Regulatory Toxicology and Pharmacology 148 (2024) 105583

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

In vivo alkaline comet assay: Statistical considerations on historical negative and positive control data

Timur Tug[a,1,*], Julia C. Duda[a,1], Max Menssen[b], Shannon Wilson Bruce[c], Frank Bringezu[d], Martina Dammann[e], Roland Frötschl[f], Volker Harm[g], Katja Ickstadt[a], Bernd-Wolfgang Igl[h], Marco Jarzombek[i], Rupert Kellner[j], Jasmin Lott[h], Stefan Pfuhler[k], Ulla Plappert-Helbig[l], Jörg Rahnenführer[a], Markus Schulz[m], Lea Vaas[g], Marie Vasquez[n], Verena Ziegler[o], Christina Ziemann[j]

[a] *Department of Statistics, TU Dortmund University, Dortmund, Germany*
[b] *Institute of Cell Biology and Biophysics, Department of Biostatistics, Leibniz University Hannover, Germany*
[c] *Inotiv, Rockville, MD, USA*
[d] *Merck Healthcare KGaA, Chemical and Preclinical Safety, Darmstadt, Germany*
[e] *BASF SE, Ludwigshafen Am Rhein, Germany*
[f] *Federal Institute for Drugs and Medical Devices (BfArM), Bonn, Germany*
[g] *Bayer AG, Berlin, Germany*
[h] *Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany*
[i] *NUVISAN ICB GmbH, Preclinical Compound Profiling, Germany*
[j] *Fraunhofer Institute for Toxicology and Experimental Medicine ITEM, Hannover, Germany*
[k] *Procter & Gamble, Cincinnati, OH, USA*
[l] *Lörrach, Germany*
[m] *ICCR-Roßdorf GmbH, Rossdorf, Germany*
[n] *Helix3 Inc, Morrisville, NC, USA*
[o] *Bayer AG, Wuppertal, Germany*

ARTICLE INFO

Handling Editor: Dr. Martin Van den berg

Keywords:
Genotoxicity
DNA damage
Rat
In vivo mammalian alkaline comet assay
OECD test guideline 489
Historical control data
Descriptive statistics
Summarizing strategies
Variance components analysis

ABSTRACT

The alkaline comet assay is frequently used as *in vivo* follow-up test within different regulatory environments to characterize the DNA-damaging potential of different test items. The corresponding OECD Test guideline 489 highlights the importance of statistical analyses and historical control data (HCD) but does not provide detailed procedures. Therefore, the working group “Statistics” of the German-speaking Society for Environmental Mutation Research (GUM) collected HCD from five laboratories and >200 comet assay studies and performed several statistical analyses. Key results included that (I) observed large inter-laboratory effects argue against the use of absolute quality thresholds, (II) > 50% zero values on a slide are considered problematic, due to their influence on slide or animal summary statistics, (III) the type of summarizing measure for single-cell data (e.g., median, arithmetic and geometric mean) may lead to extreme differences in resulting animal tail intensities and study outcome in the HCD. These summarizing values increase the reliability of analysis results by better meeting statistical model assumptions, but at the cost of information loss. Furthermore, the relation between negative and positive control groups in the data set was always satisfactorily (or sufficiently) based on ratio, difference and quantile analyses.

1. Introduction

Originally developed in the 1980s (Ostling and Johanson, 1984; Singh et al., 1988), the alkaline comet assay (also known as single-cell gel electrophoresis) is nowadays an integral part of *in vivo* genotoxicity testing strategies among others for industrial chemicals, pharmaceuticals, food ingredients, biocides and pesticides. It has been

[*] Corresponding author
E-mail address: tug@statistik.tu-dortmund.de (T. Tug).
[1] Tug and Duda share the first authorship for this article.

https://doi.org/10.1016/j.yrtph.2024.105583
Received 1 November 2023; Received in revised form 26 January 2024; Accepted 18 February 2024
Available online 22 February 2024

