*Mutagenesis*, 2022, **37**, 76–88
https://doi.org/10.1093/mutage/geab026
Advance access publication 3 June 2021
**Original Manuscript**

OXFORD

Editor's choice

Original Manuscript

# Validation of the hen's egg test for micronucleus induction (HET-MN): detailed protocol including scoring atlas, historical control data and statistical analysis

**Katrin Maul[1,*], Dagmar Fieblinger[1], Andreas Heppenheimer[2], Juergen Kreutz[3], Manfred Liebsch[1], Andreas Luch[1], Ralph Pirow[1], Albrecht Poth[2], Pamela Strauch[2], Eva Dony[2], Markus Schulz[2], Thorsten Wolf[4] and Kerstin Reisinger[3]**

[1]Department of Chemical and Product Safety, German Federal Institute for Risk Assessment (BfR), Berlin, Germany, [2]ICCR-Roßdorf GmbH (former: Harlan Cytotest Cell Research, Envigo CRS), Roßdorf, Germany, [3]Henkel AG & Co KGaA, Duesseldorf, Germany and [4]University of Osnabrueck, Osnabrueck, Germany

*To whom correspondence should be addressed. Email: katrin.maul@bfr.bund.de

Received 1 February 2021; Editorial decision 9 July 2021; Accepted 3 August 2021.

## Abstract

A validation exercise of the hen's egg test for micronucleus induction was finalised with a very good predictivity based on the analysis of micronuclei in peripheral erythrocytes of fertilised chicken eggs (Reisinger *et al.* The hen's egg test for micronucleus-induction (HET-MN): validation data set. *Mutagenesis*, this issue). For transparency reasons this complementary publication provides further details on the assay especially as it was the first validation study in the field of genotoxicity testing involving the use of chicken eggs. Thus, the experimental protocol is described in detail and is complemented by a scoring atlas for microscopic analysis in blood cells. In addition, general characteristics of the test system, which is able to mirror the systemic availability of test compounds, are delineated: the test compound passes the egg membrane and is taken up by the blood vessels of the underlying chorioallantoic membrane. Subsequently, it is distributed by the circulating blood, metabolised by the developing liver and the yolk sac membrane and finally excreted into the allantois, a bladder equivalent. In specific, the suitability of the test system for genotoxicity testing is shown by, *inter alia*, a low background DNA damage in a comprehensive historical control database. In addition, the state-of-the-art statistical method used to evaluate obtained data is delineated. It combines laboratory-specific effect threshold with the Umbrella–Williams test, a statistical model also of interest for other genotoxicity test methods.

## Introduction

Three-dimensional test systems gain increasing importance as part of *in vitro* methods for toxicological safety assessment, e.g. refs (1,2). These approaches follow recommendations of international expert groups and regulatory agencies, e.g. refs (3–5), to address downsides observed with *in vitro* methods based on the fact that '2D cultures have less than 1% of both cell density per volume and cell-to-cell contacts when compared to native tissues' (6). In consequence, cell function and physiology of 2D cell cultures as well as their response to external stimuli are limited to reflect the *in vivo* situation. This general description can be exemplified with the reduced metabolic capacity of 2D cell cultures, which is compensated in *in vitro* genotoxicity assays by adding an external metabolic activation system (usually rat liver S9 mix) to mimic liver metabolism. However, even the revised version of *in vitro* Organization for Economic Co-operation and Development (OECD) test guideline (TG) considers S9 mix as not appropriately mirroring the *in vivo* situation as it concedes several downsides (7–9). Preparation

of S9 mix implies a destruction of cell structures, resulting in reduced metabolic capacities with an imbalance towards the toxifying CYP450 phase-I system. Furthermore, S9 mix tends to cause strong cytotoxic effects, thus preventing longer incubation times, which may be required to identify certain genotoxicants. In addition, test systems for *in vitro* methods are often based on tumour cell lines, which lack normal cell cycle control.

In order to compensate for these disadvantages, single-organ 3D test systems have been developed, e.g. ref. (1), either as horizontally oriented systems to represent epithelial tissues or as spheroids to mirror solid organs. Whereas 3D tissues of the skin or the cornea have already gained regulatory acceptance (10–12), test methods based on spheroids are less advanced. For example, respective approaches face the problem of tissue heterogeneity as cells on the surface differ from those inside the spheroids. This is supported by the fact that the cells are nourished by diffusion and the supply of those in the centre could be diffusion-limited (13). Even if several cell types are used to generate spheroids, the functional unit of solid organs such as the liver lobule is not represented (14–16).

Following these considerations, fertilised chicken eggs have been introduced into genotoxicity testing and were combined with a classical read-out parameter, i.e. the analysis of the micronucleus (MN) rate in peripheral nucleated erythrocytes, to develop the hen's egg test for micronucleus induction, the HET-MN (17). As a major advantage the test mirrors the systemic availability of test compounds reflecting certain steps of Absorption, Distribution, Metabolism, Excretion (ADME): after a test compound is applied through a small hole in the eggshell, it is taken up by the inner shell membrane and the underlying highly vascularised chorioallantoic membrane (CAM). The test compound is *d*istributed via the vessel system and *m*etabolised by the developing liver and the yolk sac membrane. Finally, the parent compound and/or its metabolites are actively *e*xcreted into the allantois, a bladder equivalent, which is accessible for sampling.

The developing egg has a pronounced intrinsic metabolic capacity, which may be considered as an adaptation to the autonomy of the developing egg, which lacks a protecting maternal organism (18). Several phase-I and phase-II biotransformation characteristics were reported for the developing egg, including various cytochrome P450 enzymes and reactions such as glucuronidation, sulphation, acetylation and methylation, e.g. refs (19–27). In line with these findings, metabolic profiles of e.g. ethyl 4-hydroxybenzoate in the developing chicken eggs correlated to those in humans (26). In consequence, no S9 mix needs to be added to successfully characterise effects of promutagens with the HET-MN (17,28–31).

Further studies are underway to investigate the metabolic capacity specifically between Days 8 and 11 of egg development, the developmental window, which is utilised for the HET-MN (K. Reisinger, personal communication). First results of those studies showed that the developing liver and the yolk sac membrane exhibit a high metabolic capacity. The latter also represents the major site of erythropoiesis at that early state of egg development (32–34). Thus, test compounds that enter the test system via fenestrated blood vessels of the CAM are metabolised in close vicinity to the repository of cells, which are used to analyse the chemical's genotoxic potential. This means that a pre-systemic metabolic elimination of a toxic test compound, which is described for some orally administered drugs by the intestine/liver first-pass effect, is not expected. The sensitivity of the method is further enhanced due to the fact that both the blood volume and the erythrocyte number per blood volume increase exponentially during this developmental window (35,36), while the spleen is not functional to be able to remove damaged erythrocytes (35,37).

For HET-MN studies, standardised specific-pathogen-free (SPF) chicken eggs are utilised, which are also used for influenza vaccine production. Suppliers enable a global availability (see Section Chicken eggs: supply and incubation) of the eggs and guarantee their stable genetic identity. The eggs that can easily be bred in respective incubators are used for experimental purposes exclusively in the first half of egg development, during which the embryonic nervous system is not completely differentiated. In consequence, cerebral activities are only demonstrated in the second half (38–40). This premature physiological status is also acknowledged by legislations around the globe as they do not consider assays such as the HET-MN as animal experiment. This applies e.g. to the EU, UK, Switzerland, the USA and New Zealand (41–45).

Based on the intrinsic characteristics of SPF chicken eggs, the HET-MN underwent a thorough development phase in one laboratory (17,28–30) before the method was transferred to and further optimised together with another laboratory. Taken together, 21 chemicals, comprising different genotoxic modes of action and different chemicals classes, were tested and all correctly predicted in the two laboratories (46,31). Subsequently, the performance of the assay was further investigated in a validation exercise with 34 chemicals being tested double-blinded in three laboratories, i.e. according to OECD recommendations (47). The results on predictivity and reproducibility are reported in Reisinger *et al.* (46).

Here we provide information on the HET-MN to further support the growing confidence in the assay. A detailed description of the protocol used in the validation exercise is given—together with recommendations for regulatory testing—including a detailed image atlas for cellular and nuclear damage analysis in the chicken blood cells. Further, the general suitability of the test system for genotoxicity testing is substantiated by an extensive data set of historical controls, from which lab-specific acceptance criteria and thresholds for a positive call were derived. Finally, the validation data set was used for an evaluation of the two prediction models concluding in a recommendation of a final model to be used for future HET-MN studies.

With providing this essential information, we intend to support the transfer of the assay to interested laboratories and to finally support the regulatory acceptance of the assay.

## Materials and methods

The following sections reflect the detailed protocol used during the validation exercise; a short version of the protocol was recently published (48). Based on the thorough method optimisation and transferability phase no changes, but some amendments needed to be introduced after the study. For regulatory testing certain steps can be omitted, as indicated below.

### Chemicals

Chemicals were obtained from local suppliers: ethanol (>99.5%; e.g. Sigma-Aldrich), dimethyl sulphoxide (DMSO, >99.9%; e.g. Sigma-Aldrich), May-Gruenwald solution (eosine/methylene blue solution; Merck, Germany), Giemsa solution (azur eosine/methylene blue solution; Merck, Germany), citric acid monohydrate (e.g. Sigma-Aldrich), xylol (e.g. Sigma-Aldrich) and mounting medium (e.g. DePex, Serva) while cyclophosphamide monohydrate (99%; molecular weight: 279.1 g/mol; Sigma-Aldrich) and isopropyl myristate (IPM, >98%; Sigma-Aldrich) were distributed among laboratories during the validation process to maximise harmonisation.

## Chicken eggs: supply and incubation

Fertilised White Leghorn chicken eggs of SPF quality were used for the HET-MN validation. SPF eggs can be obtained e.g. from VALO Biomedia (www.valobiomedia.com) or Charles River (www.criver.com/products-services/avian-vaccine-services/spf-eggs), which together ensure a global availability of these standardised eggs (except for Africa). In the present validation project mainly eggs from VALO Biomedia (Osterholz-Scharmbeck, Germany) were used. In the rare cases when eggs from Charles River were used, no difference in the experimental outcome was observed (data not shown). For standardisation, only eggs within the weight range of 65 ± 4 g on Day 8 of development were used. More eggs than necessary for the experiments were ordered to account for a possible 20% loss due to damage, non-fertilised eggs or those with weights outside the acceptance range (for calculation of egg demand per experiment see also Section Main experiments with dosing regime).

