

Appendix D – Approach for assessing genotoxicity studies

1. Evaluation of reliability and relevance of results of genotoxicity studies – general considerations

Evaluation of data quality for hazard/risk assessment includes evaluation of reliability and relevance (Klimisch et al., 1997; OECD, 2005; ECHA, 2011).

The relevance of study results was categorized into high, limited or low relevance. The relevance of a study result was based on its reliability and on the relevance of the test system and study design. The cross-cutting WG Genotoxicity developed a scoring system for reliability based on the scoring system of Klimisch et al. (1997) as recommended in the EFSA Scientific Committee Guidance on genotoxicity testing strategies (EFSA Scientific Committee, 2011b). The reliability scores were:

1. reliable without restriction
2. reliable with restrictions
3. insufficient reliability
4. reliability cannot be evaluated
5. reliability not evaluated since the study is not relevant and/or not required for the risk assessment (in case the study is reported for reasons of transparency only)

These reliability scores were defined as follows:

1. Reliable without Restriction

“This includes studies or data from the literature or reports which were carried out or generated according to generally valid and/or internationally accepted testing guidelines (preferably performed according to GLP) or in which the test parameters documented are based on a specific (national) testing guideline (preferably performed according to GLP) or in which all parameters described are closely related/comparable to a guideline method.”

2. Reliable with Restrictions

“This includes studies or data from the literature, reports (mostly not performed according to GLP), in which the test parameters documented do not totally comply with the specific testing guideline, but are sufficient to accept the data or in which investigations are described which cannot be subsumed under a testing guideline, but which are nevertheless well documented and scientifically acceptable.”

*3. Insufficient Reliability*¹ “This includes studies or data from the literature/reports in which there are interferences between the measuring system and the test substance or in which organisms/test systems were used which are not relevant in relation to the exposure (...) or which were carried out or generated according to a method which is not acceptable, the documentation of which is not sufficient for an assessment and which is not convincing for an expert judgment.”

*4. Reliability cannot be evaluated*² “This includes studies or data from the literature, which do not give sufficient experimental details and which are only listed in short abstracts or secondary literature (books, reviews, etc.).”

In case a study is reported for reasons of transparency only, a further score (5) may be required:

5. Reliability not evaluated

The study is not relevant and/or not useful for the risk assessment.

¹ Klimisch et al. (1997) used the term “Not reliable” however, “Insufficient reliability” was considered more appropriate.

² Klimisch et al. (1997) used the term “Not assignable” however, “Reliability cannot be evaluated” was considered more appropriate.

The above mentioned references (Klimisch et al., 1997; OECD, 2005; ECHA, 2011; EFSA Scientific Committee, 2011b) may be consulted for more details and examples. Generally, the assignment of a reliability score is expert judgement based on defined criteria.

Each reliability box in the summary tables (see Appendices from J to P) started with the reliability score, followed by comments justifying the score. This is equally applicable for *in vitro* and *in vivo* studies.

The relevance of the test results is mainly, but not exclusively, based on:

- Genetic endpoint (high relevance for gene mutations, structural and numerical chromosomal alterations as well as results obtained in an *in vivo* comet assay (which belongs to the assays recommended by the EFSA Scientific Committee (2011b) for the follow-up of a positive *in vitro* result); lower relevance for other genotoxic effects). Other test systems although potentially considered of limited or low relevance may provide useful supporting information.
- Cell lines (e.g. mammalian vs non-mammalian) in case of *in vitro* studies.
- Animal species (e.g. mammalian vs non-mammalian, rodents vs non-rodents) in case of *in vivo* studies.
- Route of administration (e.g. oral vs intravenous, intraperitoneal injection, inhalation exposure and intratracheal instillation) in case of *in vivo* studies.
- Status of validation (e.g. for which an OECD Test Guideline (TG) exists or is in the course of development, internationally recommended protocol, validation at national level only, no validation).

Tables (reported in Appendices from J to P) were used in order to structure the outcome of the evaluations in a transparent way and to provide a possibility to consider the relevance of study results in a weight-of-evidence approach. Remarks were inserted in the columns "Reliability" and assigned relevance to the test results in order to justify the judgments. Minor and/or major deviations from OECD TGs were reported in column "Reliability" (e.g. lack of positive control, inappropriate exposure conditions, limited reporting etc.).

The studies were grouped in these tables based on genetic endpoints or test systems and chronologically within these groups. The results were evaluated by the cross-cutting WG Genotoxicity and presented as positive, negative, equivocal or inconclusive. Non-genotoxicity endpoints (e.g. ROS production) were reported, if considered relevant for the interpretation of the genotoxicity endpoints, but the results were not classified as "positive" or "negative".

Evaluation of reliability and relevance of the test system/test design was always performed irrespectively whether a study has been conducted in compliance with Good Laboratory Practice (GLP) or not. The type of a document (i.e., publication or unpublished study report) and the question if the study has been performed according to GLP or not do not necessarily have an impact on the reliability score. The details reported are key for judgment of the reliability and relevance of the information irrespectively of whether or not published in a peer-reviewed journal.

