Statistical Evaluation of In Vivo Micronucleus Assays in Toxicology

John Closter F. Olivo
Central Luzon State University

Several alternative statistical procedures have been suggested and published to statistically analyze the incidence of micronucleated polychromatic erythrocytes (MNPCs) among treatment groups, but no standard procedure has been singled out and exclusively recommended. In this study, the potential of TO2 to induce chromosomal damage is tested using both Poisson and quasi-Poisson models for the statistical evaluation of *in vivo* micronucleus (MN) assay. The genotoxic activity of T02 is assessed in the rodent bone marrow micronucleus test using male mice. Results show that MN frequencies are significantly elevated in mice exposed to any dose level of T02 administered orally in a single frequency of dose. Moreover, results indicate that T02 is tested to be a positive compound under the anticipated condition of the tests used.

Keywords: in vivo, micronucleus, MNPC, Poisson model, quasi-Poisson model, TO2

1. Introduction

Genotoxicity testing is a vital element of a product safety assessment which is highly recommended by the regulatory agencies around the globe. It is designed to assess and detect chemicals that induce genetic damage. Thus, the Fourth International Conference on Harmonization (ICH4) of Genotoxicity Guidelines recommended the micronucleus assay (MNC) as one of the standard three-test batteries for genotoxicity testing of pharmaceuticals. The two most well-established tests as screening methods of new chemical entities with a widespread acceptance in industry and authorities are the *in vivo* and *in vitro* micronucleus assays.

As a measure for chromosomal aberrations *in vivo*, data on the frequency of micronucleated erythrocytes (MN) per a certain number of scored polychromatic erythrocytes (PCE) per animal are usually determined and analyzed using an

appropriate statistical test. The attention of this paper was restricted only on the statistical procedure for the analysis of micronuclei particularly on the application of methodology suggested by Hothorn and Gerhard (2008).

HG considered different aspects in the statistical evaluation of the *in vivo* micronucleus assay. One aspect concerns the choice of the experimental unit and the potential presence of overdispersion often ignored in traditional analyses. In *in vivo*, biologically speaking, the standard protocol design is basically a randomized one-way layout including a negative control, several doses or treatment groups and, optionally, a positive control. Five to ten animals are randomly assigned to each treatment group. For each animal, a number of polychromatic erythrocytes (PCEs) are scored and evaluated for the presence of small micronuclei.

As proposed by Kim et al. (2000), a formal statistical approach is to assume a binomial distribution of the number of micronuclei observed in 2,000 PCEs. This means pooling the number of MN over the animals to have one estimate of the proportion for each treatment group. These proportions are evaluated using a Dunnet-type procedure or the Cochran-Armitage trend test. However, such approach is not recommended by HG because pooling the number of MN over all animals/cultures ignores the variability between the animals as experimental units. They propose to model this between-animal variability using extra-Poisson or extra-binomial models.

The second statistical aspect is the contradiction between statistical significance and biological relevance. Traditionally, scientists rely on the p-values from their reporting systems to draw conclusions regarding the safety of a compound. P-values represent a probability of falsification but do not provide interpretation in terms of biological relevance. HG suggest the use of confidence intervals which allow interpretations of both statistical significance and biological relevance at the same time.

The third aspect is the type of inference. HG dwell upon the type of inference that is of relevance in trying to identify an increasing dose-related trend, possibly with downturn effects at high doses (Bretz and Hothorn, 2002).

Finally, HG propose a proof of safety approach for a possible claim that a compound is not genotoxic.

The purpose of this study is to implement, apply and discuss the recommendations of HG using the genotoxicity database of the test compound T02. The code name TO2 is used in this study for confidentiality purposes. Furthermore, the results are compared with the widespread "traditional" ANOVA-like approaches that completely ignore the between-animal variability, focus on p-values and linear trend tests only, and are set up in terms of proof of hazard.

2. Motivating Example

The data considered are from an independent *in vivo* assay with compound T02. Five male mice are randomly allocated in each different dose levels of

testing groups – low, medium, high and a concurrent negative control/vehicle control (VC) group. In total, 20 male mice were sacrificed. Each mouse in the vehicle control is given demineralised water while each mouse in the remaining groups is treated with a certain dose of T02. Next, blood samples are obtained for each mouse based on the harvesting time (27h) after dosing which is specified in the protocol. Micronucleus frequencies are determined for each animal by scoring 20,000 polychromatic erythrocytes (PCEs) and the micronucleus occurrence per 20,000 PCEs was recorded.