The eggs were delivered within 1 day after egg deposition. Upon delivery they were stored at 4–8°C for a maximum of 4 days. Prior to an incubation at 37.5 ± 0.5°C and a humidity of ~70% (40–80%), eggs were kept at room temperature for a minimum of 2 h before being inspected for integrity. In the incubator, they were positioned horizontally on trays and automatically rotated (3 ± 1 h intervals) to simulate natural incubation conditions (Figure 1; incubator e.g. Ehret breeder). After 8 days of incubation in the breeder, i.e. on Day 8 of egg development (1 day was consistent with an incubation period of 24 h), fertilisation and viability were checked by candling of eggs (Figure 1). Only eggs with weights within the acceptance range and with a normal vascularisation were randomised and allocated to control and dose groups.

## Study design

The investigation of the genotoxic potential with the HET-MN follows the design of standard *in vitro* genotoxicity assays comprising (i) a solubility study, a recommended (ii) pre-test, (iii) a dose range-finding experiment and for validation purposes (iv) at least two valid main experiments. For regulatory testing, laboratories may finalise testing after a valid and positive first experiment.

### Solubility study

First priority, deionised water (aqua DI; standard 300 µl/egg, maximum 1500 µl) and IPM (50 µl/egg) were used to solubilise chemicals as they proved best compatibility with the test system and did not interfere with respective test compounds. Ethanol (10%, v/v; 100 µl/egg) and DMSO (1% or 10%, v/v; 300 µl/egg or 100 µl/egg) were of second priority.

In case additional solvents are intended to be used, a historical control data set should be established to demonstrate the same low MN rate as that obtained with the recommended ones.

A maximum of 100 mg/egg (weight: 65 ± 4 g) of test compound is applied for HET-MN studies, which corresponds to the top dose in the mammalian *in vivo* MN test, i.e. 2000 mg/kg body weight (7). The maximum dose for less soluble test compounds was determined by supplying additional small volumes of the solvent. Solubility was inspected visually and could be supported by warming or ultrasonification of the test compound. The application of the test compound in the solubilised form was the first choice. In rare cases of hardly soluble test compounds, homogeneous suspensions or emulsions can be applied (see ref. (46)), which could result in the precipitation of the test compound on the inner shell membrane. So far, the authors observed that precipitations did not interfere with the integrity of the test system and had therefore no negative impact on the experimental outcome (please note that for a valid experiment bioavailability of the test compound in the egg needs to be proven). In addition, it was demonstrated that chicken eggs tolerate an exposure to chemical solutions with pH values within a broad range between pH 2 and 12 (49).

### Pre-test

This short-time test was used to define the dose range for the subsequent dose range-finding experiment, especially for well soluble test compounds. For this purpose, a limited number of eggs, e.g. two per dose group, was exposed to a limited number of doses, e.g. the highest soluble dose and several dilutions, for 0.5 h up to 48 h (starting on Day 8; controls are not applied). It should be noted that residues of the applied solution may still be visible on the egg membrane as the penetration rate of aqua DI is ~0.5 h for 300 µl and ~1 h for 1 ml, whereas ~6 h were observed for 50 µl IPM.

### Dose range-finding experiment

The dose range-finding experiment was designed to define the maximum dose for the main experiments, which could be limited by (i) the solubility, if it is <100 mg/egg or (ii) the chemical's general toxicity (see Section Validity of experiments). For well soluble test compounds logarithmic dose spacing is recommended to cover a wide dose range, whereas for less soluble ones a closer spacing might be more appropriate. Eggs were exposed on Day 8 of egg development and viability was determined on Day 11 at the end of the experiment. The laboratories also prepared slides to investigate the MN rate in the dose range-finding experiment. In case it meets all validity criteria (see Section Validity of experiments), it can be accepted as main experiment.

### Main experiments with dosing regime

Main experiments comprised a solvent control (SC), a positive control (PC) and at least three doses of the test compound. As the recommended solvents did not interfere with the low background DNA damage of untreated eggs, negative controls were omitted. Cyclophosphamide (CP; 0.05 mg CP/egg in aqua DI) was used as PC to induce a moderate increase in MN rate without causing remarkable general toxicity.

At the end of experiments each control or dose group of valid experiments needed to be represented by six viable eggs to be subjected to the analysis of MN frequency. The number of eggs allocated to a dose or control group at the beginning of the experiment was determined by the expected viability of eggs at the end of the experiment considering the pre-defined minimum viability of ≥40% for valid dose groups. So, for control and dose groups of known low toxicity 8–10 eggs were used, whereas 15–18 eggs were allocated to dose groups of high or unknown toxicity. It is recommended to select doses that result in high, intermediate and low viability levels.

Eggs received a single dose on Day 8 of egg development. Blood samples were then taken on Day 11 of egg development, precisely before Day 10.75 (for time schedule please refer to ref. (31)). After validation a recommendation was added to the protocol to further investigate negative studies in which only low doses of the test compound, i.e. ≤1 mg/egg, could be investigated due to strong toxicity. Respective follow-up experiments start on Day 9 of egg development, i.e. the test compound is applied 24 h later than in the standard procedure while sampling remains on Day 11 (see ref. (46)).

For validation purposes, at least two main experiments were performed to obtain information on the intra-laboratory reproducibility. For regulatory testing, a study can already be finalised after a

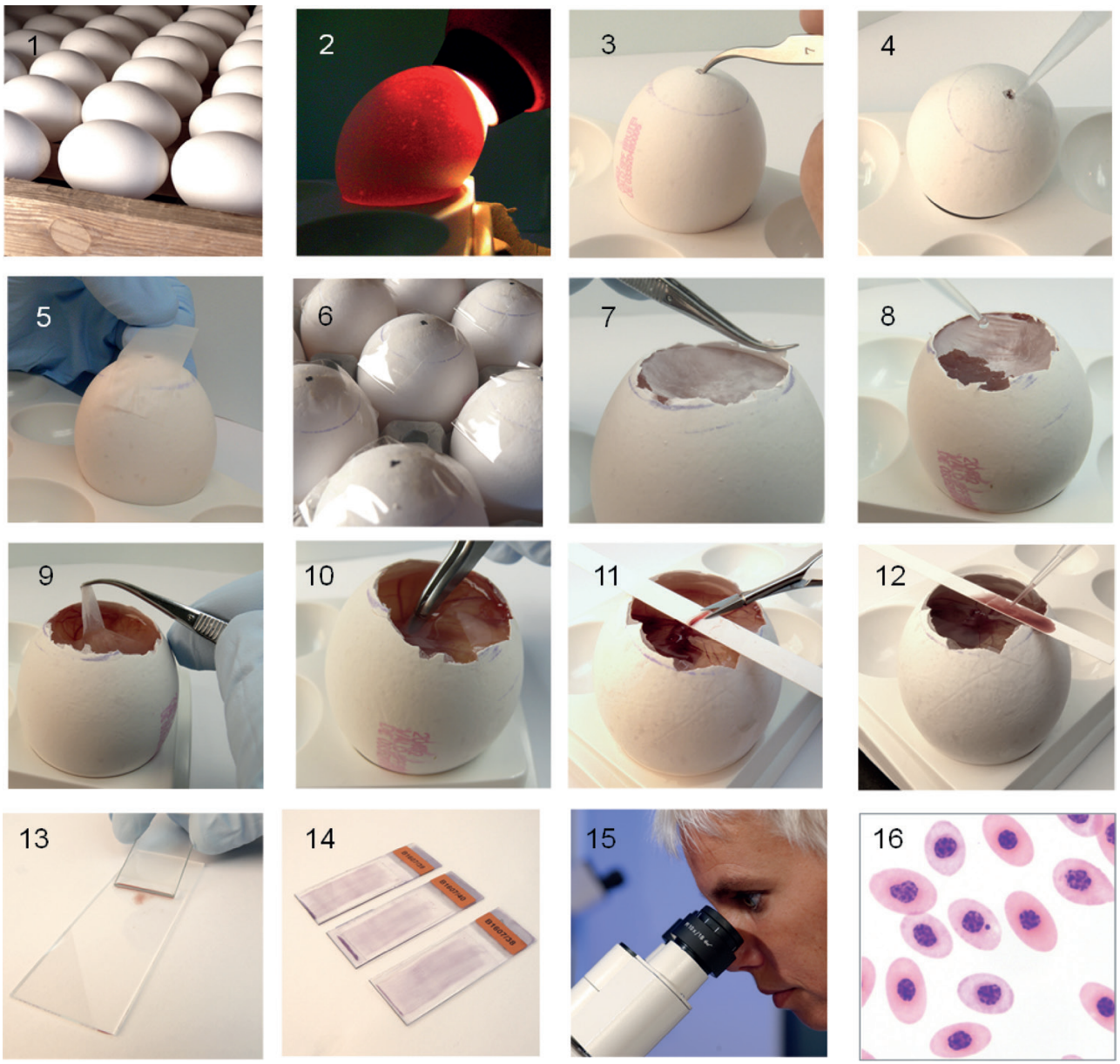


**Fig. 1.** Overview of HET-MN protocol. (1) Eggs are incubated horizontally during the first days of egg development. (2) On developmental Day 8, eggs are candled to mark the position of the air cell and to identify normally developed eggs, which are subjected to experiments. (3) First, a hole is made in the eggshell at the blunt end where the air cell is located (see blue marking). (4) Afterwards, the test compound is applied through the hole onto the inner shell membrane. (5) The hole is covered with an adhesive tape and (6) the eggs are again placed into the incubator in an up-right position. (7) At the end of the treatment period, eggs are opened widely using curved forceps. (8) Warm physiological saline is applied to the inner shell membrane (9) before it is peeled off the chorioallantoic membrane. (10) Subsequently, the first big blood vessel is identified and (11) pulled out to be positioned over a plastic strip. It is incised and (12) the first oozing blood is discarded (13) before 3–5 µl are sampled and spread on a slide. (14) Slides are stained and randomised (15) before the microscopic analysis of (16) MNs in the erythrocytes, which account for ~95% of the cells at that time of embryonic development (52). For details on the protocol, see Section Materials and methods.

clear positive call in the first main experiment. In case a second main experiment is performed, the dose spacing is usually modified by using a tighter spacing, depending on the outcome of the first main experiment.