2. Aspects specific for genotoxicity evaluation of nanoparticles to be considered as part of the evaluation

OECD TGs generally do not address nano specific aspects such as specific requirements for exposure time, treatment conditions, concentration range etc. Therefore, deviations from the standard protocols are sometimes needed and considered acceptable if justified. Detailed advice on the validity of genotoxicity tests for materials containing small particles, which is relevant for the E171 assessment, is given in the relevant sections of the EFSA Scientific Committee Guidance documents (2018a, 2020). The main considerations from the EFSA Scientific Committee Guidance on risk assessment of the application of nanoscience and nanotechnologies in the food and feed chain (2018a) are summarised below. Another aspect is the level of dispersion of the nanoparticles. Some additional considerations based on recent reviews related with the genotoxicity assessment of nanoparticles (Elespuru et al., 2018) and nano-TiO₂ in particular (Charles et al., 2018) are also included.

2.1 *In vitro* studies

More weight was given to study designs including observations confirming that cells were exposed to the nanoparticles. Negative results from studies where the cell uptake was not demonstrated were considered as inconclusive (to which only low relevance was assigned). In the case of positive results, a demonstration of exposure

at the cellular level was not required provided the study did not have major shortcomings and there was no indication for an artefact or a result due to an effect not related to the test substance.

A low weight was given to studies performed using only excessively high concentrations i.e. higher than 100 µg/ml (because of aggregation/agglomeration and precipitation of the tested nanoparticles at high concentration).

A low weight was given to studies in which no parallel toxicity evaluation was performed or an inappropriate toxicity test had been used.

The use of metabolic activation system (S9) was evaluated on a case by case basis because it may interfere with the assay reducing the nanoparticles bioavailability, especially in the case of poorly water soluble nanoparticles, such as TiO₂.

2.1.1 *In vitro* gene mutation

Bacterial reverse mutation (Ames) assay is not considered suitable for investigation of gene mutations (due to limitations in the penetration of particles through the bacterial cell wall and the lack of internalisation in bacteria), and therefore assigned low relevance. Hence a higher weight was given to mammalian cell models (OECD TG 476 and OECD TG 490).

2.1.2 *In vitro* micronucleus assay

In the evaluation of *in vitro* micronucleus assays special consideration was given to duration of the treatment, the cytochalasin B dosing regime, possible interference of the nanoparticles with the dyes applied (if flow cytometry is used) and the suitability of the cell lines.

Higher weight was given to studies with an extended treatment, covering at least one cell cycle.

A low weight was given to studies in which cytochalasin B and nanoparticles were simultaneously added (cytochalasin B needs to be added after the nanoparticles, since cytochalasin B might inhibit the cellular uptake of nanoparticles).

If flow cytometry was used, a higher weight was given to studies where possible interference of the tested material with the dye applied was investigated.

A higher weight was given to studies in which the uptake capability of the selected cell lines was demonstrated. A low weight was given to studies based on cell lines with high background micronuclei frequency (higher than 2%) in line with OECD (2014).

2.1.3 *In vitro* Comet assay

In vitro Comet assays are widely used in genotoxicity testing of nanoparticles. However, the lack of an OECD TG leads to a high variability in the test protocols used. Evaluation of the relevance of the test design included identification of possible interferences (e.g. interaction of nanoparticles with dye and lysis condition) within the comet assay at the applied test conditions.

Use of hydrogen peroxide as a positive control is recommended for detection of DNA strand breaks, but was not considered appropriate for the Comet +Fpg assay because it is not specific for detection of oxidised bases (Møller et al., 2018).

Studies in which only Olive tail moment was reported were considered to be of limited reliability.

2.2 *In vivo* studies

In the evaluation of the relevance of the tested system the respective OECD TGs should be considered.

Because TiO₂ needs to be assessed as a food additive, administration by non oral routes of exposure was considered of limited or low relevance, depending on the reliability of the study and other aspects such as information on the level of dispersion.

Regarding the *in vivo* Comet assay, a negative outcome observed in a study with 24h sampling time (without shorter sampling times) might still be considered acceptable even though not fully in line with the OECD TG and needs to be evaluated on a case by case basis.

2.3 Relevance of the results of studies with nanoparticles

The relevance of the study result (high, limited or low) is based on its reliability and on the relevance of the test system/test design (as described above) and also on the characterisation and dispersion of the test item. The relevance of the result may be different for addressing the conventional material and the fraction of nanoparticles. For the fraction of nanoparticles the relevance may vary depending on the concentration/dose (including within the same study) due to agglomeration.

The score for nanoscale considerations (NSC) related to specific aspects in the study design (dispersion and/or confirmation of internal exposure) for all studies has been performed according to Appendix E and considered in the evaluation of the relevance of the results.

Studies for which the relevance of the result was judged to be low were not considered further. The evaluation was focused on studies of high and limited relevance in a weight-of-evidence approach (EFSA Scientific Committee, 2017).