Abbreviations

ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
3 R	Replace: Reduce: Refine
OECD	Organization for Economic Co-operation and Development
TG	Test guideline
GUM	Society for Environmental Mutation Research
LMA	low melting point agarose
NMA	normal melting point agarose
BL	Blood
LI	Liver
LU	Lung
DU	Duodenum
ST	Stomach
GLP	Good Laboratory Practice
TL	Tail length
TM	Tail moment
TI	Tail intensity
JaCVAM	Japanese Center for the Validation of Alternative Methods
PI	Propidium iodide
SG	SYBR Gold
MS	Multispot
SS	Standard slide
NC	Negative/vehicle control
PC	Positive control, HCD Historical control data
ANOVA	Analysis of Variance
ArithM	Arithmetic mean
TrArithM	Trimmed arithmetic mean
GeoM	Geometric mean
TrGeoM	Trimmed geometric mean
Med	Median
AQU	Aqueous
CEB	Cellulose-based
NIS	Non-ionic surfactants
OIL	Oil
OTH	Others
IQR	Interquartile range
IWGT	International Workshop on Genotoxicity Testing
EMS	Ethyl methanesulfonate
GI	Gastrointestinal
SAS	Statistical Analysis System
VCA	Variance components analysis
CIs	Confidence intervals

introduced into several guidelines including the *ICH guideline S2 (R1) on genotoxicity testing and data interpretation for pharmaceuticals intended for human use* (ICH S2) and the *Scientific opinion on genotoxicity testing strategies applicable to food and feed safety assessment* of the European Food Safety Authority (EFSA Scientific Committee, 2011). The *in vivo* alkaline comet assay is recommended as follow-up option for chemicals with a positive result in *vitro* gene mutation tests as stipulated by the Reach Regulation (European Chemicals Agency, 2017).

The assay detects DNA damage, i.e., DNA single- and double-strand breaks, alkali-labile sites (e.g., apurinic/apyrimidinic sites), as well as lesions resulting from incomplete DNA excision repair (Speit et al., 2015). The migration of DNA fragments towards the anode during electrophoresis of agarose-embedded and subsequently lysed single-cell suspensions represents the basic principle of the alkaline comet assay. DNA damage is finally quantified by analysing the percentage of DNA in the comet tail (tail intensity, or % DNA in tail). To further specify DNA damage, enzyme-modified versions of the alkaline comet assay using an incubation step with, e.g., formamidopyrimidine DNA glycosylase or human 8-oxoguanine DNA glycosylase before electrophoresis have been developed to enable detection of DNA lesions such as oxidized pyrimidines or purines (Muruzabal et al., 2021).

The alkaline comet assay can detect nuclear DNA damage in virtually all eukaryotic cell types. Its versatility can also be seen from the wide range of species and tissues/organs that can be assessed. The most frequently used organ in the *in vivo* mammalian alkaline comet assay is the liver, due to its high metabolic capacity and ease of isolation of single-cell suspensions. Sites of first contact like stomach, intestine, skin or lung as well as kidney, bladder and other tissues have also been investigated in comet assay experiments (Sasaki et al., 2008). The assay is not limited to proliferating cells, it can be combined with other *in vivo* genotoxicity studies like the *in vivo* micronucleus assay and can be integrated into repeated-dose toxicity studies, thereby contributing to implementation of the 3 R principles according to Russel & Burch (Vasquez, 2010, Recio et al., 2010; Bowen et al., 2011, Rothfuss et al., 2010). When evaluating published *in vivo* studies with 67 carcinogens that were negative or equivocal in the *in vivo* micronucleus test, the *in vivo* alkaline comet assay demonstrated higher sensitivity (detection of >90% of carcinogens), compared to transgenic rodent studies (TGR; detection of >50%), and unscheduled DNA synthesis (UDS; detection of <20%). The authors therefore concluded that the *in vivo* comet assay should play a more prominent role in regulatory testing strategies for detection of (rodent) carcinogens than the UDS test (Kirkland and Speit, 2008). Following an extensive international validation, led by the Japanese Center for the Validation of Alternative Methods (JaCVAM) (Uno et al., 2015), a respective OECD Guideline, i.e., No. 489 ("In Vivo Mammalian Alkaline Comet Assay") was adopted in 2014 and later updated in 2016.