## Protocol

### Treatment

On Day 8 of egg development, the eggs were candled to check their viability and to mark the air cell. Intact eggs with normal development were randomised and allocated to dose and control groups. By using curved forceps, a small hole was inserted into the eggshell at the blunt end where the air cell is located. The freshly prepared dosing solution was applied through the hole onto the inner shell membrane using a pipette. The applied solution was distributed homogeneously across the membrane by slightly tilting the egg while rotating the egg back and forth along its longitudinal axis. Finally, the hole was covered with adhesive tape and the incubation continued with the egg positioned in an up-right position until blood sampling on Day 11 of egg development. During validation the eggs

were additionally checked for viability on Days 9 and 10; for routine testing candling on Days 8 and 11 is regarded as sufficient (Figure 1).

**Sampling**

On Day 11 of development, eggs were candled and inspected for integrity. Six viable eggs per group were opened widely around the small hole used for application with curved forceps without damaging the inner shell membrane. Any remains of the test compound on the inner shell membrane or other peculiarities were recorded and the inner shell membrane was rinsed at least twice with warm (25°C) physiological saline. The liquid film could remain on the membrane for several minutes to wet the inner eggshell membrane, which could then be easily peeled away from the CAM using curved forceps without damaging the CAM vessels. A small incision (~1 cm long) in a weakly vascularised region of the CAM gave access to the one big blood vessel, i.e. *Arteria umbilicalis*. A loop of the vessel was pulled out and positioned on a plastic strip (e.g. pH strip) lying across the rim of the opened eggshell. The vessel was punctured using microscissors. The initial oozing blood accumulating on the plastic strip was discarded, while the later blood was sampled from the vessel opening using a pipette. A sample of 3–5 µl was spread on one (unprepared) glass slide. Three slides were prepared per egg; one was used for the analysis, two served as back-up (Figure 1).

**Slide staining**

Slides were air-dried overnight and subjected to a modified Pappenheim staining procedure: On a staining rack slides were covered for 3 min with 0.4 ml filtered May-Gruenwald solution. After adding 0.8 ml disodium citrate buffer (0.1 M; pH 5.2) for not longer than 5 min the solution was removed after a metallic gleam became visible. Then slides were thoroughly rinsed with aqua DI. Subsequently, 2 ml Giemsa working solution [3.8 ml filtered Giemsa solution mixed with 36 ml disodium citrate buffer (0.1 M; pH 5.2) for 18 slides] was applied for 20 min. The slides were thoroughly rinsed with aqua DI for 10 min, which also enabled a swelling of cells, and then air-dried. The last two steps can optionally be performed in staining dishes. Finally, slides were incubated in xylol for 20 min and mounted with cover slips using a quick-hardening mounting medium for microscopy.

**Slide analysis**

Before analysis, slides were randomised and coded to prevent operator bias. Analysis was harmonised among laboratories taking published standards into consideration (32,50,51). Using 100× objective with immersion oil, 1000 cells per egg (i.e. 1000 cells per slide) were analysed for the presence of MNs in more than 20 different regions of interest, in which the cells were well spread (see Figures 1 and 2). MNs were only considered for determining the genotoxic damage if they appeared in polychromatic and normochromatic erythrocytes (PCE and NCE, see below). Their appearance in other cell types or the occurrence of other cellular and nuclear effects was recorded and considered as alert parameters. To facilitate a transfer of the assay, an overview of the blood cell types, guidance on MN evaluation including a description of other cellular and nuclear aberrations is given below.

*Cell types*

Erythrocytes represent the vast majority of blood cells on Day 11 of egg development. Two different lineages can be observed, i.e. primitive and definitive erythrocytes, which are subdivided in different maturation stages. The processes during maturation apply to both lineages and are linked to an increasing haemoglobin concentration in the cytoplasm mediating a stronger affinity to acidic (eosinophilic) stains. In parallel, the affinity to basophilic stains declines due to a decrease in mRNAs and ribosomes in the cytoplasm. This causes the normochromatic appearance (52).

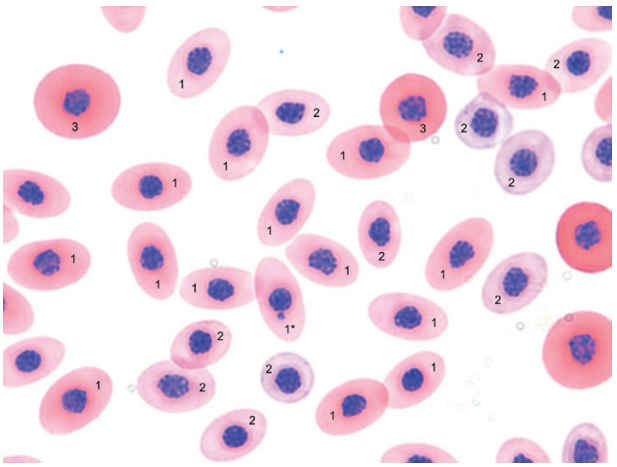

**Fig. 2.** Typical cell-type profile for erythrocytes on Day 11 of incubation. The selected region of interest shows different types of blood cells normally appearing on Day 11: (1) NCE, (1*) NCE with MN, (2) PCE and (3) E1. (Olympus light microscope, 100× objective, immersion oil.)

Definitive erythrocytes

Between Day 8 and Day 11 both blood volume and the erythrocyte number increase exponentially (35,36), the latter being mainly represented by definitive erythrocytes: They start to appear on Day 5 in the blood with a life span of 5–6 days (36). Between Day 8 and Day 11 they are mainly recruited from blood islands dispersed in the yolk sac membrane, which are surrounded by epithelial cells, which possess the metabolic capacity (33,34). This makes the egg at this developmental stage a high proliferative and very sensitive test system.

Several maturation stages of definite erythrocytes can be observed during slide analysis (32,51,52). (i) Erythroblasts, immature cells, appear only sporadically, having a greater size than all other blood cells. They have an irregular but mainly round cell shape with a large nucleus containing loosely packed chromatin (Supplementary Figure S1D–F). (ii) PCE are subdivided into early (Supplementary Figure S1G–I), middle (Supplementary Figure S1J–L) and late cells (Supplementary Figure S1M–O). During PCE maturation, the cell shape changes from round to oval and the nucleus evolves to a more compact structure. (iii) NCE represent the mature type of the definitive erythrocytes. They are oval shaped and normally have a round, centrally located (pyknotic) nucleus and show a smaller nucleus-to-cytoplasm ratio compared to PCE (Supplementary Figure S1S–U). PCE and NCE are the cells to be analysed for the presence of MNs.

Primitive erythrocytes

Prior to the occurrence of PCE and NCE, primitive erythrocytes (E1; Supplementary Figure S1A–C) develop from 36 h until Day 7 of egg development with a life span of 8 days (17,36). In consequence, only mature non-dividing primitive erythrocytes exist between Days 8 and 11 (32), which are therefore not considered for MN analysis. They can be distinguished from definitive erythrocytes by a larger and rounder shape and the smaller nucleus-to-cytoplasm ratio (36,52).

### Additional cell types

In rare cases thrombocytes and granular leukocytes can be observed (Supplementary Figure S1V–X) (36,51,52). The cytoplasm of the granular leukocytes shown here has a pale blue colour and may contain brightly red granules. The cells can widely vary in size with an irregular shape and possess a segmented nucleus (51,52). Thrombocytes can occur alone or sticking together in groups and are identified by their small size.

*Criteria for scoring MNs*

PCE and NCE were analysed regarding the presence of MNs according to published standards (17,50,53). A MN was identified based on its three-dimensionality and similarity to the main cell nucleus in terms of morphology, staining characteristics and texture. The size of a MN did not exceed one-third of the cell nucleus' size from which it was clearly separated, and the MNs were round to oval shaped (52). Cells with MNs were not further differentiated in respect to the number of MNs per cell (Supplementary Figure S2).

*Alert parameters*

Beside the appearance of MNs, other nuclear or cellular effects can be observed, which may possibly be related to the bioavailability of test compounds. Those effects were recorded as alert parameters (31,52) but were not taken into account during evaluation and to describe the genotoxic damage of a test compound. They are briefly described in the supplementing data to complete the blood picture description (Supplementary Figure S3).

## Evaluation of data

Two read-out parameters are generated with HET-MN experiments: egg viability and MN frequency. The viability (%) in a dose or control group was calculated based on the number of viable eggs at the end of the experiment on Day 11 in comparison to the number on Day 8 (Figure 3A–D). To calculate the MN frequency (‰), 1000 cells (sum of PCE and NCE only) per egg (i.e. in total 6000 cells of six eggs per group) were analysed for the presence of MNs (Figure 3E–H). This experimental design used (i.e. six eggs per dose group, scoring of 1000 cells per egg) was confirmed (54).

