tail probabilities α and β, respectively. The minimum detectable fold-increase over the spontaneous group is then

$$f = \frac{\mu_1}{\mu_0} = \frac{\mu_0 + \delta}{\mu_0} = 1 + \frac{\delta}{\mu_0}.$$

While spontaneous MN-RET frequencies determined from counting 2000 RETs from different animals are not often normally distributed, it has been our experience that spontaneous frequencies determined from counting 20,000 RETs from different animals are approximately normally distributed. Therefore, the assumptions of normality that we made above are most likely reasonable.

It should be noted that even if the counting error of the MN-RET frequency in each individual animal could be eliminated, the sensitivity of detection of changes in the observed mean group frequency would still be limited by the inter-animal variability (represented in Table 3 by the line in which an infinite number of cells is scored). It is clear that the regulatory assay as currently conducted is relatively insensitive to changes in the spontaneous frequency, especially when the spontaneous frequency is low. For example, when the spontaneous frequency is 0.05% and only 2000 RETs are scored, even a 6.8-fold increase would fail to be detected at a confidence level of $p \leq 0.01$ in 10% of experiments conducted. Even at the more commonly reported spontaneous frequency of 0.1% a 4.8-fold increase would fail to be detected 10% of the time at this same confidence level. The use of flow cytometric scoring to achieve a sufficient cell count to allow individual animal frequencies with adequate certainty (i.e., certainty of the individual value relative to the inter-animal variation) would increase the sensitivity

Table 4
Number of reticulocytes required to be scored to reduce counting error to less than the observed inter-animal coefficient of variation

Spontaneous frequency (% MN-RET)	Species/tissue with this approximate spontaneous frequency	Inter-animal %CV[a]	No. of RETs 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
0.05	Rat BM & PB(microscopy, some reports)	NA[b]	NA	NA	NA
0.10	Rat PB (data cited above)	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

[a] Experimentally determined inter-animal %CV by flow cytometric scoring of 20,000 peripheral blood RETs, at the approximate spontaneous frequency tabulated.

[b] Inter-animal %CV has not been determined at the spontaneous frequency of 0.05%; no reported experiments have scored sufficient cells to determine the inter-animal variability.

such that a doubling of a spontaneous frequency of 0.1% among 20,000 RETs scored would be detected nearly 90% of the time at a confidence level of $p \leq 0.05$. It should also be noted that, regardless of the spontaneous frequency, the sensitivity achieved by scoring 20,000 RETs is close to the optimal sensitivity that could be achieved if no counting error were present.

4. Number of reticulocytes required to be scored to reduce counting error to less than inter-animal variability

Table 4 summarizes the number of cells required to be scored to reduce the counting error of individual animal values (Table 2, %CV) to the observed inter-animal variation or less (Table 1, %CV). These numbers were calculated by setting a multiple ($m = 1.0$, 0.5, or 0.2) of the inter-animal %CV equal to the binomial counting error %CV and solving for the required sample size, n. Mathematically, if p is the percent of MN-RETs among all RETs within an animal, then

$$\%\mathrm{CV}_{\text{binomial error}} = \frac{\sqrt{p(1-p)/n}}{p} = m \times \%\mathrm{CV}_{\text{inter-animal}}.$$

Solving for n, we get

$$n = \frac{1-p}{p(m \times \%\mathrm{CV}_{\text{inter-animal}})^2}.$$

The numbers of RETs required are prohibitively laborious to obtain by conventional microscopic scoring, but are easily achieved by automated procedures such as flow cytometry.

Acknowledgements

We thank Ronald Fiedler of Pfizer, Inc. for providing the data on spontaneous frequencies and inter-animal variability of the frequency of micronucleated reticulocytes in the bone marrow of Sprague–Dawley rats. This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

References

[1] B.H. Margolin, B.J. Collings, J.M. Mason, Statistical analysis and sample-size determinations for mutagenicity experiments with binomial responses, Environ. Mutagen. 5 (1983) 705–716.

[2] B.H. Margolin, K.J. Risko, The statistical analysis of in vivo genotoxicity data: case studies of the rat hepatocyte UDS and mouse bone marrow micronucleus assays, in: J. Ashby, F.J. de Serres, M.D. Shelby, B.H. Margolin, M. Ishidate Jr., G.C. Becking (Eds.), Evaluation of Short-term Tests for Carcinogens: Report of the International Programme on Chemical Safety's Collaborative Study on In Vivo Assays, vol. 1, Cambridge University Press on Behalf of WHO, Cambridge, 1988, pp. 1.29–1.42.

