

The assessment and communication of genotoxicity test results: moving beyond binary

Takashi Omori¹, Makoto Hayashi^{2*}

¹Department of Clinical Biostatistics, Graduate School of Medicine, Kyoto University

²Representative, makoto international consulting, 4-23-3-1, Kami-imaizumi, Ebina, 243-0431 Japan

*Corresponding author: Representative, makoto international consulting, 4-23-3-1, Kami-imaizumi, Ebina, 243-0431 Japan, mkoto_hayashi@wine.ocn.ne.jp

Highlights

- Genotoxicity study results have usually been evaluated on a binary (positive or negative) basis
- Statistical/biological evaluation cannot classify all results in binary terms
- Because many evaluations leave room for ambiguity/uncertainty, we believe that it is time to retire the binary paradigm
- Ambiguity/ uncertainty of test results must be communicated clearly

Abstract

Potential genotoxicity is one of the essential considerations in the safety assessment of chemicals to which humans may be exposed. Several endpoints are used to evaluate genotoxicity, but, in each case, a binary assessment (negative/positive) is demanded by regulators. The use of binary assessment has rarely been questioned, although we have pointed out some questions and difficulties with regard to the statistical methods used and the evaluation of biological significance, both of which inform the calls of negative/positive. Here, we discuss these issues further, focusing on ambiguity and uncertainty in the binary paradigm, and we seek a new direction for genotoxicity assessment. To this end, we need to understand, acknowledge, and accept these ambiguities and study-related uncertainties and then to consider new strategies for safety assessment. We also discuss the communication of ambiguity and uncertainty in risk communication.

Keywords:

Human health risk; Study interpretation; Uncertainty; Risk communication

Introduction

To ensure validity, safety assessment tests must be designed and conducted by personnel with appropriate technical and scientific skills. Even when the highest standards are met (e.g., following Test Guideline protocols such as OECD TG-471, 473, and 474 [1-3] and maintaining Good Laboratory Practice (GLP) [4]), the general public may be skeptical about reported risk assessment results. The evaluation and interpretation of test results obtained according to such guidance documents have been the subject of scientific debate [5-7]. Currently, the most common evaluation method is to choose a group as the evaluation unit, to evaluate study results using a statistical approach, and to consider biological validity, before making a final determination. Nonetheless, we must never stop questioning and, where appropriate, revising our testing and evaluation framework. Otherwise, we risk loss of confidence in safety evaluation as a robust science.

Currently, those who design and conduct the tests and produce the measurements are responsible for evaluating and interpreting the results. The test must be understandable and acceptable to all parties involved, including the test operators, regulators, and consumers. The data must be evaluated comprehensively so that sound judgement calls can be made. Knowing clearly the objectives that are to be derived from the data is essential, to ensure that the quality of the target measurements is not an issue. Although data quality will not be discussed in this paper, there are many situations in which it becomes an issue. As mentioned above, TG and GLP play important roles in maintaining data quality. Adherence to them is not, in itself, sufficient, but we should always pay attention to data quality and be open to its improvement [7].

In many cases, uncertainties arise with regard to test results, and those who conduct safety assessments should – and do – strive to communicate them. In contrast, regulatory officials and members of the general public often look for binary outcomes – safe or unsafe. Undoubtedly, in the evaluation of genotoxicity tests, there is a third category: “equivocal” or inconclusive outcomes. However, even in such cases, experts may be pressured to make binary judgements.

In the present paper, we assume that the quality of the measurements is sufficient, and we consider how their evaluation and interpretation can be made more appropriate and more reliable for safety assessment. In our previous papers, we discussed the problems associated with statistical evaluation [6] and the use of biological relevance [7] in the evaluation process. Here, we emphasize ambiguity/uncertainty of genotoxicity test measurements, discuss binary judgments from a critical standpoint, and argue for the need for revised evaluation strategies.

2. General considerations for safety testing: the planning stage

When designing a study, one must consider the purpose of the safety tests. For example, the study design may vary, depending on whether the aim is to identify chemical hazards or to assess human safety risk. Will individual endpoints be considered separately, or will they be placed in the context of other endpoints, such as mutation or chromosomal aberrations? Will individual studies be evaluated or will independent repeat studies (if available) be considered? Certainly, the degree of confidence in a test result will differ, depending on whether it is based on a single study versus two or more consistent studies. The reproducibility of a test is a prerequisite for informed decision-making [8, 9]: those who interpret the results expect them to be

reproducible. Occasionally, a subsequent study is performed, specifically to test the reproducibility of a previous finding. Even if the results are the same, the degree of confidence will vary between the results obtained from a single test and reproducible results from multiple tests.