### Data processing

To prepare the statistical evaluation of data, individual MN counts (per 1000 cells per egg) were subjected to a Freeman–Tukey (FT) root transformation (55) to ensure variance homogeneity across dose groups (54) and to address a statistically important point arising from the presence of over-dispersed near-to-zero counts in the SC and in dose groups in the absence of genotoxic effects. The FT-transformed value $z$ was calculated from an individual MN count $x$ as

$$z = \sqrt{x} + \sqrt{x+1}$$

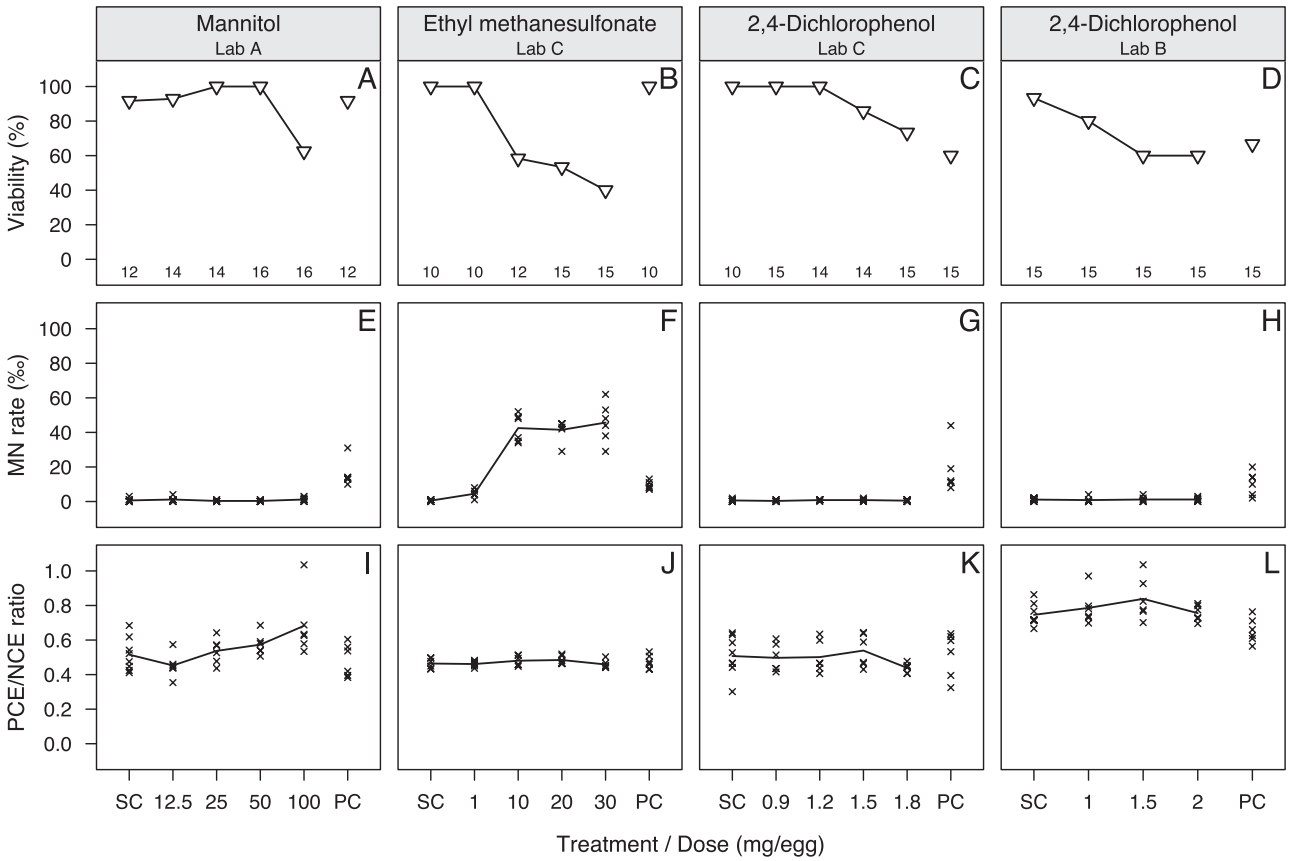


**Fig. 3.** Read-out parameters of the HET-MN assay. Example data for three substances from three laboratories showing the egg viability, the MN rate and the PCE/NCE ratio in relation to the treatment, including a SC, three-to-four dose groups, and a PC. The number of eggs used to determine the egg viability is indicated. The raw data (six samples per group) are shown for the MN rate and the PCE/NCE ratio. *Note.* The PCE/NCE ratio is calculated as the quotient $r$ = [percentage of polychromatic erythrocytes]/[percentage of normochromatic erythrocytes (including E1)] (31).

**Table 1.** Descriptive statistics of the historical control data for the MN rate

| Set no. | Data from | Solvent control (SC) | | | Positive control (PC) | | |
|---|---|---|---|---|---|---|---|
| | | ($m_{hSC}$ ± $sd_{hSC}$) | | | ($m_{hPC}$ ± $sd_{hPC}$) | | |
| | | Lab A | Lab B | Lab C | Lab A | Lab B | Lab C |
| I | Transfer phases 1 + 2 | 1.65 ± 0.37 (20) | 1.84 ± 0.33 (26) | 1.49 ± 0.38 (19) | 6.58 ± 0.94 (15) | 5.42 ± 0.96 (14) | 5.89 ± 0.92 (14) |
| II | Transfer phase 2 and validation phase 1 | 1.65 ± 0.39 (23) | 1.73 ± 0.30 (20) | 1.71 ± 0.25 (24) | 6.40 ± 1.41 (23) | 5.08 ± 0.55 (12) | 5.63 ± 1.01 (19) |
| III | Transfer phases 1 + 2 and validation phases 1–4 | 1.67 ± 0.35 (52) | 1.81 ± 0.38 (70) | 1.60 ± 0.28 (76) | 6.45 ± 1.20 (51) | 5.30 ± 0.91 (50) | 5.68 ± 0.79 (68) |

Data are given as mean ± standard deviation (FT-transformed scale) with the sample size given in parentheses.

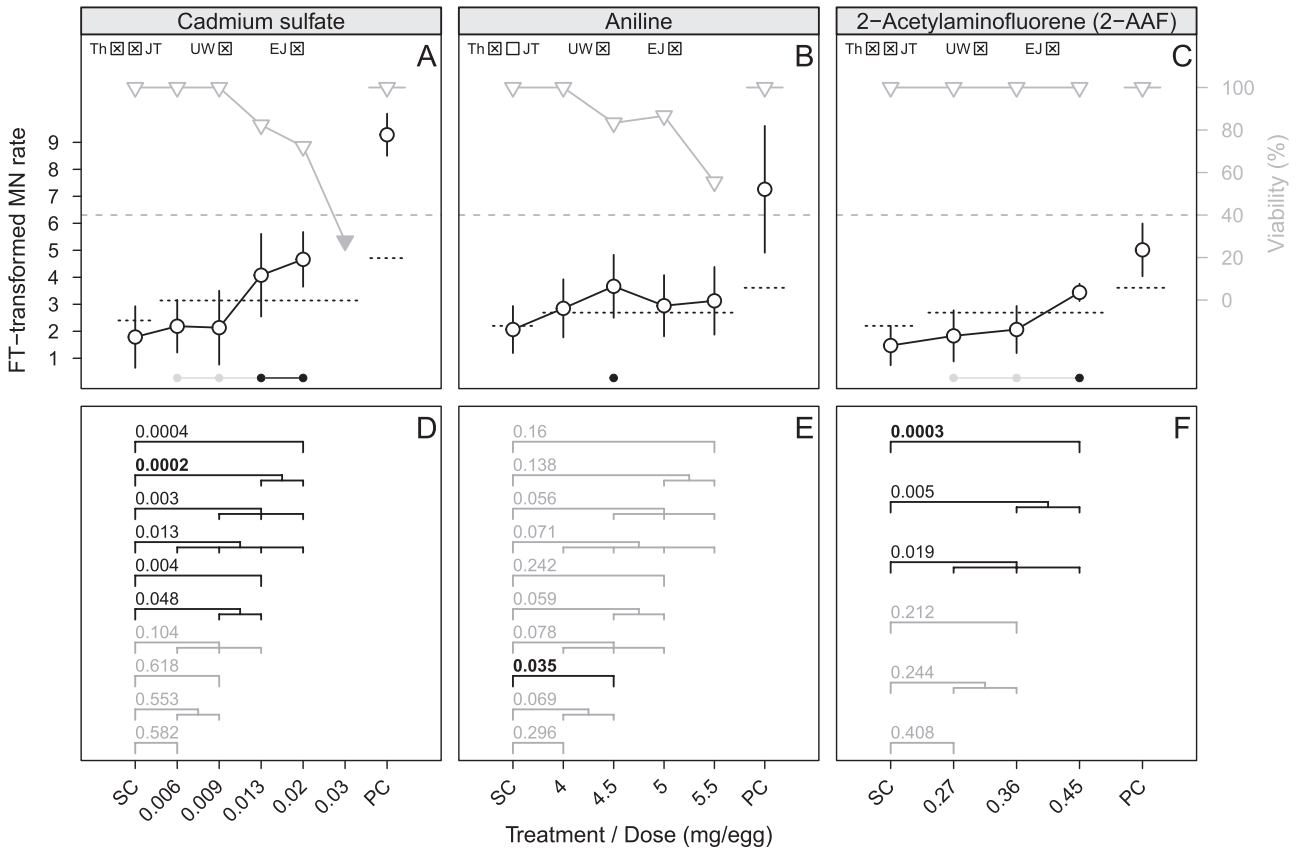


**Fig. 4.** Acceptance and statistical evaluation of test results. Upper row: FT-transformed MN rate (large open circles; left axis) and egg viability (triangles; right axis) in relation to the treatment, including a SC, several dose levels and a PC. Filled triangles indicate viabilities below 40%. MN data are given as mean ± standard deviation. Dotted horizontal lines refer to the MN rate and indicate the upper acceptance limit for the SC, the threshold for a positive call and the lower acceptance limit for the PC. MN data were tested for an increase above the threshold (Th) and for a linear trend using the JT test (prediction model 1, PM1). MN data were also analysed using the UW procedure (PM2). Finally, the result of the expert judgement (EJ) is indicated. For each test, a positive outcome is indicated by a crossed check box. The filled (individual or linked) circles at the bottom indicate single or pooled dose groups for which the UW test indicated a statistically significant increase. Lower row: *P*-values obtained by the one-sided UW test for the comparison of single and pooled dose groups with the SC. Grey colour is used for *P*-values > 0.05. The smallest significant *P*-value is shown in bold-face type.

The FT transformation is directly applied to the count raw data, i.e. before group means are calculated.