[3] M.A. Kastenbaum, K.O. Bowman, Tables for determining the statistical significance of mutation frequencies, Mutat. Res. 9 (1970) 527–549.

[4] M. Hayashi, I. Yoshimura, T. Sofuni, M. Ishidate Jr., A procedure for data analysis of the rodent micronucleus test involving a historical control, Environ. Mol. Mutagen. 13 (1989) 347–359.

[5] M. Hayashi, J.T. MacGregor, D.G. Gatehouse, D.H. Blakey, S.D. Dertinger, L. Abramsson-Zetterberg, G. Krishna, T. Morita, A. Russo, N. Asano, H. Suzuki, W. Ohyama, D. Gibson, *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 (2007) 10–30.

[6] N. Asano, D. Torous, C. Tometsko, S. Dertinger, T. Morita, M. Hayashi, Practical threshold for micronucleated reticulocyte

induction for low doses of mitomycin C, Ara-C and colchicine, Mutagenesis 21 (2006) 15–20.
[7] J. Grawe, L. Abramsson-Zetterberg, G. Zetterberg, Low dose effects of chemicals as assessed by the flow cytometric in vivo micronucleus assay, Mutat. Res. 405 (1998) 199–208.
[8] D. Torous, N. Asano, C. Tometsko, S. Sugunan, S. Dertinger, T. Morita, Performance of flow cytometric analysis for the micronucleus assay—a reconstruction model using serial dilutions of malaria-infected cells with normal mouse peripheral blood, Mutagenesis 21 (2006) 11–13.
[9] F. Romagna, Current issues in mutagenesis and carcinogenesis; fractionation of a pure PE and NE population from rodent bone marrow, Mutat. Res. 206 (1988) 307–309.
[10] S.K. Weiner, R.D. Fiedler, M.J. Schuler, Development and evaluation of a flow cytometric method for the analysis of micronuclei in rat bone marrow in vivo, Environ. Mol. Mutagen. 43 (2004) 236.
[11] S.D. Dertinger, D.K. Torous, K. Tometsko, Simple and reliable enumeration of micronucleated reticulocytes with a single-laser flow cytometer, Mutat. Res. 371 (1996) 283–292.
[12] S.D. Dertinger, D.K. Torous, N.E. Hall, C.R. Tometsko, T.A. Gasiewicz, Malaria-infected erythrocytes serve as biological standards to ensure reliable and consistent scoring of micronucleated erythrocytes by flow cytometry, Mutat. Res. 464 (2000) 195–200.
[13] S.D. Dertinger, K. Camphausen, J.T. MacGregor, M.E. Bishop, D.K. Torous, S. Avlasevich, S. Cairns, C.R. Tometsko, C. Menard, T. Muanza, Y. Chen, R.K. Miller, K. Cederbrant, K. Sandelin, I. Ponten, G. Bolcsfoldi, Three-color labeling method for flow cytometric measurement of cytogenetic damage in rodent and human blood, Environ. Mol. Mutagen. 44 (2004) 427–435.
[14] J.T. MacGregor, M.E. Bishop, J.P. McNamee, M. Hayashi, N. Asano, A. Wakata, M. Nakajima, A. Aidoo, M.M. Moore, S.D. Dertinger, Flow cytometric analysis of micronuclei in peripheral blood reticulocytes II. An efficient method of monitoring chromosomal damage in the rat, Toxicol. Sci. 94 (2006) 92–107.
[15] D.K. Torous, N.E. Hall, A.H. Illi-Love, M.S. Diehl, K. Cederbrant, K. Sandelin, I. Pontén, G. Bolcsfoldi, L.R. Ferguson, A. Pearson, J.B. Majeska, J.P. Tarca, G.M. Hynes, A.M. Lynch, J.P. McNamee, P.V. Bellier, M. Parenteau, D. Blakey, J. Bayley, B.M. van der Leede, P. Vanparys, P.R. Harbach, S. Zhao, A.L. Filipunas, C.W. Johnson, C.R. Tometsko, S.D. Dertinger, Interlaboratory validation of a CD71-based flow cytometric method (MicroFlow®) for the scoring of micronucleated reticulocytes in mouse peripheral blood, Environ. Mol. Mutagen. 45 (2005) 44–55.
[16] J.A. Heddle, M.F. Salamone, M. Hite, B. Kirkhart, K. Mavournin, J.T. MacGregor, G.W. Newell, The induction of micronuclei as a measure of genotoxicity, Mutat. Res. 123 (1983) 61–118.
[17] OECD, Guideline for the testing of chemicals. Mammalian erythrocyte micronucleus test. Guideline 474, July 1997.
[18] FDA (US Food and Drug Administration), Office of Food Additive Safety, Redbook 2000, Toxicological principles for safety assessment of food ingredients, Updated November 2003, www.cfsan.fda.gov/~redbook/red-toca.html.
[19] EPA (US Environmental Protection Agency), Health Effects Test Guidelines OPPTS 870.5395, Mammalian Erythrocyte Micronucleus Test, Office of Prevention, Pesticides and Toxic Substances (7101) EPA 712-C-98-226, August 1998. www.epa.gov/opptsfrs/publications/OPPTS_Harmonized/870_Health_Effects_Test_Guidelines/Series/870-5395.pdf.
[20] G.W. Snedecor, W.G. Cochran, Statistical Methods, 6th ed., The Iowa State University Press, Ames, IA, 1976, pp. 113–116.
[21] R. Fiedler, Pfizer, Inc., Personal communication (2007).
[22] E. Beutler, T. Gelbart, The mechanism of removal of leukocytes by cellulose columns, Blood Cells 12 (1986) 57–64.