During the last 20 years, different expert groups developed protocols for the conduct of the *in vivo* alkaline comet assay (Tice et al., 2000; Hartmann, 2003; Burlinson et al., 2007; Speit et al., 2015). They mainly cover aspects like doses, tissues, slide preparation, lysis, electrophoresis, measures of cytotoxicity, image analysis, and scoring. Besides recommendations on experimental procedures and regarding minimum information for reporting of comet assay studies (MIRCA; Møller et al., 2020), some publications also deal with specific aspects of the statistical evaluation of comet assay data. Amongst others, the importance of the experimental unit, summary measures, distributions, data transformations, and confidence intervals (CIs) were discussed (Lovell et al., 1999; Wiklund and Agurell, 2003; Lovell and Omori, 2008; Bright et al., 2011). Several of these statistical considerations have become an integral part of the OECD Test guideline (TG) 489 (OECD, 2016), which highlights the importance of historical control data and statistical analyses, and, albeit not detailed, gives practical advice on data processing and feasible statistical methods.

As given in OECD TG 489, study acceptance requires that "*the concurrent negative control is considered acceptable for addition to the laboratory historical negative control database*" and "*concurrent positive controls should induce responses that are compatible with those generated in the historical positive control database*". Thus, to assess the validity of an *in vivo* comet assay study, compilation and analysis of historical control data is key. To finally evaluate the outcome of a study (positive or negative) it must be analysed whether any "*test concentrations exhibit a statistically significant increase compared with the concurrent negative control*", whether "*there is (a) concentration-related increase when evaluated with an appropriate trend test*" and whether the "*results are inside or outside the distribution of the historical negative control data*". This is even more important when results do not fulfill all criteria for a clear negative or positive result. In such cases, the outcome of a study is largely influenced by the quality of historical control data and the type of statistical methods used.

For these reasons, the working group "Statistics" of the German-speaking Society for Environmental Mutation Research ("Gesellschaft für Umwelt-Mutationsforschung e.V.", GUM), consisting of genetic toxicologists and statisticians from academia, the regulatory body, and industry, set out to provide recommendations for statistical analysis of *in vivo* alkaline comet assay data. A large set of male and female rat *in vivo* comet assay data (>200 studies) from five different laboratories/companies were collected, including single-cell data from several organs (liver, lung, stomach, duodenum, and blood) of negative and positive control animals. Finally, mainly male rats of different strains were used for statistical analysis comprising in total 1081 negative and 940 positive control animals (for more details see 2.2). Using this comprehensive "real-world" data set, empirical data distributions were depicted, the impact of different summary measures was analysed, and the interrelationship between negative and positive control data was described. In addition, the handling of zero-values was critically reflected, and variance components analysis was carried out based on linear random (or mixed) effect models in order to provide insights in different sources of random variation (within and between laboratory variation). Compared to previously mentioned work, we used a large real data set rather than purely simulated data or simulated data based on a few individual studies.

These in-depth investigations allowed us to better specify several aspects of data handling and statistical analysis, only briefly mentioned in OECD TG 489, and to provide further recommendations for scientists and statisticians regarding the design, data processing, and statistical analysis of *in vivo* comet assays studies.

2. Materials & methods

2.1. Data/experimental procedure

Our data set was collected from 5 laboratories (labelled "A" to "E" in the following sections) covering a period of 10 years (2008–2018). Notably, about 48% of the data were generated according to OECD TG 489, and, thus, from 2014 onward (see Table 1). There were some differences in certain slide preparation steps and in the final analysis of DNA damage between laboratories and/or over the time of performance. Details of the respective protocols are given in Table 1. In general, slide preparation from organs was performed following the recommendations of the respective versions of OECD TG 489. Experiments, which were performed before issuance of OECD TG 489 in 2014, were included in the statistical analyses, when the used protocol fulfilled the requirements of OECD TG 489.