### Validity of experiments

The validity of experiments as a prerequisite for statistical evaluation was determined on the following four criteria. (i) The experiment followed the pre-defined experimental design, i.e. SC, PC and at least three dose groups of the test compound, with six eggs per group and 1000 cells scored per egg. (ii) The viability of control groups and three dose groups on Day 11 should be ≥40% (Figure 4A). (iii) For the FT-transformed MN frequency, the mean of the concurrent SC ($m_{cSC}$) should be equal to or lower than the mean of the historical SC ($m_{hSC}$) plus two times of standard deviation ($sd_{hSC}$) ($m_{cSC} \leq m_{hSC} + 2\ sd_{hSC}$; Table 1 and Supplementary Table S1; Figure 4A–C). The mean of the concurrent PC ($m_{cPC}$) needed to be equal to or higher than the mean of the historical positive control ($m_{hPC}$) minus two times the standard deviation ($sd_{hPC}$) ($m_{cPC} \geq m_{hPC} - 2\ sd_{hPC}$; Table 1 and Supplementary Table S1; Figure 4A–C). (iv) The bioavailability of

the test compound was either demonstrated by a dose-dependent decrease in viability or by an increase in MN frequency (Figure 4A–C). In the absence of these indicative effects, it is recommended to take samples from e.g. blood and/or allantoic fluid to prove the bioavailability of the parent compound, either directly or from the presence of possible metabolites (17,35,56).

Following the approach of Wolf *et al.* (30), historical control data for the PCE/NCE ratio, i.e. the ratio of polychromatic (immature) erythrocytes to normochromatic (mature) cells reflecting proliferation, were collected during the validation study (Supplementary Figure S4) to support the establishment of the HET-MN in other laboratories. The mean levels of the historical SC and PC differed somewhat between laboratories (Supplementary Figure S4). The testing of more than 30 compounds showed the PCE/NCE ratio to be very stable, and only merely changing even under dosing conditions leading to strong general toxicity (Figure 3I–L), thus confirming previous results of Wolf, Greywe and colleagues (17,29–31,52). Since this parameter turned out to be largely insensitive to the different treatments, the PCE/NCE ratio was considered as not sufficiently sensitive and therefore not being appropriate to monitor the bioavailability of test compounds. The evaluation therefore focussed on egg viability and MN frequency, which proved to be sensitive in terms of general toxicity and genotoxic damage, and thus sufficient for demonstrating bioavailability.

#### Statistical analysis of experiments

The data of valid experiments were analysed by two prediction models. The first prediction model (PM1) checked the exceedance of a pre-defined threshold; i.e. the mean of the historical negative control ($m_{hSC}$) plus four times the standard deviation ($sd_{hSC}$) (Supplementary Table S1). In addition, the Jonckheere–Terpstra (JT) test was used to check for a dose-dependent, monotone increase at a significance level of 0.025 (Figure 4). The latter is of special relevance in case of moderate increases, which do not exceed the pre-defined threshold. The outcome of PM1 was positive if the threshold was exceeded and/or if the JT test indicated a statistically significant increase (Figure 4C: JT test with *P*-value of 0.0004). PM1 was adopted from Greywe *et al.* (31) with three modifications: The model was applied to FT-transformed (instead of untransformed) data, the threshold was based on taking four (instead of three) times the standard deviation and the JT test used a significance level of 0.025 (instead of 0.05). The latter two modifications were implemented during the transfer phase to further reduce the number of misclassifications.

The second prediction model (PM2) used the one-sided Umbrella–Williams (UW) test, which is able to detect additional types of dose–response curves as it compares single as well as pooled dose groups against the SC (significance level: 0.05) (Figure 4) (54). Specifically, the UW test integrates a test for any increase in individual treatment levels (Dunnett procedure) with a test for an increasing trend against the control (standard Williams procedure) that is additionally protected against downturn effects at high doses. The outcome of PM2 was positive if at least one of the individual comparisons signalised a statistically significant increase (Figure 4).

The statistical analysis was carried out in the statistical computing environment R (57) by use of the R packages multcomp (58) and coin (59). The R package lattice (60) was used to generate the graphical figures.

#### Criteria for a positive call for experiments and studies

HET-MN data were fed into a prediction model to obtain a statistically based decision about the test outcome (positive or negative). The validation exercise used two prediction models (PM1 and PM2) in an attempt to identify the model with the best predictive performance. The statistically based test outcome was then subjected to an expert judgement, which took the biological relevance of effects into account. Specifically, the expert judgement checked whether the observed MN frequency increased above the historical control range, whether the increase showed a dose dependency and whether the effect occurred without critical reduction in viability (Figure 4). That means that the expert judgement could overrule the prediction-model decision if the response was not biologically relevant.

A positive call in the first experiment needed to be confirmed in a second experiment by demonstrating the reproducibility of the effect. A confirmatory result made the final call for the entire study positive. Specific criteria were established for dealing with discordant experiments. For example, a positive call in the first experiment was sufficient to consider the entire study as positive even if the second experiment was negative. A specific case is that of an increase (statistically significant and/or above threshold) in a single dose group without dose dependency. Such an experiment would be generally judged as not biologically relevant. However, if this effect is reproducible in the second experiment, the final call for the study would be positive.

#### Criteria for negative test results

A test result was considered negative if the experiment was valid, the criteria for a positive test result were not fulfilled and the bioavailability of the test compound was shown. That means that the dose groups did not show a biologically relevant increase in MN frequency across different experiments.

A negative result has to be critically scrutinised if at least one of the following alert parameters applies (31,52): anaemic effects, necrotic alterations or sealing effects of the CAM, test compound residues on the egg membrane, an increase of E1 (i.e. primitive erythrocytes), an increased proportion of cells in metaphase or a higher frequency of nuclear defects (e.g. budding, binucleated or multi-nucleated cells, nuclear aberration).

In case the obtained data did not provide sufficient information for a final call, the study was considered equivocal and a follow-up testing was required.

## Results and discussion

Following the development of the HET-MN as an assay in genotoxicity testing by using the MN frequency as read-out parameter (17,28–30), an inter-laboratory trial was subsequently conducted to confirm the initial results (31). The obtained data were promising and triggered the further investigation of the HET-MN in a validation study with three participating laboratories. As this was the first validation study in the field of genotoxicity applying chicken eggs the performance of the test system is elucidated in the following. The results of the validation study are presented elsewhere (see ref. (46)).

### Historical control data sets

After the first inter-laboratory trial the HET-MN was transferred to two test-naïve laboratories prior to the validation exercise. During the first transfer phase SC and PC were tested for implementing the assay in each laboratory. The second transfer phase included a first analysis of three test compounds shared blinded (data not shown). Afterwards more than 30 test compounds were being tested double-blinded in three laboratories (see ref. (46)). All valid data sets were collected for establishing a historical control data set.

During the transfer phases and the initial validation phase (validation phase 1), 7,12-dimethylbenz[*a*]antracene (DMBA) had been used as PC (0.04 mg DMBA in 50 µl IPM/egg) in experiments with test compounds formulated in IPM (Supplementary Table S1). In two out of three participating laboratories the viability of DMBA-treated control eggs was notably reduced (in exceptional cases to 40%) in comparison to CP-treated control eggs (data not shown; see ref. (46)). Because of these viability issues, and since the use of the same solvent for both PC and test compounds is not requested in genotoxicity testing, DMBA was excluded as PC during the later validation phase 1. In the following phases of the validation exercise only CP served as PC and remains the recommended PC (see Section Main experiments with dosing regime).

The historical SC mainly relies on data generated with aqua DI and IPM. These solvents are recommended with first priority for HET-MN experiments whereas methanol, acetone and pure DMSO turned out to be unsuitable as they induced local effects and/or could be toxic at higher concentrations (see also refs (17,30,46)). The use of diluted DMSO and ethanol is still advised (see Section Solubility study), but these SC data are not included in the historical SC data set due to the limited number of samples obtained during the validation exercise.

Figure 5 summarises the historical data for mean MN frequencies in the PC (only CP) and SC (only aqua DI, IPM and ethanol) as obtained by the three participating laboratories during the transfer and validation phases. For a given set of historical control data, the overall means of the historical SC ($m_{hSC}$) and the historical PC ($m_{hPC}$) as well as the corresponding standard deviations ($sd_{hSC}$, $sd_{hPC}$) were calculated. Table 1 provides the descriptive statistics for two sets of historical control data from different study phases (Set No. I and II, blue-shaded and orange-shaded 2-sigma range in Figure 5) and for the final set after finishing the validation exercise (Set No. III, not coloured in Figure 5), which includes all data for SC and PC.

The data of Set No. I (blue-shaded areas in Figure 5) were obtained in the transfer phases 1 and 2, and data of Set No. II (orange-shaded areas in Figure 5) in transfer phase 2 and validation phase 1. There was some overlap in the data of both sets, and consequently in the blue-shaded and orange-shaded areas in Figure 5, because the data from transfer phase 2 were included in both Set No. I and Set No. II. Set No. III comprised the data from all phases (up to validation phase 4). The sample size is lower for the PC since data obtained with DMBA (instead of CP) were excluded. Additionally, the sample size of Set No. III for Lab A is lower than those for the other laboratories because Lab A did not participate in the validation phase 4.

The descriptive statistics of the historical control data for the SC and PC (Table 1) were used to derive acceptance criteria and effect thresholds for the different study phases (Figure 5, Supplementary Table S1). The acceptance criteria comprised the upper limit for the concurrent SC and the lower limit for the concurrent PC (visualised as blue-shaded and orange-shaded areas in Figure 5). For example, Set II determined the lab-specific upper limit for the SC at 2.33 for Lab B (Supplementary Table S1) that is equivalent with the upper border of the lower orange-shaded area in Figure 5. The acceptance criteria and threshold for a treatment-related increase are also presented in the example data in Figure 4.

Figure 5 indicates that across the different study phases, the acceptance criteria and thresholds as derived from the historical SC and PC were generally quite stable but are influenced by the sample size. The range of the acceptance criterion for PC was clearly separated from the low background MN frequency of the SC. In Lab B, e.g. the acceptance criterion for SC at 2.33 in Set No. II was lower than in Set No. I and Set No. III (2.50 and 2.57; Supplementary Table S1) due to the smaller variation of SC data as indicated by the orange-shaded 2-sigma range. Differences of Set No. I and Set No. II may also be caused by initial training and learning effects during the transfer phase. Lab-specific differences in acceptance criteria and thresholds, being consistent across the study phases, support the decision for deriving lab-specific rather than general criteria, as recommended in OECD TG 487 (8).