The Journal of Toxicological Sciences (J. Toxicol. Sci.)
Vol.33, No.1, 97-104, 2008

Original Article

Evaluation of statistical tools used in short-term repeated dose administration toxicity studies with rodents

Katsumi Kobayashi[1], K. Sadasivan Pillai[2], Yuki Sakuratani[1], Takemaru Abe[1], Eiichi Kamata[3] and Makoto Hayashi[3]

[1]Chemical Management Center, National Institute of Technology Evaluation, 49-10 Nishihara-Nichome, Shibuya, Tokyo 151-0066, Japan
[2]Orchid Research Laboratories Ltd., R&D Centre, Plot No. 476/14, Old Mahabalipuram Road, Sholinganallur, Chennai 600119, India
[3]National Institute of Health Sciences 18-1 Kamiyoga-1chome, Setagaya-ku, Tokyo 158-8501, Japan

(Received November 20, 2007; Accepted November 27, 2007)

ABSTRACT — In order to know the different statistical tools used to analyze the data obtained from twenty-eight-day repeated dose oral toxicity studies with rodents and the impact of these statistical tools on interpretation of data obtained from the studies, study reports of 122 numbers of twenty-eight-day repeated dose oral toxicity studies conducted in rats were examined. It was found that both complex and easy routes of decision trees were followed for the analysis of the quantitative data. These tools include Scheffe's test, non-parametric type Dunnett's and Scheffe's tests with very low power. Few studies used the non-parametric Dunnett type test and Mann-Whitney's *U* test. Though Chi-square and Fisher's tests are widely used for analysis of qualitative data, their sensitivity to detect a treatment-related effect is questionable. Mann-Whitney's *U* test has better sensitivity to analyze qualitative data than the chi-square and Fisher's tests. We propose Dunnett's test for analysis of quantitative data obtained from twenty-eight-day repeated dose oral toxicity tests and for qualitative data, Mann-Whitney's *U* test. For both tests, one-sided test with p=0.05 may be applied.

Key words: Statistics; 28-day repeated toxicity study; Rodents; Dunnett's test; Mann-Whitney's *U* test

INTRODUCTION

Short-term repeated oral toxicity study conducted for 14 or 28 days is aimed to (1) predict appropriate doses of test substance for future subchronic or chronic toxicity studies, (2) determine NOELs for some toxicology endpoints and (3) to allow future studies in rodents to be designed with special emphasis on identified target organs (USFDA, 2000). This study also provides information on the possible health hazards likely to arise from repeated exposure over a relatively limited period of time (USEPA, 2000; OECD, 1995). Though these guidelines provide all the information required for the conduct of the study, no information is provided on the appropriate statistical tools to be used to analyze the data obtained from the study. Use of right statistical tool to analyze the data obtained from theses studies is very crucial as the interpretation of the data is mostly based on the results of the statistical analysis.