The selection of a method of evaluation influences interpretation of test results. As long as humans interpret test results, biases associated with personal cognition cannot be avoided; even artificial intelligence systems may replicate human biases. Many reports show that scientific results can be influenced by analysts' opinions or beliefs [10]. To avoid such biases, researchers should declare, in advance, the purpose of the study is, the procedures that will be used to conduct it, and the evaluation methods.

3. Data visualization and summarization

Graphics (histograms, dot charts, etc.) and summary tables are often used to aid understanding of results or to highlight outliers. While useful, aggregation of results should not be allowed to obscure detailed scrutiny of individual tests/ measurements; such scrutiny is important, to ensure accurate assessments.

Data visualization is an important technique and a first step in data analysis [11, 12]. The same data can be graphed to give quite different visual impressions, depending on the choices of axis scales and origins, linear vs logarithmic scales, etc. Computer technology makes it easy to play with these choices. The purpose of visualizing data is not to make it look beautiful, but to express its meaning.

4. Assumptions for statistical data analysis

Statistical methods of data analysis are used routinely to evaluate test results.

When applying statistical models, we must decide, in advance, the unit of evaluation [13-15]. For example, in the Ames test, the *assessment unit* could be either a single agar plate (of the three used per dose), or the group of three plates. In the *in vivo* micronucleus test, using five animals per group, the statistical model depends on whether the assessment unit is the individual animal or the group. Statistical models that analyze variation at the individual unit level (among plates, individual animals, etc.) are preferred and recommended. Simplifications always involve assumptions. The results of statistical analysis depend on the methods and models used. These choices should always be justified; choices should not be subjective or arbitrary.

It can be difficult to judge whether an outlier value obtained in a single test should be attributed to chance or to a technical error. Analysts who use statistical models should understand the models used and the assumptions made. Insight, based on observation and experience, is fundamental to scientific interpretation of results. Reports rely on insights and judgments, which should be described correctly, clearly, and concisely.

5. Dose-response relationship

Establishing a dose-response relationship (a change in response with dose) is an essential component of any safety evaluation. Rigorous approaches are required [7]. Response shapes can differ for each chemical substance tested and the strategy for setting the spacing of the doses should be explained.

Does a statistical model, which merely expresses the mathematical relationship between dose and response, contribute to the evaluation of safety? There have been

many discussions concerning the determination of point-of-departure (POD) or virtually-safe doses (VSD), based on model fitting and extrapolation [17-19]. However, no method has been established for assessing the uncertainties associated with extrapolating dose-response relationships beyond the ranges that were tested. There is no guarantee that such extrapolations accurately evaluate the safety of a chemical substance at human-exposure levels.

6. Historical control data

The variability of measurements obtained from previous tests needs to be considered: do the values (and the variability) of measurements on a concurrent control group differ from those of historical controls [13,14]? Knowledge of previous data is helpful for evaluating test quality. Accumulating and maintaining a data set of (mainly negative) control group results, obtained under standard conditions, can provide the study sponsor with objective information about response fluctuations.

Although guidelines advise the accumulation and analysis of historical control data [15], the methods for interpretation of such data are still debated. To evaluate the test data, we must understand the study: kind of data has been obtained and what values and variations do they show? The accumulation and management of high-quality historical data are essential for reliable safety evaluations [20].

7. The ambiguity of binary decisions about genotoxicity

A binary decision – is a chemical safe in the dose range used – yes or no? – is usually required. However, there are difficulties in making such binary decisions, as has been discussed previously [6,21]. For the genotoxicity test, the regulatory requirement is

a binary negative-or-positive result, with little room for ambiguity or uncertainty. Those who conduct safety assessments need to be forthright concerning uncertainties that may affect the evaluation of the test results. Even in cases where the results are considered to be clear, statistical procedures may have been followed without adequate justification. In clear-cut cases, independent evaluators may quickly draw comparable conclusions. Problems arise when results are ambiguous and it is difficult to make a binary call. In some cases, a statistically significant positive result is obtained, yet the person who conducted the test is uncomfortable and remains unconvinced of an unequivocal positive call. In such cases, additional factors should be considered: the quantity and quality of data; whether the response itself is borderline; whether there is too much variability in the data; whether additional tests were conducted, to confirm reproducibility; whether peripheral information was sought; and so on.