The acceptance criteria and thresholds of Set No. I were applied to the evaluation of test compounds in transfer phase 2 and validation phase 1. Those of Set No. II were applied in validation phases 2–4, the main part of the validation study regarding the number of analysed test compounds in Lab B and Lab C (Supplementary Table S1). Based on the acceptance criteria the experiments were evaluated in order to determine their validity (see Section Validity of experiments). In case, e.g. the effect in the PC was not large enough to exceed the acceptance limit, the experiment had to be repeated. The evaluation of experiments during the validation phases 1–4 is presented and discussed in Reisinger *et al.* (46).

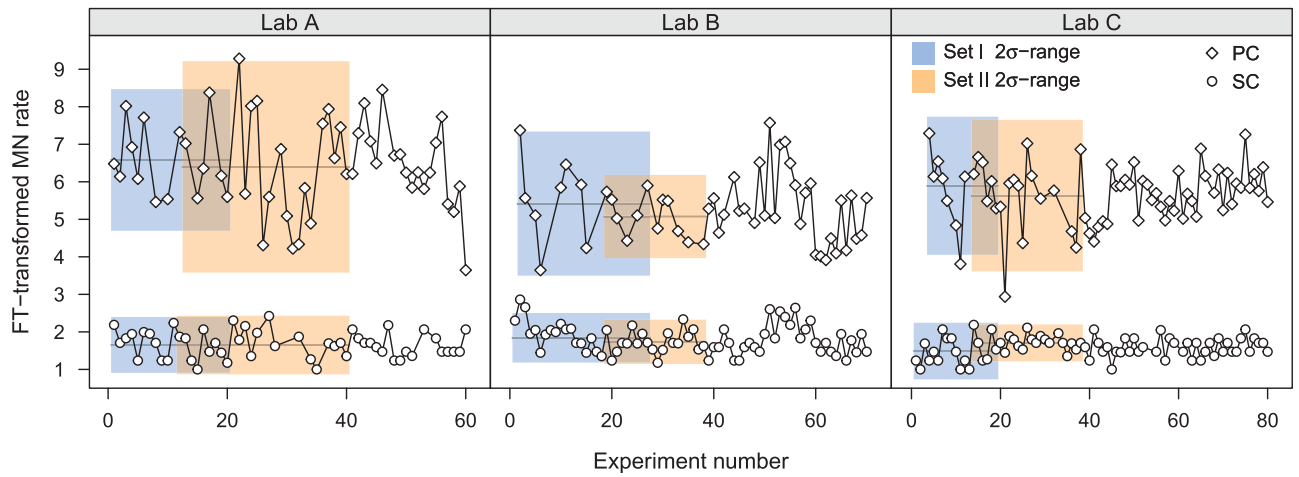


**Fig. 5.** Historical control data for the MN rate. Diamond and circle symbols represent the means of the PC and SC of individual runs as obtained by the three participating laboratories. Data are shown as FT-transformed values. The blue-shaded and orange-shaded areas indicate two-sigma ranges (i.e. mean ± 2 SD) for two subsets of historical control data, which were used to establish acceptance and threshold criteria for the different phases of the validation.

The final data set, i.e. Set No. III, consists of all historical data for mean MN frequencies in the SC (only aqua DI, IPM and ethanol) and PC (only CP) during the transfer and validation phases. The data include a range of possible sources of variation such as seasonal effects on the test system (higher temperature during egg delivery in summer), change of the egg supplier or the training of new evaluators. Hence, as a final result the Set No. III of the validation study with three laboratories can be used as guidance for future HET-MN studies.

## Evaluation of the two prediction models used

The HET-MN validation used two prediction models in an attempt to identify the model with the best predictive performance. The first prediction model (PM1) checked the HET-MN data for the exceedance of a pre-defined threshold (derived from historical SC data) and for a dose-dependent increase by using the JT test. The second prediction model (PM2) used the one-sided UW test to compare single as well as pooled dose groups against the concurrent SC (54). The UW procedure integrates a test for any increase in individual dose groups with a test for an increasing trend against the SC that is additionally protected against downturn effects at high doses.

In the HET-MN validation, 34 test compounds were tested in 123 experiments (for results and detailed discussion, please refer to Reisinger *et al.* (46)). They comprised 16 true positives (TPs), 11 misleading positives (MPs) and 7 true negatives (TNs). For 107 experiments (of 30 test compounds), the two integral procedures of PM1 and the UW procedure of PM2 gave concordant results, meaning that the respective outcomes were either all negative (77 experiments) or all positive (32 experiments). Discordant outcomes were obtained for 14 cases, comprising 13 experiments with TPs and 1 experiment with a MP. There was no case involving TNs. The overrepresentation of TPs among the experiments with discordant outcomes indicates a differential sensitivity of the three procedures in detecting certain TP test compounds. The experiments with discordant outcomes (Figure 6) were used to critically evaluate the procedures implemented in PM1 and PM2 in order to arrive at a recommendation on the best prediction model.

Several classes of response patterns/profiles were identified to be associated with discordant outcomes among the three procedures. This included sub-threshold increases as well as non-linear increases with, e.g. plateau shapes or umbrella-like profiles (Figure 6). Another cause for a discordant outcome was experiments where the response in the SC was above the acceptance range.

Sub-threshold increases were detected by the JT test (part of PM1) and/or the UW test (PM2) (Figure 6A–H). Almost all of these response patterns occurred in experiments with TPs. This suggests a

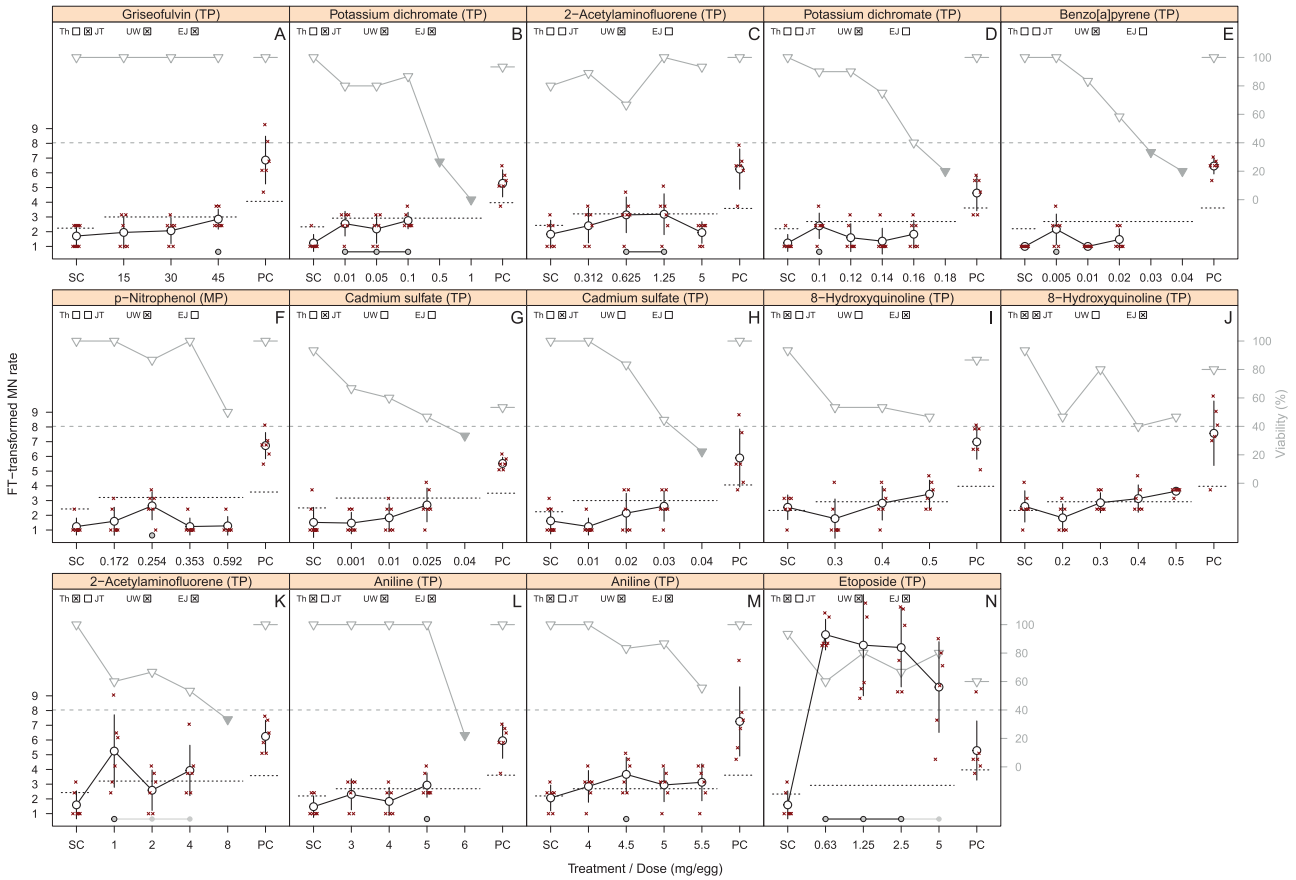


**Fig. 6.** Experiments with discordant prediction-model outcomes. The FT-transformed MN rate (circles; left axis) and the egg viability (triangles; right axis) are given in relation to the different treatments. Filled triangles indicate viabilities below 40%. MN data are given as mean ± standard deviation and as raw data (small cross symbols). Dotted horizontal lines refer to the MN rate and indicate the upper acceptance limit for the negative control (SC), the threshold for a positive call and the lower acceptance limit for the PC. MN data were tested for an increase above the threshold (Th) and for a linear trend using the JT test (prediction model 1, PM1). MN data were also analysed using the UW procedure (PM2). Finally, the result of the expert judgement (EJ) is indicated. For each test, a positive outcome is indicated by a crossed check box. Filled (individual or linked) circles above the *x*-axis indicate single or pooled-dose groups for which PM2 indicated a statistically significant increase; circles with a black outline indicate single or pooled dose groups with the smallest significant *P* value.

certain insensitivity of the threshold procedure being the other part of PM1. This PM1 threshold is a stringent one, implying a very low probability for an exceedance by chance. It was calculated from historical SC data from the mean plus four times the standard deviation. The decision to base the threshold on four times the standard deviation, instead of three times (see ref. (31)), was made at a very early stage of the validation exercise in an attempt to reduce the false-positive rate. Indeed, the probability to generate a 'false alarm', by exceeding the 4-sigma threshold by chance, is very low. For a normally distributed response variable and an experimental design with four dose groups, the probability of threshold exceedance by at least one dose group is 0.0001. False-positive rate reduction is usually gained at the expense of reduced sensitivity. For PM1, however, the reduced sensitivity of the threshold procedure turned out to be compensated by inclusion of the JT trend test, which was able to recognise sub-threshold increases (Figure 6A and B). It is to be noted that also the UW test could serve this compensatory function if combined with the threshold procedure. There were two cases, where the positive outcome of the JT trend test was overruled by expert judgement as the sub-threshold increase was only slight and not reproducible in follow-up experiments (Figure 6G and H).