3. Statistical Methodology

The number of micronucleated erythrocytes (MN) constitute the primary endpoint. The micronucleus frequencies are determined by analyzing the number of micronuclei from at least 20,000 PCEs per animal.

3.1 Poisson model

The statistical evaluation of the number of micronuclei is primarily focused on multiple contrast tests for comparisons versus the vehicle control. This number of micronuclei as counts are analyzed using the classical approach Poisson model. This means that for each treatment group only one count is estimated by pooling the number of MN over animals. According to Parodi and Bottarelli (2006), the Poisson regression model is often applied to study the occurrence of small number of counts or events as a function of a set of predictor variables.

Let Y_{ij} be the number of micronuclei observed in the i^{th} animal at dose d_j, $j = 0, 1, 2,\ldots, m$. If Y_{ij} can be assumed to be distributed independently with

$$Y_{ij} \sim \text{Poisson}\,(\mu(x_j)) \tag{1}$$

then, the log-linear Poisson regression can be applied such that

$$\mu(x_j) = \exp\,(\alpha + \beta^T x_j) \tag{2}$$

where α is the log of mean of the reference group, i.e., vehicle control, β^T is a vector of the parameter estimates and x_j is a vector of the covariates. In this case, the covariate is the indicator variable for the dose group.

Accordingly, to observe any significant differences amongst the dosage sets in assessing the genotoxicity effects, the Dunnet's-type procedure was used to determine if any differed from the vehicle control.

The key assumption of this model is that mean and variance are equal, i.e., $V(\mu) = \mu$. This approach, however, does not recognize animals as experimental units and is therefore not recommended since it results in too liberal decisions (HG, 2009).

3.2 Quasi-Poisson model

HG highlight the importance of taking into account the between-animal variability since the individual animal is the experimental unit. Without accounting for extra-variability, the simple Poisson approach becomes more liberal with increasing overdispersion. Its implications include underestimation of standard errors and thus, wrongly inflating the level of significance (Lee et al., 2012). They recommend fitting a quasi-Poisson model for counts. A quasi-Poisson model is a type of generalized linear models where instead of the maximization of the Poisson likelihood, a more relaxed relationship of the mean-to-variance dependency is assumed. Ramon et al. (2002) suggested that, with generalized linear models, overdispersion can be accounted for by fitting the Poisson model but adjust the standard errors and test statistics. This can be done by introducing an additional parameter to the model, representing the deviation from the Poisson variance assumption. To account for extra variation, the quasi-likelihood methods can be fitted under the assumption $V(\mu) = \rho\mu$ where a value of ρ larger than one indicates overdispersion (Hothorn and Gerhard, 2009).

3.3 The trend test

Traditionally, to evaluate the number of micronuclei, it is suggested by Margolin and Risko (1988), cited by Hayashi et al. (1989), to use the Cochran-Armitage (CA) trend test (Cochran, 1954; Armitage, 1955) to verify the dose-response trends of MNPCEs. This tests the null hypothesis of no trend, i.e., the number of MNPCEs is the same for all dose levels versus the alternative that there is a linear trend across increasing levels of dosage. However, the Cochran-Armitage trend test should not be recommended because it is underpowered when the true trend is not linear, i.e., when the dose and response exhibit a convex and concave relationship (Bretz and Hothorn, 2002). Moreover, this should not be recommended because of its ignorance of the between-animal variability (HG, 2009). For this reason HG propose to use a Williams-type procedure for trend test over Cochran-Armitage (CA) trend test. The CA trend test is particularly sensitive for near-linear shape, whereas, the Williams trend test is sensitive to several shapes. This procedure tests the null hypothesis of no difference among the counts against the alternative that the counts are increasing with increasing dosage compared to the control group. To achieve this comparison, higher concentration groups are successively pooled and compared to the control (Herberich and Hothorn, 2012).