In brief, animals were sacrificed according to the applicable local animal welfare regulations. For keeping inter-sample variability to a minimum, tissues and cell suspensions were kept ice-cold until slide preparation, and slide preparation was done within 1 h from animal sacrifice. Shortly after sacrifice the abdomen was opened and the organ of choice was prepared carefully, placed in pre-cooled buffer, and single-cell suspensions were generated from liver and lung tissue by e.g., cutting the tissue into small pieces by mincing with a scissor. In case of duodenum and stomach the cell layer of the inner surface was carefully wettened with pre-cooled buffer and scratched with a cell scraper after fixation on a preparation board, or the tissues were processed by mincing in pre-cooled mincing buffer. The minced or scratched tissues were transferred into a tube already containing pre-cooled tissue buffer, whereas blood was directly pipetted into pre-cooled tissue buffer for washing. Samples where then washed by centrifugation, subsequently mixed with low melting point agarose (LMA, about 37 °C) and pipetted onto a microscopic slide, pre-coated with normal melting point agarose (NMA). The cell-containing agarose layer was then covered with a coverslip and cooled for hardening. After removal of the coverslip a third agarose layer (LMA) was added and again a coverslip was placed on top until hardening. In the next step the coverslip was removed, and the slides were transferred into lysis buffer and incubated for at least 1 h in the refrigerator. The slides were randomly placed into a horizontal electrophoresis chamber and were covered with alkaline electrophoresis buffer. After an unwinding period, electrophoresis was started at a constant V/cm, initially adjusting the buffer volume to achieve the desired current. At the end of electrophoresis, slides were transferred to neutralization buffer followed by an optional dehydration step using ethanol and air drying or storage in a humid box before direct analysis. All steps from the removal of the organ out of the organism until electrophoresis were done on ice using pre-cooled solutions (about 4 °C) and protected from direct sunlight. The slides were analysed regarding DNA strand break induction after staining with a certain fluorescence dye using a fluorescence microscope. The slides were scored by using image analysis systems, i.e., Comet Assay III or IV (INSTEM, UK) or Komet GLP (Andor, UK).

The used data set contained raw data on a single-cell level from negative/vehicle (NC) and positive control (PC) animals of 215 comet assay experiments. Notably, the number of cells per slide (50–250), the number of slides per animal (2–5), and the number of animals per group (5–10) varied between the laboratories and certain experiments. As the data set comprised studies from 2008 to 2018, a minimum number of cells per animal according to OECD TG 489 in its updated version (150 cells per tissue per animal) was present in 48 % (420/882) of the animals only (see Table 1 and S2).

A total of 1126 animals in the negative control groups and 1109 animals in the positive control group were collected and analysed in the five organs as listed in Tables 1 and i.e., blood (BL), liver (LI), lung (LU), duodenum (DU), and stomach (ST). As the data set was most comprehensive for liver tissue, the liver was used as primary organ for most of the statistical analyses.

In general, comet assay data are structured on 3 levels, i.e., cell, slide, and animal level, with 50–100 cells per slide, and 2–3 slides per animal. In the current data set, typically 3 slides were available per animal with

Table 1
Data set overview focussing on negative (NC) and positive control (PC) animals after removal of studies with mice and female animals (see chapter 2.2).

Laboratory	Organ	Total number of animals	Number of studies	Median number and range of animals per study	Study performance	Number of NC and PC animals with analysis of				
						50 cells/ animal	100 cells/ animal	150 cells/ animal	200 cells/ animal	250 cells/ animal
A	BL	57	7	4 (3–8)	2009–2011	0	0	57	0	0
A	LI	64	8	4 (3–8)	2009–2012	0	7	51	0	6
A	LU	52	6	5 (3–8)	2009–2010	0	1	45	0	6
B	DU	230	21	5 (5–7)	2011–2017	6	93	121	10	0
B	LI	659	62	6 (5–10)	2014–2017	6	89	563	1	0
B	ST	176	16	6 (5–6)	2014–2017	5	24	147	0	0
C	LI	50	5	5 (5–5)	2008–2010	0	50	0	0	0
C	ST	50	5	5 (5–5)	2008–2010	1	49	0	0	0
D	LI	96	8	6 (6–6)	2013–2018	0	0	96	0	0
E	LI	587	53	6 (3–12)	2004–2013	0	587	0	0	0

Blood (BL), liver (LI), lung (LU), duodenum (DU), and stomach (ST).