There is a difference between qualitative judgements (positive or negative) and quantitative evaluations (e.g., setting POD). A binary evaluation is required for regulatory applications in genotoxicity and carcinogenicity tests: it is efficient for regulatory purposes. Nevertheless, test guidelines developed (and widely accepted) with the goal of making binary decisions may prove to be obstacles to the reliable evaluation of safety-test outcomes. Judgment is biased toward qualitative evaluation because conservative thinking is preferred, but such decisions can be flawed, with wide-ranging implications for the use of specific chemicals that have potential benefits for humans. The ‘precautionary principle’ is often invoked by decision-makers confronted with ambiguous data: such data is regarded as positive, in an attempt to minimize risk (22). However, experimenters, evaluators, regulators, and consumers should recognize the existence of ambiguity in borderline results and it is the responsibility of scientists to

communicate ambiguity/uncertainties adequately.

8. Acceptance of ambiguity and uncertainties affecting binary decisions

When a test result is clear - positive for anyone to evaluate, or negative without doubt - there is no need to ponder evaluation methods; interpretation is easy. What is difficult is the “in-between part”, where evaluation/interpretation are difficult because of ambiguity, doubt, inability to conclude, etc., and fluctuations due to bias may become large. Regulatory judgments must be made from an overall and broad perspective. Is it possible, necessary, or worthwhile to binarize this gray zone? Factors that may lead to ambiguous results include the quality of the test itself, significant variability in each evaluation unit, considerable variability in lower-level units (e.g., individual units rather than group units), bias already introduced at the time of observation, etc.

In some cases, repeating the test may lead to concurrence and a binary decision can be made with confidence. The problem arises when, despite all efforts, we still find ourselves in a “gray zone”. The late John Ashby once said that “grey is always grey” [personal communication]. Ambiguous results are not compatible with binary decisions. Indeed, it can be difficult to draw a line between the “gray zone” and the clear-cut positive and negative zones. Interpretation of results becomes more complicated and complex because of the gray zone.

9. Safety and security

It is crucial to consider safety and security in the relationship between chemicals and people's daily lives. Here, we define *security* (“anshin”, in Japanese) as a situation in which one feels less anxious - separated from danger. Safety is an issue in the natural sciences, while security is a matter for the humanities, including psychology and sociology. Can we draw a line between safety and non-safety based on knowledge of

natural science, or is safety simply a binary value: safe or unsafe? Perceptions of safety depend on individual attitudes and are not always fixed or definitive. Individual sensitivity to the chemicals in question needs to be considered. With our growing understanding of the immense individual genetic variations in the population, it becomes clear that individual responses to chemicals, including potential genetic toxins, are also extremely variable, especially in relation to ADME factors. Some argue that tailor-made “personalized” toxicology should be developed, and the safety of each individual should be evaluated. If this were to happen, chemical substance regulation targeting a population would not be possible, with current technology and human resources. How representative for each individual is the evaluation of the safety of the population ? While security is a more complicated issue and is not the target of this article, the topic of “safety and security” needs to be addressed for human safety risk assessment.

10. Beyond the ambiguity issue

The safety evaluation, interpretation, and judgment regarding a chemical substance is not based on the results of a single test. Several tests with different endpoints are conducted, to provide a comprehensive view of safety, considering also the utility, benefits, and daily exposures to chemicals. Even if the result of one safety test is in the “gray zone”, results of other tests may lead to a correct decision. However, if the majority of the results of various tests are in the gray zone, we should express conclusions based on the factors considered and how the final judgment is drawn, with full transparency. As mentioned above, the utility/benefits and amount of daily exposure are important factors. In addition, the marketing authorization of a product using a

particular chemical substance requires consideration from many perspectives, such as the availability of alternatives and their safety (risk trade-off) [23]; usage; range of uses; past experience of use; and any track record of use overseas. The same judgment can be made if the safety evaluation is based on conventional binary judgment. However, by thoroughly examining the situation from various perspectives and clarifying ambiguous facts in the gray zone, we can increase the potential for using beneficial products more wisely, thereby enriching our lives while, at the same time, reducing potential harm caused by chemical substances.

11. Risk communication

“Risk communication” is a relatively new term, defined as “the mutual exchange of information and opinions among risk assessors, risk managers, consumers, businesses, researchers, and other interested parties during the entire risk analysis process.” Are communications from experts (here, risk assessors and researchers), which establish the “mutual exchange” noted in this definition, successful in communicating uncertainty? Even if uncertainty is expressed, this should not be a statement regarding the avoidance of responsibility. Risk assessment must be accurate for proper risk management. Experts should not simply give a binary judgment but should explain their viewpoints and how their conclusions were reached. If ambiguity exists, the expert is responsible for informing about that ambiguity. The recipient of the risk analysis should be sufficiently sophisticated to accept ambiguity. The development and dissemination of risk communication is essential. However, it must be avoided that the discussion disregards science, as indicated in “Thinking on the Safe Side,” to avoid ambiguity [7]. We should know that there are ambiguities, uncertainties, and

fluctuations in science, especially in biology, and we should consider evaluating/interpreting safety test results based on such ambiguities, accurately communicating these facts to third parties.