The statistical tools used to analyze the data obtained from 122 numbers of twenty-eight-day repeated dose oral toxicity tests in rats were examined in the present study. The objective of the study was to know the different statistical tools that are used in these studies and the possible impact of these statistical tools on interpretation of the data. A brief discussion on the use and the property of the different statistical tools used in the studies are also given. The purpose of this article wished for the standardization of statistics and the analysis methods. Finally, the authors made an attempt to suggest statistical techniques that may best suit twenty-eight-day repeated dose oral toxicity studies in rodents.

Correspondence: Katsumi Kobayashi (E-mail: kobayashi-katsumi@nite.go.jp)

MATERIALS AND METHODS

Studies examined

A total number of 122 studies conducted in various test facilities in Japan were examined (MHLW, 2006). The chemical of these examinations was executed with existing chemical substances by the guideline of the Chemical Substance Control Law (1986). The number of studies conducted in each test facility is given in parenthesis: Food and Drug Safety Center, Kanagawa (22), An-Pyo Center, Shizuoka (22), Mitsubishi Chemical Safety Institute Ltd., Ibaraki (18), Safety Research Institute for Chemical Compounds Co., LTD, Hokkaido (15), Bozo Research Center Inc., Shizuoka (12), Research Institute for Animal Science in Biochemistry & Toxicology, Kanagawa (11), Panapharm Laboratories, Kumamoto (10), Nihon Bioresearch Inc., Gifu (9) and National Institutes of Health, Tokyo (3).

Quantitative and qualitative items

Several quantitative and qualitative items are evaluated in twenty-eight-day repeated dose oral toxicity tests in rats, as per the regulatory guidelines. The quantitative items that require statistical analysis are body weight, food consumption, water consumption, leucocytes, erythrocytes, hemoglobin, hematocrit, platelets, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, differential leucocyte counts, prothrombin time, activated partial thromboplastin time, total protein, albumin, albumin/globulin ratio, total bilirubin, alanine aminotransferase, aspartate aminotransferase, γ-glutamic transaminase, alkaline phosphatase, acetylcholinesterase, total cholesterol, tryglycerides, phospholipids, glucose, blood urea nitrogen, creatinine, inorganic phosphorous, calcium, sodium, potassium, chlorides, urine volume, specific gravity of urine, absolute organ weights and relative organ weights. Qualitative items that require statistical analysis are mortality, functional observation battery, clinical signs, urinalysis (color, pH, protein, glucose, ketone bodies, bilirubin, occult blood, urobilinogen, epithelial cells, erythrocytes, leucocytes, casts and crystals) and pathological findings (macroscopic and microscopic). But the regulatory guidelines do not indicate the specific statistical techniques to be used to analyze these data.

Which test to be used - One-sided or two-sided?

When the *t*-test and Dunnett's multiple comparison test (Dunnett's test) are used, the significant difference detection rate of a two-sided test is about 85% as compared with a one-sided test (Kobayashi, 1997a). In toxicological studies, usually a dosed group is compared with the control group. For this comparison, one-sided test is ideal, hence Yoshimura and Ohashi (1992) recommend the one-sided test for comparing a dosed group with the control group.

Is analysis of variance (ANOVA) necessary?

It is a common practice to subject the data, if they are from more than two groups, ANOVA. If ANOVA shows a significant difference among the groups, multiple comparison tests are used to find the significant difference between any two groups. In recent years, several authors suggested that the error of the second kind can be prevented by carrying out direct multiple comparison tests, without subjecting the data to ANOVA (Hamada *et al.*, 1998; Kobayashi *et al.*, 2000a; Sakaki *et al.*, 2000). It may be worth mention in this context that Dunnett (1964) did not recommend ANOVA prior to multiple comparison tests.

Is Bartlett's homogeneity test necessary?

Generally Bartlett's test is used to examine the homogeneity of variance if the number of animals in a group is 10 or more. Therefore, this test is not used in the toxicity studies with dogs, where the number of animals in the group is less. According to Kobayashi *et al.* (1998), Bartlett's test is not required to examine the homogeneity of variance, when the number of animals in a group is less.

Non-parametric type Dunnett's test

The non-parametric Dunnett's multiple comparison test has two techniques - 'joint type' and 'separate type' or Steel's test. When the Steel's test shows the highest dosage correlation, the number of animals required in the dosage groups to detect a significant difference in the low dosage group is four (Inaba, 1994; Kobayashi *et al.*, 1995). On the contrary, 'joint type' needs 15 animals in each group.