50 cells analysed per slide. However, there were single studies, e.g., lab “E" for liver or lab “C" for liver, where, due to the common practice at that time, 100 cells from 2 slides were used per animal. Before starting statistical analyses, data quality was checked and finally found to be good. Notably, all experiments were carried out according to the principles of Good Laboratory Practice (GLP) (Chemicals Act,), Appendix 1.

On the single-cell level, different readouts can be used such as tail intensity (TI; synonyms: tail DNA or % DNA in tail), tail length (TL), and tail moment (TM; formula considering both tail intensity and tail length) (Wiklund and Agurell, 2003; Uno et al., 2015). For the present statistical analyses, TI was used as main measure, based on its linearity over a wide range and its direct correlation with the amount of DNA strand breaks, both justifying its recommendation by OECD TG 489.

In addition to cell data, the laboratories were asked to fill for each individual study two methods questionnaires asking for animal strain, study design, slide preparation, cell lysis, electrophoresis, staining and analysis and study design, vehicle, positive controls, test species, organ used, and slide preparation, respectively (see Supplement Table S1 “Method and Experiment”).

Prior to data analysis, the quality of the data was checked. Data quality was assessed using the R package dataquieR, which is described in Schmidt et al. (2021). In addition, as recommended by OECD TG 489, control charts were generated to examine stability of the historical negative control data (HCD) over time. These were created for each laboratory and organ using the ggQC R package (Grey, 2018) and can be found as Supplemental Fig. S3. Notably, most of the studies were within the control limits set by company and organ, thus supporting appropriate data quality. Within the quality control context control limits here refer to means plus/minus three standard deviations. In the rest of the manuscript, in this paper the term “control limit” will be used generically.

Many variables were collected from the almost 50 methodological questions. Unfortunately, some variables were dropped after the first analysis, because they were not identifiable (different settings between laboratories but always the same setting within a laboratory) or highly imbalanced (within a laboratory, almost always the same setting is used): Duration of electrophoresis (20, 30 or 40 min), lysis time (12, 16, 24 min), microscopic magnification (20-, 40-, 200-fold), sandwich method (Yes/No), staining (Propidium iodide (PI)/SYBR Gold (SG)), system version (4.11, 4.1.1, 7.1.0.23, Comet IV, Version 3.0), tissue sampling within 10 min (Yes/No), electrophoresis buffer (13, alkaline buffer 300 mM NaOH, 1 mM EDTA, pH > 13), type of slide (Multispot (MS)/Standard slide (SS)), unwinding time (20, 30 min) and voltage (0.7, 1 V/cm). But other factors such as vehicle type were considered in further analyses. The types of vehicles were grouped into five categories. For details, see Table S2.

2.2. Data processing

To improve homogeneity of data and, thus, interpretability of results, 88 of the 1169 (7.5%) NC animals and 81 (8.6 %) of the 1021 PC animals were removed prior to statistical analyses (Table S4). In more detail, initially, a small number of mice (6 NC and 6 PC animals) were removed, because all other experiments were performed with rats (99.5%), resulting in 2178 animals, 1163 NC and 1015 PC animals. In a second step, all experimental data of the few female rats, i.e., 82 (7.0%) NC and 75 (7.3%) PC animals were excluded. The remaining 2021 animals (1081 NC and 940 PC animals) were all male rats (strains: Sprague-Dawley, Fisher 344 or Wistar HAN), and were finally included in statistical analyses.

2.3. Basic statistical methods

Basic statistical methods used for data description refer to descriptive statistical terms and methods such as (shapes of) distributions and summary measures. An understanding of the concepts is required to grasp statistical analyses that follow, which is why a short recapitulation is offered below. It is not required to understand in detail the more elaborate statistical methods, such as mixed effect models (Section 2.4), to follow the discussions. We refer interested readers to Heumann et al. (2016) for details on basic statistical concepts and to Brown (2021) for details on mixed-effect models.