12. Proposal and further consideration

Accurate, fair, and transparent human safety assessment is critical for risk assessment, management, and communication. Currently, emphasis is placed on safety testing using animals; however, other evaluation methods should be considered. From the viewpoint of animal welfare, it is necessary to respect the “3R's” philosophy and to reconsider whether animal testing is essential.

One strategy is to separate the group in charge of the test that produces the data from the department in charge of objectively evaluating the results, a “third-party” standpoint. Each person in charge should be a professional in his/her field and each department should make its best efforts. At the time of reporting, it would be appropriate to interpret the evaluation results based on the overall conclusions of both parties, and all other parties capable of making appropriate evaluations obtained up to that point in time.

Science must always be open to new knowledge. The humble acceptance of uncertainties in the data quality at the time of the evaluation and the assumptions made for the statistical methodology, etc., should be respected. We are awaiting the construction of a system that builds expertise in each area, discloses information accurately, fairly, and transparently, and uses it wisely.

Acknowledgments

The authors thank Dr. David Tweats, Stephen Dertinger and Dr. Takeshi Morita for their precise manuscript review and valuable suggestions. We also thank the reviewers and especially the Editor for their constructive and invaluable comments and suggestions..

Funding

This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no conflict of interest.

~~Declaration of generative AI and AI-assisted technologies in the writing process~~

~~During the preparation of this work the authors used Paperpal in order to improve consistency of grammar, vocabulary, and expression in English writing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.~~

References

- [1] OECD, Test No. 471 Bacterial Reverse Mutation Test, OECD Publishing, Paris, 2020. <https://www.oecd.org/chemicalsafety/risk-assessment/1948418.pdf>
- [2] OECD, Test No. 473: *In vitro* Mammalian Chromosomal Aberration Test, , OECD Publishing, Paris, 2016. <https://doi.org/10.1787/9789264264649-en>.
- [3] OECD, Test No. 474 Mammalian Erythrocyte Micronucleus Test, OECD Publishing, Paris, 2016. https://read.oecd-ilibrary.org/environment/test-no-474-mammalian-erythrocyte-micronucleus-test_9789264264762-en#page1
- [4] OECD, OECD series on principles of good laboratory practice and compliance monitoring. 1. OECD Principles on Good Laboratory Practice, Paris, 1998. [https://one.oecd.org/document/ENV/MC/CHEM\(98\)17/en/pdf](https://one.oecd.org/document/ENV/MC/CHEM(98)17/en/pdf)
- [5] M. Hayashi, Opinion: regulatory genotoxicity: past, present and future, Genes Environ. 44 (2022) 13, <https://doi.org/10.1186/s41021-022-00242-5>.
- [6] M. Hayashi, “Statistical significance” and other important considerations in genotoxicity safety testing, Mutat. Res. 888 (2023a), 503627. <https://doi.org/10.1016/j.mrgentox.2023.503627>
- [7] M. Hayashi, Evaluation and interpretation of cytogenetic test results based on biological relevance, Mutat. Res. 891 (2023b) 503639. <https://doi.org/10.1016/j.mrgentox.2023.503693>
- [8] B. Pulverer, Reproducibility blues, EMBO J. 34 (2015) 2721-2724. <https://doi.10.15252/emboj.201570090>
- [9] M.D. Twa. Scientific integrity and the reproducibility crisis. Optom Vis Sci. 96 (2019) 1-2. <https://doi.10.1097/OPX.0000000000001339>.