3.4 Proof of hazard vs. proof of safety

The purpose of toxicology testing is to assess the safety of a new test substance relative to a control whether it is harmless up to a specified dose, or harmful. Based on this aim, statistical test of the classical null hypothesis of no difference are usually performed. Failing to reject the hypothesis, i.e., either a nonsignificant p-value or when the point-one hypothesized value is greater than

the lower bound confidence limit of relative risk, often leads to the conclusion that the compound has no harmful effect. This is the traditionally used criterion for harmlessness which demonstrates proof of hazard. The major drawback of this approach is the fact that what is controlled by a pre-specified level is the probability of erroneously concluding hazard. In fact the primary control of the false decision rate, i.e., confidence in negative results, should be preferred in toxicology (Herberich and Horthorn, 2012). Summing up, proof of hazard is an indirect approach and often leads to the problem that statistical significance does not necessarily mean toxicological relevance and statistical nonsignificance does not necessarily mean toxicological irrelevance (Hauschke et al., 1999). In short, be confident with negative results (Kirkland, 2000). Thus, the so-called proof of hazard is inappropriate simply because absence of evidence is not evidence of absence (Altman and Bland, 1995). For this reason, HG recommend proof of safety to demonstrate harmlessness where the probability of erroneously concluding safety is directly controlled. The differences to the proof of hazard with the control of the familywise error rate are (1) the estimation of the upper confidence limits instead of the lower limits, and (2) interpreting interval inclusion instead of superiority interpretation by means of point estimator and confidence limit (Hothorn and Gerhard, 2009).

In this study, for proof of hazard, the common decision is to conclude harmlessness if the p-value of the test for any dose versus control is nonsignificant (p-value > 0.05), otherwise harmfulness is concluded. This classical test problem is formulated as follows:

$$Ho: \pi_{dose} / \pi_{control} \leq \text{harmless}$$
$$Ha: \pi_{dose} / \pi_{control} > \text{harmful}$$

For the proof of safety approach, the specified direction of harmlessness is defined, i.e., only an increasing number of micronuclei are considered to be harmful. Thus, the noninferiority test, a one-sided hypotheses test, is used for proof of safety to demonstrate the possible harmfulness of a certain dose assuming a three-fold threshold of tolerability. Harmlessness is declared for at least one dose if the upper limit of the relative risk is below the three-fold threshold.

The following hypotheses are evaluated:

$$Ho: \pi_{dose} / \pi_{control} \geq 3 \text{ harmful}$$
$$Ha: \pi_{dose} / \pi_{control} < 3 \text{ harmless}$$

4. Results and Discussion

Figure 1 illustrates the results of the mice bone marrow erythrocyte micronucleus assay. The assay with T02 posted a positive assay. An increase in the frequency of micronucleated polychromatic erythrocytes (MNPCEs) in

treated mice is an indication of induced chromosome damages (Krishna and Hayashi, 2000).

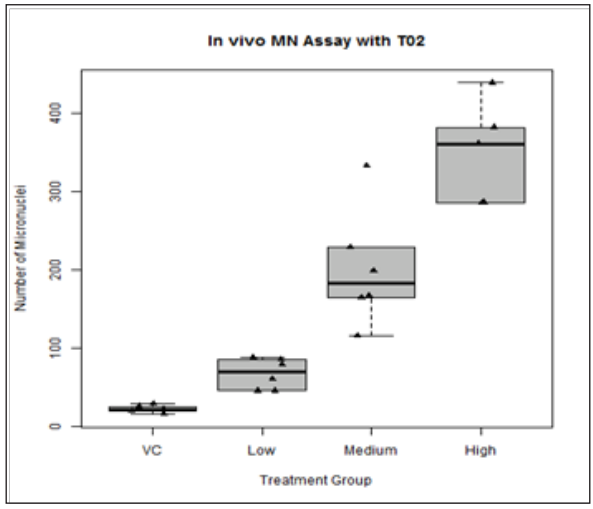

Figure 1. Boxplot for T02

Traditional approach: Poisson model

Table 1 summarizes the results of the in vivo micronucleus test for each dosage group per sacrifice interval. Given in the tables are the relative risk (RR) and the corresponding lower confidence limits estimated by the Poisson model (lower RR).

Genotoxicity activity is indicated by statistically significant dose-related incidence of MNPCEs in the treatment group. A compound is traditionally considered mutagenic in the test system if, at any of the preparation intervals, a statistically significant increase in the number of MNPCEs is found in comparison to the negative control (Shahrim et al., 2006).