In experiments with sub-threshold increases involving TP compounds, the UW test showed comparable linear trend detection and additionally recognised umbrella-like profiles, which the JT trend test failed to detect (Figure 6C). The UW test also recognised sub-threshold increases in single dose groups, which were overruled by expert judgement as being not biologically relevant due to the lack of dose dependency (Figures 6D–F). This concerned experiments with TPs and a MP.

There were two experiments with above-threshold increases where the UW test failed to detect an increase in the dose groups (Figure 6I and J). This was because of an elevated MN frequency in the SC against which the response in the dose groups is compared to. In both cases, the response in the SC was above the acceptance range. The experiments were nevertheless considered valid with reference to recommendations given in OECD TG 489 (9) (please refer to Reisinger *et al.* (46)).

In four experiments with plateau-shaped increases above the threshold, the JT test outcome was negative (Figure 6K–N). This failure is to be expected since the JT test is only able to detect near-to-linear increases. The UW test flagged all responses. It seems surprising that the JT test was unable to even detect the extreme increase above the SC in case of Etoposide (Figure 6N). However, this behaviour results from the fact the JT test is a non-parametric procedure that uses the ranks of the data rather than the data themselves. As a consequence, the extreme quantitative differences between the responses in the treatment groups and the SC are shrunk to rank differences.

To summarise, the procedures implemented in PM1 and PM2 have certain weaknesses and strengths. The chosen trigger level of the threshold procedure implicates a certain insensitivity, which, however, is compensated in PM1 by combining the threshold procedure with the JT trend test. The JT test is able to detect near-to-linear increases but it fails in case of plateau shapes or umbrella-like profiles, which can have a biological relevance. As a non-parametric, rank-based test, it also fails to detect the extreme increase above the SC. The UW test of PM2 exhibits valuable diagnostic properties to detect increases in individual doses, a monotone trend or a trend up to a peak point. Increases in individual dose groups relative to the concurrent SC could be a source of overprediction. However, spurious responses such as in the lowest or intermediate dose group would then be overruled by a well-founded expert judgement, which takes the biological relevance into account. Information on the historical SC range is essential for assessing the biological relevance.

Against the background of the identified weaknesses and strengths, the question arises which of the two prediction models is to be recommended for future HET-MN studies. The above critical evaluation, however, suggests a third option, which is a revised prediction model that combines the threshold procedure with the UW test. By benefiting from the superior trend detection capabilities of the UW test, the proposed prediction model is able to detect TP compounds, which induce shallow and/or non-linear increases in MN frequency. And by benefiting from threshold mechanism, the proposed prediction model is able to put statistically significant, sub-threshold increases in individual dose groups into the perspective of biological relevance.

## Conclusions

During the last years the HET-MN was developed and performed in different laboratories with the focus on evaluating the general performance characteristics and the predictivity of this assay. These initial studies showed a good intra- and inter-laboratory reproducibility as well as promising results concerning the predictivity. Within the scope of the validation exercise the strengths and weaknesses of the HET-MN were further evaluated and discussed.

The fertilised chicken egg is a highly complex biological system, which is assumed to be physiologically closer to the *in vivo* situation than the 2D cell systems, which are classically used in *in vitro* genotoxicity testing. The reflection of particular steps of ADME and the intrinsic metabolic capacity of the developing egg are the major advantages in comparison to classical cell culture systems.

Since the chicken egg as biological system can show variations between laboratories like other test systems used for *in vitro* genotoxicity assays, lab-specific acceptance criteria and thresholds are recommended over general criteria, which is in line with provisions in OECD TG 487 (8). For the historical control database and, thus, for the threshold for a treatment-related increase (derived from historical SC data), the sample size and the biological variation within the selected database are decisive.

Two prediction models were thoroughly evaluated for the best predictive performance within the scope of the validation exercise. The first prediction model (PM1) combined a threshold procedure with the JT test for dose-dependent increases. The second prediction model (PM2) used the one-sided UW test to compare single and pooled dose groups against the concurrent SC; this procedure integrates a test for any increase in individual dose groups with a test for an increasing trend against the SC that is additionally protected against downturn effects at high doses, a feature of umbrella-like profiles. The evaluation identified a weakness in PM1 in respect to the identification of TP compounds with shallow and/or non-linear increases in MN frequency. As a result of this critical evaluation, a revised prediction model is suggested that combines the threshold procedure with the UW test. By benefiting from the superior trend detection capabilities of the UW test, the proposed prediction model is able to detect TP compounds, which induce shallow and/or non-linear increases in MN frequency. And by benefiting from threshold mechanism, the proposed prediction model is able to put statistically significant, sub-threshold increases in individual dose groups into the perspective of biological relevance.

The open communication of the detailed protocol and the historical control data as well as the publication of the validation results may support the regulatory acceptance of the HET-MN.

## Supplementary data

Supplementary data are available at *Mutagenesis* Online.

## Funding

This work was funded by the German Ministry for Education and Research (BMBF Funding Priority 'Replacement methods of animal experiments', funding code: 0315803) during transfer phase and validation phases 1–3 as well as by Cosmetics Europe during validation phase 4.

## Acknowledgements

The authors would like to thank Julian Tharmann (German Federal Institute for Risk Assessment, Berlin, Germany) and Dirk Dressler (BioTeSys, Esslingen, Germany) for their work related to coding and sharing of the blinded compounds. We thank Kristin Fischer (German Federal Institute for Risk Assessment, Berlin, Germany) for her excellent technical assistance and the colleagues of the BfR-lab for their support during the validation project. We thank Christel Niehaus-Rolf (University of Osnabrueck, Osnabrueck, Germany) for her support during establishing the method at the BfR. We thank Anja Prellberg (ICCR Roßdorf GmbH, Roßdorf, Germany) for excellent technical assistance.
Conflict of interest statement: None declared.