Transformation of data

If the data show heterogeneity of variance as per Bartlett's test, sometimes the data are transformed, for example to logarithmic values and then they are subjected to non-parametric tests. According to Finney (1995), "when a scientist measures a quantity such as concentration of a chemical compound in body fluid, his interest usually lies in the scale, perhaps mg/ml, that he has used; he is less likely to be interested in a summary of results relating to a transformed quantity such as the logarithm of blood concentration. If he analyzes in terms of logarithms, encouraged perhaps by an elementary but uncritical statistical textbook or by a convenient software package, he may find significant differences but to express his conclusions in meaningful numbers may be impossible. I do not assert

that a scientist should never transform data before analysis; I urge that data should be transformed only after careful consideration of all consequence". Therefore, transformation should be done cautiously.

Power of Scheffe's test

Use of Scheffé's test is discouraged in recent years because this test may not show a significant difference in the dosage groups even if the dosage groups show a difference of 60-53% compared to control group (Kobayashi *et al.*, 1997b).

Power of non-parametric tests using ranked data

In four groups setting with the highest dosage correlation, the minimum numbers of animals required in the low-dose group to detect a significant difference, compared to control, using the statistical tools of Scheffé's type, Dunn's test, Tukey type, Dunnett type, Williams-Wilcoxon test, Steel test and Mann-Whitney's *U* test are 22, 19, 18, 15, 8, 4 and 3, respectively. Therefore, in the twenty-eight-day repeated dose oral toxicity tests in rats, where the number of animals is 5/sex/group, except Steel and Mann-Whitney's *U* tests, other tests are not used. Inaba (1994) made a similar observation on the power of the above tests.

Power of Chi-square and Fisher's tests

When a finding in the animals of a control group is 0, in order to find a significant difference of the finding between the control group (n=5) and dosage group (n=5) by chi-square test, all the 5 animals in the dosage group (n=5) should show the finding, whereas by Fisher's test 4 animals should show the finding. When 1 animal in the control group shows a finding, even if the finding is seen in all the animals in the dosage group, a significant difference is not detected by chi-square test, but it is detected by Fisher's test. In the light of the above it may be stated that power of one-sided Fisher's test is better than the Chi-square test.

Dunnett's test is the expanded version of *t*-tests

Dunnett's test becomes *t*-test when two groups are analyzed (Kobayashi *et al.*, 1997c). Therefore, when comparing the recovery groups in the twenty-eight-day repeated dose oral toxicity tests in rats, where number of the groups is 2, it does not make any difference, whether the analysis is carried out by Dunnett test or *t*-test.

Power of Mann-Whitney's *U* test

This test is generally used for the analysis of pathology data (Kobayashi *et al.*, 1997d). A significant difference by a one-sided test is detected if the calculated *U* value is four or less. Since one-side is expected in studies like twenty-eight-day repeated dose oral toxicity tests in rats, a one-sided Mann-Whitney's *U* test is used to analyze pathology data obtained from these studies.

RESULTS

Quantitative data

Out of 122 studies examined, 79 studies used statistical tools that follow a complicated course (tool numbers; 2, 3, 4, 5, 8, 9, 10, 12, 15, 16 and17) and 43 studies used statistical tools that follow simple course (tool numbers; 1, 6, 7, 11, 13 and14) (Table 1; Fig. 1). The statistical tools describing the method of analyzes, in the case of three or more groups and two groups were mentioned in 6 studies, whereas this description was not found in 11 studies. Only eight studies used trend test (Jonckheere, 1954). In the tool number 10, the significance level of ANOVA and Kruskal-Walis's *H* test were set at p=0.10. For comparing with the control, this tool set the significance level of p=0.05. Tool numbers 13 and 14 did not perform Bartlett's test for testing the homogeneity of variance. Use of one-sided or two-sided test is not indicated in 87 studies. Only one study indicated use of non-parametric test.

Qualitative data

Since urinalysis data were classified into many grades, chi-square test was used to analyze these data in most of the studies. For macro- and microscopic pathological findings, Mann-Whitney's *U* test, Fisher's test and Chi-square test were used. Most of the studies did not indicate the alpha. Only the pathological findings of 3 studies were examined for dose-relationship (Table 2).

Use of a one-sided test was more common than a two-sided test in the case of analysis of both quantitative and qualitative data (Table 3).