Table 1 shows a significant increase in MNPCE observed in all level of doses of T02 over the vehicle control since the point one hypothesized value is less than the estimated lower limit of relative risk. This compound had an indication to induce clastogenic effects and therefore was evaluated as positive at any level of dose administered.

To further explain the genotoxic properties of T02, it is very important to assess the dose-response relation to confirm the toxicity of chemicals (Hayashi et al., 1989). The Cochran-Armitage trend test supports a significant trend (p-value < 0.000) which implies that MN occurrence increases with increasing dose score. The trend test p-values, however, did not provide any information about the biological relevance of the results.

Table 1. Dunnett-type contrasts for relative risks (RR) on MN for T02

		Poisson		Quasi-Poisson	
Comparison	RR	Lower RR	Pvalue	Lower RR	Pvalue
Low-Vehicle	3.13	2.58	0.00*	1.61	0.00*
Med-Vehicle	9.42	7.89	0.00*	5.13	0.00*
High-Vehicle	16.43	13.80	0.00*	9.04	0.00*

** significant at 5% level of significance*

Hothorn-Gerhard recommended approach: Quasi-Poisson model

To properly account for the between-animal variations, a quasi-Poisson model was fitted. The estimated dispersion parameter was 12.0 which indicate the occurrence of extra-variability.

Table 1 provides the estimated relative risk alongside with the one-sided lower confidence limit. Apparently, this model yielded similar conclusions as what was previously discussed from using the Poisson model. There was a significant increase in the micronucleus frequency compared with the vehicle control.

Furthermore, HG propose Williams-type contrasts for the relative risk as an alternative to Cochran-Armitage test since the main objective was to demonstrate a possible dose-related trend. This test is sensitive to several shapes as compared to CA test which is sensitive for near linear shape. The direction of interest, i.e., increasing micronuclei induces chromosomal damage, and only the lower limit is required for a conclusion regarding the trend. Therefore we are only after in the performance of the lower bound confidence limits. The multiple comparisons were done by successively pooling the higher dose groups and compared to the control group. In Table 2 the relative risk estimates and their lower simultaneous confidence limits for Williams-type contrasts for the quasi-Poisson model are given. Since all three comparisons were significantly larger than 1, it was concluded that there is a significant increase in the number of MN with increasing dosage wherein the High dose group led to the most pronounced change in the number of MN compared to the control group.

Table 2. Williams-type contrasts for relative risks (RR) on MN for T02

Comparison	RR	Lower RR
C1: Vehicle vs. High	16.43	9.41
C2: Vehicle vs. Medium and High	12.13	6.99
C3: Vehicle vs. All doses	7.52	4.32

Proof of hazard vs. Proof of safety

Using the traditional approach, the evaluation of the number of micronuclei using proof of hazard revealed that none of the doses of T02 found to be harmless, i.e., signicant p-values was observed in all tests (see Table 1). In this study, noninferiority can be determined by one-sided upper confidence limits because increasing micronuclei is of interest. Assuming a-priori definition of an acceptance threshold θ equals to three for proof of safety, a noninferiority could not be claimed in all dosage groups of T02 (see Table 3). This coincides with the previous findings that T02 is a positive compound.

Table 3. Proof of Safety: One-sided Confidence Limit on Noninferiority (T02)

Comparison	RR	Upper RR
Low-Vehicle	3.13	6.08
Med-Vehicle	9.42	17.29
High-Vehicle	16.43	29.87

5. Conclusions and Recommendations

The ultimate goal of this study is to implement, apply and discuss the recommendations of HG to explain the induced chromosomal damage potential of T02 using *in vivo* micronucleus assay. The two different statistical approaches, i.e., Poisson and quasi-Poisson models lead to the same conclusion – a positive result for T02. However, the crucial part in the analysis is the consequence of failing to account for overdispersion, as in the case of fitting the Poisson model, which may lead to incorrect inferences. Evidently, the comparison of incidences of micronucleated polychromatic erythrocytes (MNPCE) using a quasi-Poisson model is shown to be more appropriate than a similar procedure using a Poisson model. The latter statistical approach becomes more liberal with increasing overdispersion, giving confidence limits for significance that are too narrow.