We here focused on the descriptive nature of empirical distributions and did not focus on more theoretical concepts. The empirical distribution of an observed set of data can typically be described in qualitative terms such as symmetry, (right) skewness, and uni- or bimodality. Additionally, outliers or extreme values might be present in a data set. In Fig. S5, generated data for such scenarios are depicted with the corresponding mean and median values.

In the following, relevant summary measures are explained that will be of relevance in section 3.3. For a set of observations $x_1, \dots, x_n$, the (arithmetic) mean is calculated as $\overline{x} = \frac{1}{n}(x_1 + \dots + x_n)$. For the calculation of a median, data are arranged in ascending order such that $x_{(1)} \leq \dots \leq x_{(n)}$ and the median x_{med} is the value $x_{((n+1)/2)}$ in the middle if n is odd, or the average of the two values in the middle $(x_{(n/2)} + x_{(n/2\,+\,1)})/2$ if n is even. The median is a robust measure since it is not influenced by outliers. Figure S5 a) - c) depicts how the mean is influenced by extreme values, but the median remains the same regardless of the presence of extreme values.

The geometric mean x_{geom} is calculated as $\sqrt[n]{x_1 \bullet \dots \bullet x_n}$. It averages the values on a multiplicative scale and is appropriate for right skewed data. A practical challenge for the geometric mean is the occurrence of zero-values in comet assay data sets on the single-cell level, as discussed in detail in Section 3.2.

Both arithmetic and geometric mean are directly influenced by a single value change. If the largest value is extreme or an outlier, these summarizing measures become relatively large (compared to the median). To render the estimate for arithmetic or geometric mean more robust against extreme values, one can trim or remove a certain fraction of $p\%$ of the largest and smallest values. For NC data in the comet assay, it is only necessary to trim the largest $p\%$ of the data, as the smaller values are naturally bounded by zero.

2.4. Variance components analysis

To understand the variability between and within the studies, statistical modelling was done for the TI values obtained from the livers of male rats only, as all laboratories delivered respective data. Based on linear random (or mixed) effects models (Searle et al., 2006), the total variance of the observations in each laboratory was decomposed into variance components that can explain the between vs. the within study variation following the idea of Dertinger et al. (2023).

It is noteworthy that the hierarchical structure of the HCD (different studies, several animals in each study, several slides per animal) leads to the violation of one of the key assumptions on which simple linear models (e.g., used for Analysis of Variance (ANOVA)) are based on, i.e., the assumption that all observations are (stochastically) independent from each other and hence, are uncorrelated, because all cells that belong to a certain randomization unit (e.g., a certain animal) might tend to show a similar reaction (e.g., above average TI). A common way to model such dependencies between the experimental units is the inclusion of random effects that reflect the experimental design. In so called random (or mixed) effects models, each random factor is assumed to follow a normal distribution, such that the total variance of the data can be expressed as the sum of variance components that correspond to a certain randomization structure. This kind of model can thus answer the question how much variance can be explained by the different studies, by the different animals in each study and by the slides per animal.

Generally, measured percentages (such as TI) tend to be heavily right skewed (Fig. S6) and cannot always be centered by log10-transformations, especially if they contain many zero-values (Fig. 1A,