DISCUSSION

National Toxicology Program, USA published technical reports of long-term carcinogenicity studies and short-term toxicity tests carried out with more than 500 substances in rat and mouse (NIH, 2006). Most of these studies used the statistical tools almost similar to the ones currently used to analyze the data obtained from the toxicity tests of agricultural chemicals and medical drugs (Kobayashi *et al.*, 2000b).

On examination of 122 studies, it was found that complex and easy courses of analytical techniques were used for the analysis of the quantitative data. These tools may be classified into 4 different categories. Five tools (tool

numbers; 4, 5, 8, 16 and17) are the advanced type of the algorithm, similar to the one developed by Yamazaki *et al.* (1981). These tools include Scheffé's test, non-parametric type Dunnett's and Scheffé's tests with very low power. Six tools (tool numbers; 3, 7, 9, 10, 12 and 15) are again advanced type of algorithm developed by Sano and Okayama (1990), which can be used even if the number of animals in the groups are different. Use of the non-parametric Dunnett type test with low power is also seen in few studies. Mann-Whitney's *U* test was also used (tool number; 9) in 14 studies in order to retain the power. Three tools (tool numbers; 2, 6 and 11) are an improved version of non-parametric type Dunnett's test ('joint type') and Steel's test ('separate type'). Dunnett's or Scheffé's tests is independently used for 3 tools (tool numbers; 1, 13 and 14). Though use of Scheffé's test has the advantage of comparison of groups in various combinations, for example, control+mid dose *vs.*high dose, low dose+mid dose *vs.* high dose, etc., it has extremely low detection power. Hence, this test is not widely used in recent years.

Yoshimura (1987) used Bartlett's test to analyze the difference in distribution of variance among the groups, where number of animals in the group is more than 10. The power of Bartlett's test decreases when the number of animals in the group is less.

Dunnett's test is the expanded version of *t*-tests, hence, it becomes *t*-test when two groups are analyzed by Dunnett's test. Therefore, for the comparison of two groups either Dunnett test or *t*-test can be used.

The most important purpose of applying statistical analysis in toxicity studies is to know whether the items estimated in the experimental group has increased or decreased compared to the control. Therefore, a one-sided test is used. Detection rate of two-sided test is half of the one-sided test, hence it is important to mention in the study report whether a one-sided or two-sided test is used. It may

Table 1. Classification of number of studies based on the statistical tools used for the analysis of quantitative data.

Tool. No.	Description of statistical tools	Number of studies
1	Dunnett's test: Three groups or more; Student or Aspin-Welch's *t*-test: Two groups	5
2	Bartlett's test, ANOVA, Dunnett's test, Kruskal-Walis's *H* test, Steel's test	7
3	Bartlett's test, ANOVA, Dunnett's test, Kruskal-Walis's *H* test, non-parametric type Dunnett's test: Three groups or more; Student or Aspin-Welch's *t*-test: Two groups	9
4	Bartlett's test, ANOVA, Dunnett's test, Scheffé's test, Kruskal-Walis's *H* test, Non-para type Dunnett's test, non-parametric type Scheffe's test: Three groups or more; Student or Aspin-Welch's *t*-test: Two groups	10
5	Bartlett's test, NOVA, Dunnett's test, Duncan's test, Kruskal-Walis's *H* test, non-parametric type Dunnett's test	9
6	Bartlett's test, Dunnett's test, Steel's test	20
7	Bartlett's test, Dunnett's test, non-parametric type Dunnett's test	10
8	Bartlett's test, ANOVA, Dunnett's test, Scheffé's test, Kruskal-Walis's *H* test, non-parametric type Dunnett's test, non-parametric type Scheffé's test	23
9	Bartlett's test, ANOVA, Dunnett's test, Kruskal-Walis's *H* test, Mann-Whitney's *U* test	14
10	Bartlett's test, ANOVA (p=0.10)、Dunnett's test, Kruskal-Walis's *H* test (p=0.10), Mann-Whitney's *U* test, When compared with control setting (p=0.05)	1
11	Bartlett's test, Dunnett's test, Steel's test	3
12	Bartlett's test, ANOVA, Dunnett's test, Kruskal-Walis's *H* test, non-parametric type Dunnett's test: Three groups or more; Student's *t*-test or Mann-Whitney's *U* test: Two groups	1
13	Dunnett's test: Three groups or more; *t*-test or Mann-Whitney's *U* test: Two groups	4
14	Dunnett's or Scheffé's tests: Three groups or more; *t*-test or Mann-Whitney's *U* test: Two groups	1
15	Bartlett's test, ANOVA, Dunnett's test, Kruskal-Walis's *H* test, non-parametric type Dunnett's test	3
16	Bartlett's test, ANOVA, Dunnett's test, Jaffé's test, Kruskal-Walis's *H* test, non-parametric type Dunnett's test, non-parametric type Jaffé's test	1
17	Bartlett's test, ANOVA, Dunnett's test, Scheffé's test, Kruskal-Walis's *H* test, non-parametric type Dunnett's test, non-parametric type Scheffé's test: Three groups or more; Student's *t*-test: Two groups	1
	Jonckheere's trend test (Not included in the number of tools)	8
	Total	**122**