- [10] J.P. Simmons, L.D. Nelson, U. Simonsohn. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci.* 22 (2011) 1359-1366. [https://doi: 10.1177/0956797611417632](https://doi.org/10.1177/0956797611417632).
- [11] E.R. Tufte. *The Visual Display of Quantitative Information*, 2nd ed., Graphics Press, 2001.
- [12] W.S. Cleveland. *Visualizing Data*. Hobart Press, 1993.
- [13] M. Hayashi, I. Yoshimura, T. Sofuni and M. Ishidate, Jr., A procedure for data analysis of the rodent micronucleus test involving a historical control, *Environ. Mol. Mutagen.* 13 (1989) 347-356. <https://doi.org/10.1002/em.2850130412>
- [14] I.D. Adler, J. Bootman, J. Favor, G. Hook, G. Schriever-Schwemmer, G. Welzl, E. Whorton, I. Yoshimura, M. Hayashi. Recommendations for statistical designs of in vivo mutagenicity test with regard to subsequent statistical analysis, *Mutat. Res.* 417 (1998) 19-30. [https://doi.org/10.1016/S1383-5718\(98\)00091-6](https://doi.org/10.1016/S1383-5718(98)00091-6).
- [15] OECD, Overview of the set of OECD Genetic Toxicology Test Guidelines and updates performed in 2014-2015. ENV Publications. Series on Testing and Assessment, No. 238; 2nd ed., OECD, Paris. 2017. [https://one.oecd.org/document/ENV/JM/MONO\(2016\)33/REV1/en/pdf](https://one.oecd.org/document/ENV/JM/MONO(2016)33/REV1/en/pdf)
- [16] B.B. Gollapudi, G.E. Johnson, L.G. Hernandez, L.H. Pottenger, K.L. Dearfield, A.M. Jeffrey, E. Julien, J.H. Kim, D.P. Lovell, J.T. Macgregor, M.M. Moore, J. van Benthem, P.A. White, E. Zeiger, V. Thybaud. Quantitative approaches for assessing dose–response relationships in genetic toxicology studies. *Environ Mol Mutagen* 54 (2013) 8–18. <https://doi.org/10.1002/em.21727>

- [17] G.E. Johnson, L.G. Soeteman-Hernandez, B.B. Gollapudi, O.G. Bodger, K.L. Dearfield, R.H. Heflich, J.G. Hixon, D.P. Lovell, J.T. MacGregor, L.H. Pottenger, C.M. Thompson, L. Abraham, V. Thybaud, J.Y. Tanir, E. Zeiger, J. van Benthem, and P.A. White. Derivation of Point of Departure (PoD) Estimates in genetic toxicology studies and their potential applications in risk assessment, *Environ. Mol. Mutagen.*, 55 (2014) 609-623. <https://doi.org/10.1002/em.21870>
- [18] J.T. MacGregor, R. Frötschl, P. A. White, K. S. Crump, D. A. Eastmond, S. Fukushima, M. Guérard, M. Hayashi, L.G. Soeteman-Hernández, T. Kasamatsu, D.D. Levy, T. Morita, L. Müller, R. Schoeny, M. J. Schuler, V. Thybaud, G.E. Johnson, IWGT report on quantitative approaches to genotoxicity risk assessment I. Methods and metrics for defining exposure–response relationships and points of departure (PoDs), *Mutat. Res.* 783 (2015), 55-65.
<http://dx.doi.org/10.1016/j.mrgentox.2014.09.011>
- [19] J.T. MacGregor, T. James, R. Frötschl, P. A. White, K. S. Crump, D.A. Eastmond, S. Fukushima, M. Guérard, M. Hayashi, L. G. Soeteman-Hernández, G.E. Johnson, T. Kasamatsu, D.D. Levy, T. Morita, L. Müller, R. Schoeny, M.J. Schuler, V. Thybaud, IWGT report on quantitative approaches to genotoxicity risk assessment II. Use of point-of-departure (PoD) metrics in defining acceptable exposure limits and assessing human risk, *Mutat. Res.* 783 (2015), 66-78.
<http://dx.doi.org/10.1016/j.mrgentox.2014.10.008>
- [20] S.D. Dertinger, D. Li, C. Beevers, G.R. Douglas, R.H. Heflich, D.P. Lovell, D.J. Roberts, R. Smith, Y. Uno, A. Williams, K.L. Witt, A. Zeller, C.-h. Zhou, Assessing the quality and making appropriate use of historical negative control data: A report of the International Workshop on Genotoxicity Testing (IWGT), *Environ. Mol.*

Mutagen. (2023).

<https://doi.org/10.1002/em.22541>

- [21] V. Amrhein, S. Greenland, B. McShane, et al., Retire statistical significance, Nature, 567 (2019) 305-307. <https://www.nature.com/articles/d41586-019-00857-9>.
- [22] Guidance on genotoxicity testing and data interpretation for pharmaceuticals intended for human use. In: ICH Guideline S2(R1), 2011.
<https://database.ich.org/sites/default/files/S2%28R1%29%20Guideline.pdf>
- [23] J.D. Graham and J.B. Wiener. Risk vs. Risk: Tradeoffs in Protecting Health and the Environment, Harvard University Press, 1995.