In terms of claiming a possible genotoxicity of the given compounds, a three-fold threshold of tolerability for the noninferiority test is assumed. However, this chosen threshold indicates the difficulty of claiming harmlessness with sample size of five. HG discuss that for a primary claim of being not genotoxic, more animals than five are needed. In toxicological studies, the sample size is often determined on the basis of regulatory guidelines. Moreover, aside from using non-inferiority test for proof of safety, an alternative approach such as the step-up estimation of the maximal safe dose can also be performed (Hothorn and Gerhard, 2009).

In terms of the interpretation of the results in in vivo, there are several criteria for determining a positive response, one of which is a statistically significant dose-

related increase in the number of micronucleated polychromatic erythrocytes. Another criterion may be based upon detection of a reproducible and statistically significant positive response for at least one of the test substance concentrations. If none of the criteria are satisfied, then the test substance is considered to be nonmutagenic in this system. Sofuni et al. (1990) considered the dose response to be (strong) positive, if it had two significant doses out of three dose groups and decided it to be weakly positive if it had only one significant dose and there was a significant trend. However, both biological and statistical significance should be considered together in an evaluation. Lastly, it is highly recommended to confer with toxicologist for the biological relevance of the findings.

References

ALTMAN, D.G. and BLAND, J.M., 1995, Absence of Evidence is Not Evidence of Absence, *British Medical J*., 311(7003): 485.

ARMITAGE, P., 1955, Test for Linear Trends in Proportions and Frequencies, *Biometrics* 11(3): 375-386.

BRETZ, F. and HOTHORN, L.A., 2002, Detecting dose-response using contrasts: Asymptotic power and sample size determination for Poisson data, *Stat Med*. 21(22): 3325-3335.

COCHRAN, W.G., 1954, Some methods for strengthening the common χ^2 tests. *Biometrics* 10: 417-451.

HAYASHI, M. et. al.,1989, A procedure for data analysis of the rodent micronucleus test involving historical control, *Environmental and Molecular Mutagenesis* 13: 347-356.

HERBERICH, E. and HOTHORN, L., 2012, Statistical evaluation of mortality in long-term carcinogenicity bioassays using a Williams-type procedure, *Regulatory Toxicology and Pharmacology* 64: 26-34.

HAUSCHKE, D. et.al., 1999, Proof of safety in toxicology based on the ratio of two means for normally distributed data, *Biometrical Journal* 41(3): 295-304.

HOTHORN, L.A. and GERHARD, D., 2008, Statistical evaluation of the in vitro micronucleus assay, Reports of the Institute of Biostatistics No. 09/2008.

HOTHORN, L.A. and GERHARD, D., 2009, Statistical evaluation of the in vivo micronucleus assay, *Arch Toxicol* 83: 625-634.

KIM, B.S. et. al., 2000, Statistical analysis of in vivo rodent micronucleus assay, *Mutation Research* 469: 233-241.

KIRKLAND, D.J., 2000, *Statistical Evaluation of Mutagenicity Test Data*, Cambridge University Press.

KRISHNA, G. and HAYASHI, M., 2000, In vivo rodent micronucleus assay: Protocol, conduct and data interpretation, *Mutation Research* 455(1-2): 155-166.

LEE, J. et. al., 2012, Analysis of overdispersed count data: Application to the human Papillomavirus Infection in Men (HIM) study, *Epidemiol Infect* 1087–1094.

MARGOLIN, B.H. and RISKO, R.J., 1988, The statistical analysis of in vivo genotoxicity data: Case studies of the rat hepatocyte UDS and mouse bC (eds): "Evaluation of Short-Term Test for Carcinogens. Report of the International Program on Chemical Safety's Collaborative Study on In Vivo Assays." Cambridge: Cambridge University Press vol. 1 pp. 129-142.

PARODI, S. and BOTTARELLI, E., 2006, Poisson regression model in epidemiology - An introduction, *Medic. Vet.* 25-44.

SHAHRIM, Z. et. al., 2006, The in vivo rodent micronucleus assay of Kacip Fatimah (Labisia pumila) extract, *Tropical Biomedicine* 23(2): 214-219.

SOFUNI, T. et. al., 1990, A comparison of chromosome aberration induction by 25 compounds tested by two Chinese hamster cells (CHO and CHL) systems in culture, *Mutation Research* 241:175-213.