Statistical tools used in short-term toxicity studies.

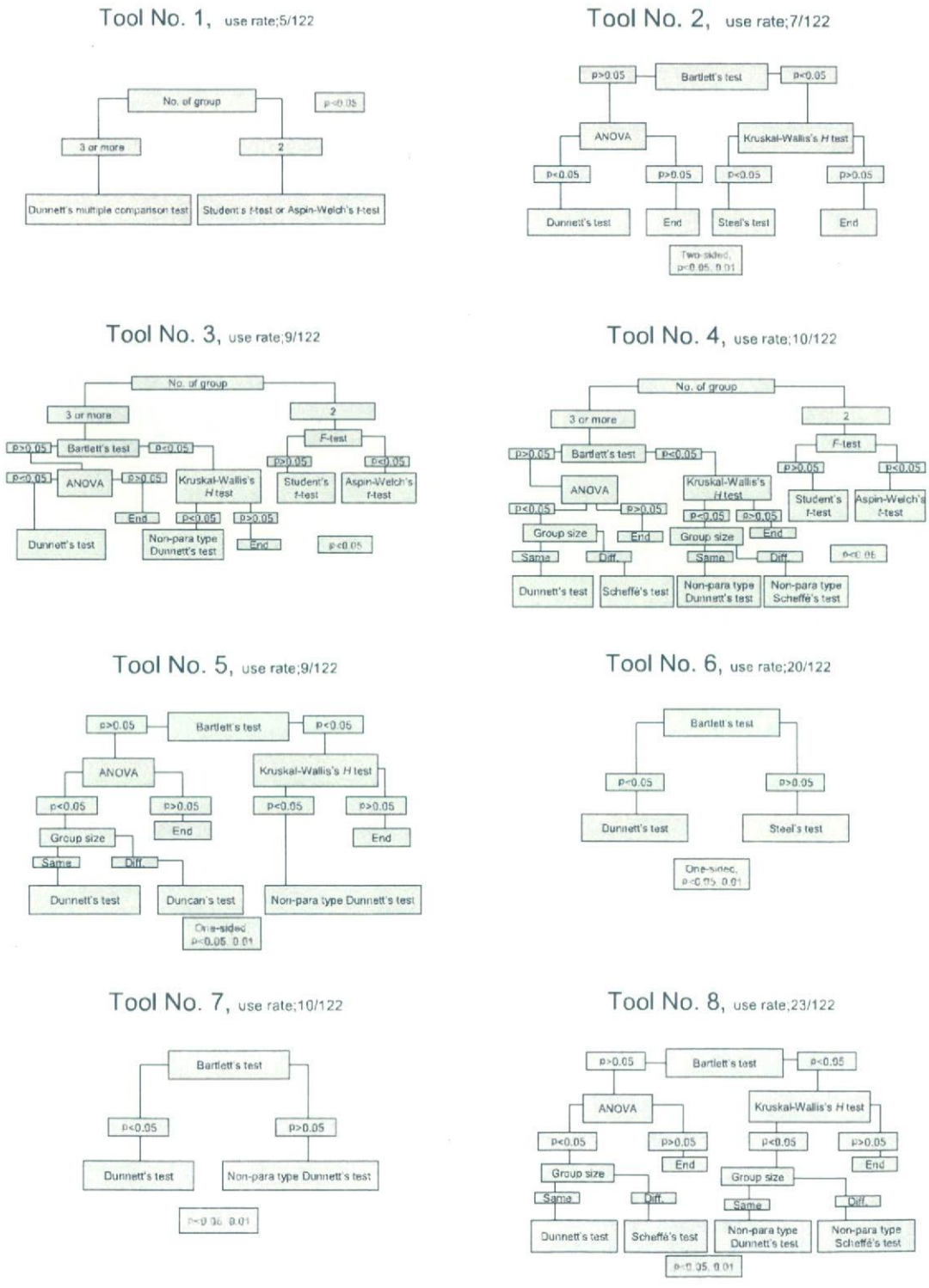

Fig. 1. Classification of number of studies based on the statistical tools used for the analysis of quantitative data.