## References

1. Alépée, N., Bahinski, A., Daneshian, M., *et al.* (2014) State-of-the-art of 3D cultures (organs-on-a-chip) in safety testing and pathophysiology. *ALTEX*, 31, 441–477.
2. Jennings, P., Corvi, R. and Culot, M. (2017) A snapshot on the progress of in vitro toxicology for safety assessment. *Toxicol. In Vitro*, 45, 269–271.
3. Pfuhler, S., Fellows, M., Van Benthem, J., *et al.* (2011) In vitro genotoxicity test approaches with better predictivity: summary of an IWGT workshop. *Mutat. Res.*, 723, 101–107.
4. Zeiger, E., Gollapudi, B., Aardema, M. J., *et al.* (2015) Opportunities to integrate new approaches in genetic toxicology: an ILSI-HESI workshop report. *Environ. Mol. Mutagen.*, 56, 277–285.
5. Corvi, R. and Madia, F. (2017) In vitro genotoxicity testing—can the performance be enhanced? *Food Chem. Toxicol.*, 106, 600–608.
6. Hartung, T. (2014) 3D—a new dimension of in vitro research. *Adv. Drug Deliv. Rev.*, 69–70, vi.
7. OECD (2016) *OECD Guidelines for the Testing of Chemicals, Test No. 474: Mammalian Erythrocyte Micronucleus Test, Section 4*. OECD Publishing, Paris, https://doi.org/10.1787/9789264264762-en (accessed July 29, 2021).
8. OECD (2016) *OECD Guidelines for the Testing of Chemicals, Test No. 487: In Vitro Mammalian Cell Micronucleus Test, Section 4*. OECD Publishing, Paris, https://doi.org/10.1787/9789264264861-en (accessed July 29, 2021).
9. OECD (2016) *OECD Guidelines for the Testing of Chemicals, Test No. 489: In Vivo Mammalian Alkaline Comet Assay, Section 4*. OECD Publishing, Paris, https://doi.org/10.1787/9789264264885-en (accessed July 29, 2021).
10. OECD (2019) *OECD Guidelines for the Testing of Chemicals, Test No. 431: In Vitro Skin Corrosion: Reconstructed Human Epidermis (RHE) Test Method, Section 4*. OECD Publishing, Paris, https://doi.org/10.1787/9789264264618-en (accessed July 29, 2021).
11. OECD (2021) *OECD Guidelines for the Testing of Chemicals, Test No. 439: In Vitro Skin Irritation: Reconstructed Human Epidermis Test Method, Section 4*. OECD Publishing, Paris, https://doi.org/10.1787/9789264242845-en (accessed July 29, 2021).
12. OECD (2019) *OECD Guidelines for the Testing of Chemicals, Test No. 492: Reconstructed Human Cornea-like Epithelium (RhCE) Test Method for Identifying Chemicals not Requiring Classification and Labelling for Eye Irritation or Serious Eye Damage, Section 4*. OECD Publishing, Paris, https://doi.org/10.1787/9789264242548-en (accessed July 29, 2021).
13. Shah, U. K., Mallia, J. O., Singh, N., Chapman, K. E., Doak, S. H. and Jenkins, G. J. S. (2018) A three-dimensional in vitro HepG2 cells liver spheroid model for genotoxicity studies. *Mutat. Res. Genet. Toxicol. Environ. Mutagen.*, 825, 51–58.
14. Messner, S., Fredriksson, L., Lauschke, V. M., Roessger, K., Escher, C., Bober, M., Kelm, J. M., Ingelman-Sundberg, M. and Moritz, W. (2018) Transcriptomic, proteomic, and functional long-term characterization of multicellular three-dimensional human liver microtissues. *Appl. In Vitro Toxicol.*, 4, 1–12.
15. Hurrell, T., Kastrinou-Lampou, V., Fardellas, A., *et al.* (2020) Human liver spheroids as a model to study aetiology and treatment of hepatic fibrosis. *Cells*, 9, 964.
16. Li, F., Cao, L., Parikh, S. and Zuo, R. (2020) Three-dimensional spheroids with primary human liver cells and differential roles of Kupffer cells in drug-induced liver injury. *J. Pharm. Sci.*, 109, 1912–1923.
17. Wolf, T. and Luepke, N. P. (1997) Formation of micronuclei in incubated hen's eggs as a measure of genotoxicity. *Mutat. Res.*, 394, 163–175.
18. Sinclair, J. F. and Sinclair P. R. (1992) Avian cytochrome P450. In Schenkmann, J., Greim, H. (eds), *Handbook of Experimental Pharmacology*, Vol. 105. Springer, Heidelberg, pp. 259–277.
19. Ignarro, L. J. and Shideman, F. E. (1968) Catechol-O-methyl transferase and monoamine oxidase activities in the heart and liver of the embryonic and developing chick. *J. Pharmacol. Exp. Ther.*, 159, 29–37.
20. Burchell, B., Wishart, G. J. and Dutton, G. J. (1974) Relation between the induction of hydroxylation and of glucuronidation in chick liver. *FEBS Lett.*, 43, 323–326.
21. Collett, R. A. and Ungkitchanukit, A. (1979) Phenol sulphotransferase in developing chick embryo [proceedings]. *Biochem. Soc. Trans.*, 7, 132–134.
22. Hamilton, J. W. and Bloom, S. E. (1986) Correlation between induction of xenobiotic metabolism and DNA damage from chemical carcinogens in the chick embryo in vivo. *Carcinogenesis*, 7, 1101–1106.
23. Lorr, N. A. and Bloom, S. E. (1987) Ontogeny of the chicken cytochrome P-450 enzyme system. Expression and development of responsiveness to phenobarbital induction. *Biochem. Pharmacol.*, 36, 3059–3067.
24. Gilday, D., Gannon, M., Yutzey, K., Bader, D. and Rifkind, A. B. (1996) Molecular cloning and expression of two novel avian cytochrome P450 1A enzymes induced by 2,3,7,8-tetrachlorodibenzo-p-dioxin. *J. Biol. Chem.*, 271, 33054–33059.
25. Bentivegna, C. S., Ihnat, M. A., Baptiste, N. S. and Hamilton, J. W. (1998) Developmental regulation of the 3-methylcholanthrene- and dioxin-inducible CYP1A5 gene in chick embryo liver in vivo. *Toxicol. Appl. Pharmacol.*, 151, 166–173.
26. Kiep, L. (2005) Metabolism of xenobiotics in the incubated hen's egg: investigations with ethyl 4-hydroxybenzoate. *ALTEX*, 22, 135–141.
27. Wessel, J. C., Matyja, M., Neugebauer, M., Kiefer, H., Daldrup, T., Tarbah, F. A. and Weber, H. (2006) Characterization of oxalic acid derivatives as new metabolites of metamizol (dipyrone) in incubated hen's egg and human. *Eur. J. Pharm. Sci.*, 28, 15–25.
28. Wolf, T., Niehaus-Rolf, C. and Luepke, N. P. (2002) Some new methodological aspects of the hen's egg test for micronucleus induction (HET-MN). *Mutat. Res.*, 514, 59–76.
29. Wolf, T., Niehaus-Rolf, C. and Luepke, N. P. (2003) Investigating genotoxic and hematotoxic effects of N-nitrosodimethylamine, N-nitrosodiethylamine and N-nitrosodiethanolamine in the hen's egg-micronucleus test (HET-MN). *Food Chem. Toxicol.*, 41, 561–573.
30. Wolf, T., Niehaus-Rolf, C., Banduhn, N., Eschrich, D., Scheel, J. and Luepke, N. P. (2008) The hen's egg test for micronucleus induction (HET-MN): novel analyses with a series of well-characterized substances support the further evaluation of the test system. *Mutat. Res.*, 650, 150–164.
31. Greywe, D., Kreutz, J., Banduhn, N., Krauledat, M., Scheel, J., Schroeder, K. R., Wolf, T. and Reisinger, K. (2012) Applicability and robustness of the hen's egg test for analysis of micronucleus induction (HET-MN): results from an inter-laboratory trial. *Mutat. Res.*, 747, 118–134.

32. Bruns, G. A. and Ingram, V. M. (1973) The erythroid cells and haemoglobins of the chick embryo. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.*, 266, 225–305.
33. Dieterlen-Lievre, F. (1988) Birds. In Rourley, A. F., Ratcliffe, N. A. (eds), *Vertebrate Blood Cells*. Cambridge University Press, Cambridge, pp. 257–336.
34. Heinrich-Hirsch, B., Hofmann, D., Webb, J. and Neubert, D. (1990) Activity of aldrinepoxidase, 7-ethoxycoumarin-O-deethylase and 7-ethoxyresorufin-O-deethylase during the development of chick embryos in ovo. *Arch. Toxicol.*, 64, 128–134.
35. Romanoff, A. (1960) *The Avian Embryo, Structural and Functional Development*. McMillan, New York.
36. Lemez, L. (1964) The blood of chick embryos: quantitative embryology at a cellular level. *Adv. Morphog.*, 4, 197–245.
37. Jones, J. F. (1983) Development of the spleen. *Lymphology*, 16, 83–89.
38. Hamburger, V. and Hamilton, H. L. (1951) A series of normal stages in the development of the chick embryo. *J. Morphol.*, 88, 49–92.
39. Peters, J. J., Vonderahe, A. R., and Powers, T. H. (1956) The functional chronology in developing chick nervous system. *J. Exp. Zool.*, 133, 505–518.
40. Rosenbruch, M. (1997) The sensitivity of chicken embryos in incubated eggs. *ALTEX*, 14, 111–113.
41. EU (2010) *Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the Protection of Animals used for Scientific Purposes*. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32010L0063&from=EN (accessed July 29, 2021).
42. United Kingdom (2006) Animal Welfare Act. http://www.legislation.gov.uk/ukpga/2006/45/contents (accessed July 29, 2021).
43. Der Schweizer Bundesrat (2008) 455.1 Tierschutzverordnung. https://www.admin.ch/opc/de/classified-compilation/20080796/index.html (accessed July 29, 2021).
44. United States of America (2013) Animal Welfare Amendment Bill. http://www.legislation.govt.nz/bill/government/2013/0107/latest/whole.html#DLM5174807 (accessed July 29, 2021).
45. New Zealand (2008) Animal Welfare Act. https://www.nal.usda.gov/awic/animal-welfare-act (accessed July 29, 2021).
46. Reisinger K, Fieblinger D, Heppenheimer A, *et al*. (2022) The hen's egg test for micronucleus-induction (HETMN): validation data set. *Mutagenesis*, 37, 61–75.
47. OECD (2005) *OECD Series on Testing and Assessment Number 34 Guidance Document on the Validation and International Acceptance of New or Updated test Methods for Hazard Assessment*. https://ntp.niehs.nih.gov/iccvam/suppdocs/feddocs/oecd/oecd-gd34.pdf (accessed July 29, 2021).
48. Reisinger, K., Dony, E., Wolf, T., Maul, K. (2019) Hen's egg test for micronucleus induction (HET-MN). In Dhawan, A., Bajpayee, M. (eds), *Genotoxicity Assessment. Methods in Molecular Biology*, Vol. 2031. Humana, New York, NY, USA.
49. Scheel, J., Heppenheimer, A., Lehringer, E., Kreutz, J., Poth, A., Ammann, H., Reisinger, K. and Banduhn, N. (2011) Classification and labeling of industrial products with extreme pH by making use of in vitro methods for the assessment of skin and eye irritation and corrosion in a weight of evidence approach. *Toxicol. In Vitro*, 25, 1435–1447.
50. Fenech, M., Chang, W. P., Kirsch-Volders, M., Holland, N., Bonassi, S. and Zeiger, E.; HUman MicronNucleus Project (2003) HUMN project: detailed description of the scoring criteria for the cytokinesis-block micronucleus assay using isolated human lymphocyte cultures. *Mutat. Res.*, 534, 65–75.
51. Lucas, A. M. and Jamros, C. (1961) In U.S. Department of Agriculture (ed.), *Atlas of Avian Hematology*. U.S. Government Printing Office, Washington, DC, USA.
52. Wolf, T. (2003) *Evaluierung der versuchstierfreien Genotoxizitätsprüfung am angebrüteten Hühnerei (Hen's Egg Test for Micronuclei-Induction, HET-MN)*. Tectum, Osnabrück, pp. 1–418.
53. Müller W. U. and Streffer, C. (1994) Micronucleus assays. In Obe, G. (ed.), *Advances in Mutagenesis Research*, Vol. 5. Springer, Berlin, pp. 1–134.
54. Hothorn, L. A., Reisinger, K., Wolf, T., Poth, A., Fieblinger, D., Liebsch, M. and Pirow, R. (2013) Statistical analysis of the hen's egg test for micronucleus induction (HET-MN assay). *Mutat. Res.*, 757, 68–78.
55. Freeman, M. F. and Tukey, J. W. (1950) Transformations related to the angular and the square root. *Ann. Math. Stat.*, 21, 607–611.
56. Neugebauer, M. (1995) Biotransformation of (+)-methamphetamine in fertile hen's eggs. *Pharmazie*, 50, 201–206.
57. R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (accessed July 29, 2021).
58. Hothorn, T., Bretz, F. and Westfall, P. (2008) Simultaneous inference in general parametric models. *Biom. J.*, 50, 346–363.
59. Hothorn, T., Hornik, K., van de Weil, M. A., Zeileis, A. (2006) A Lego system for conditional inference. *Am. Stat.*, 60, 257–263.
60. Sarkar, D (2008) *Lattice: Multivariate Data Visualization with R*. Springer, New York, NY, USA.