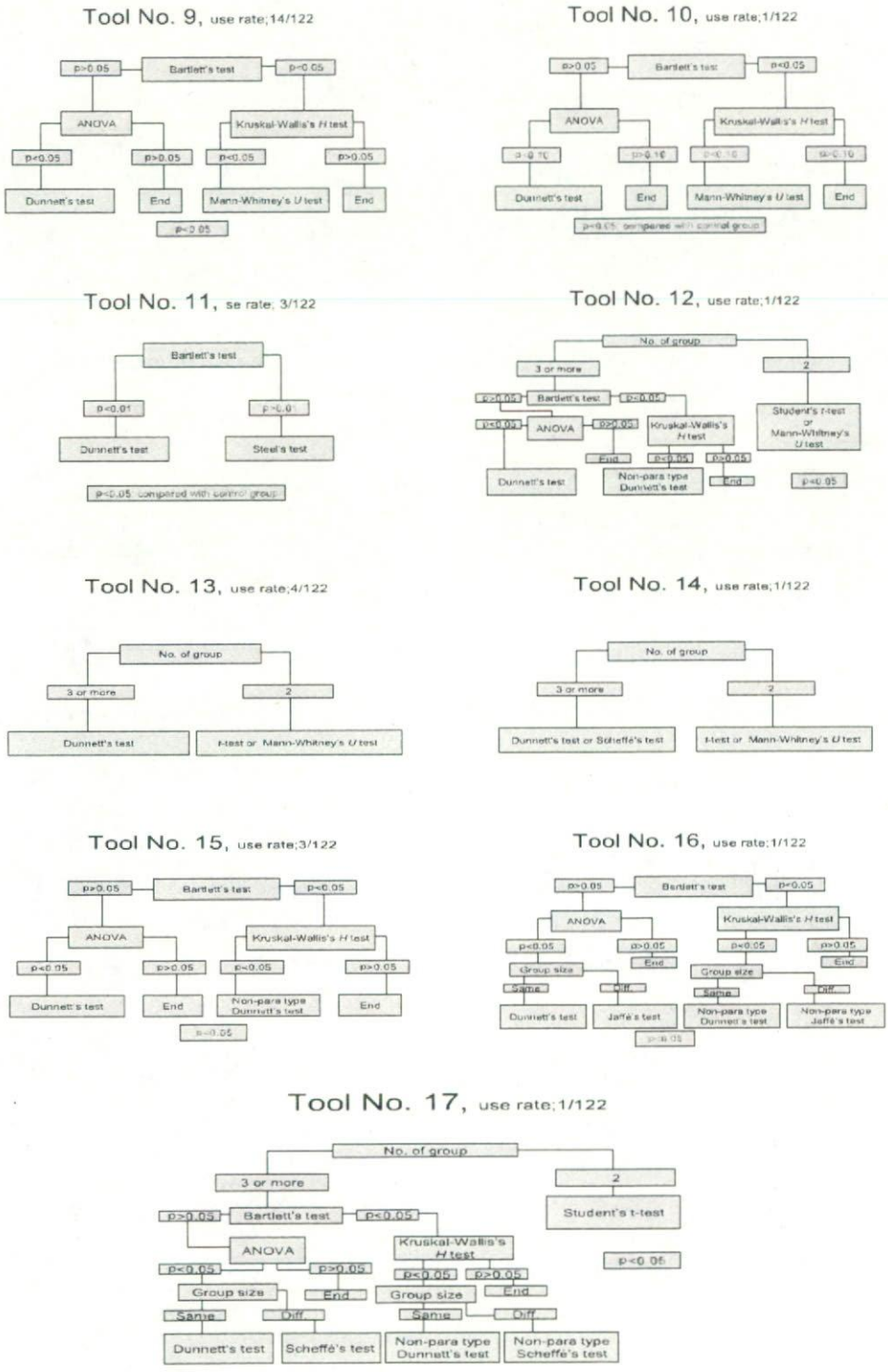

Fig. 1. Continued.