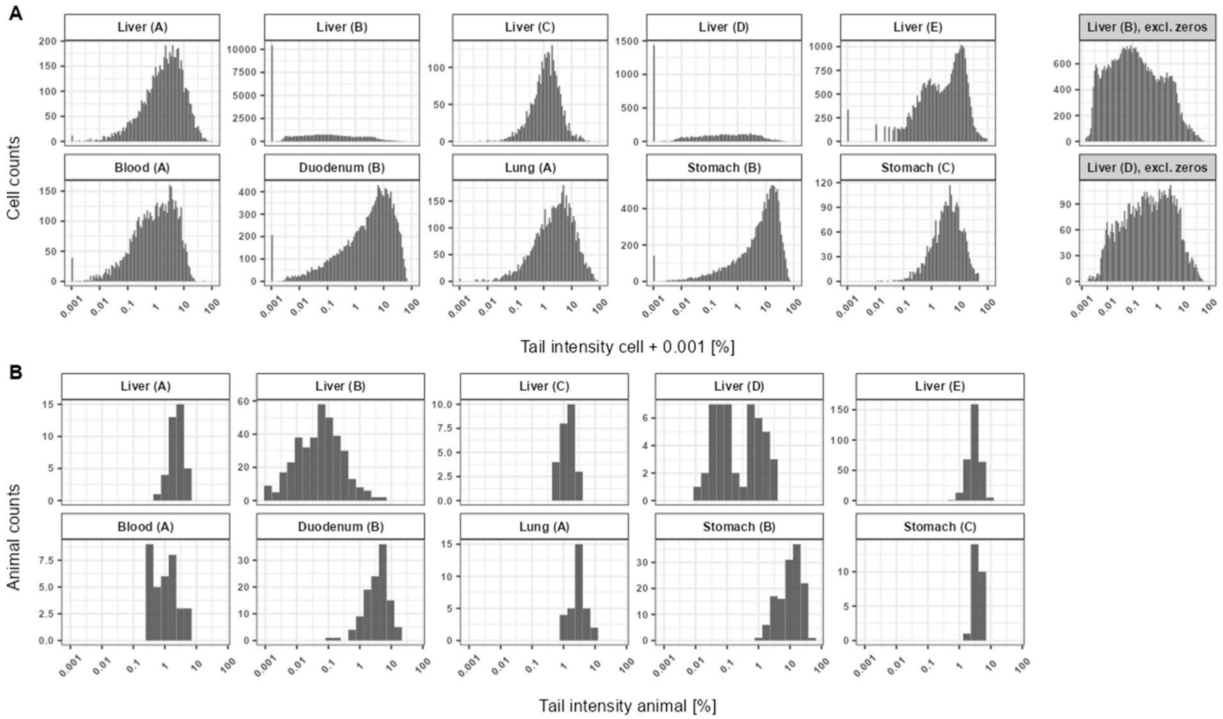


Fig. 1. (**A**) Left: Distribution of TI values for negative controls after adding + 0.001 on the cell level and log-transformation, stratified by organs and laboratories (A–E). For better interpretability, the values on the x-axis indicate the values prior to log-transformation, i.e., the original values with constant 0.001 added. Right: For laboratory B and D, the liver values are displayed again but without the peak of (original) zero-values; (**B**): Distribution of TI values for negative controls after summarizing cell-level + 0.001 values using the median and then the resulting slide-level values using the arithmetic mean on a logarithmic scale.

chapter 3.1). Hence, the assumption of normal distributed and variance homogeneous random effects is violated, if one fits a random (or mixed) effects model to the log10-transformed data on cell level (see supplementary material S7 and S7a-e).

Therefore, in a first step, the observations were aggregated as median TI per slide. Based on these data, random effects models were fit, taking studies, animals, and slides as random factors. But this procedure did not cure violation of the model assumptions up to a satisfactory level (since the right skewness persists) and therefore, was not pursued further.

In a last step, the data was aggregated on the animal level, as recommended in OECD TG 489. Initially, the median per slide was computed and then the slide medians were averaged resulting in one value per animal. Linear random (or mixed) effects models were fit to log10-transformed TI values on animal level, taking the studies as random effects, as recommended by Bright et al. (2011). If a laboratory (B and E) used different rat strains, the strains were modelled as fixed effects in order to test if the strains have an impact on TIs (via type three ANOVA using Kenward Roger approximation of degrees of freedom).

The aggregation on animal level could cure violation of the model assumptions up to a certain extent, but for some laboratories the log10-transformation led to slight right skewness (laboratory A) or slight bimodality (laboratory D). Diagnostic plots for each of the fitted models are given in the supplementary material (S7 and S7a-e). By modelling TI values aggregated on animal level, it is possible to decompose the total variance of the aggregated observations into two variance components: The between study variance that reflects the variability between historical studies and the within study (residual) variance which reflects the variability that cannot be explained by the differences between the different historical studies.

For each laboratory, the estimates for both variance components are depicted in Fig. 7 B together with their 95% confidence intervals. In a recent publication of Dertinger et al. (2023) it was stated that, if the between study variance is the major source of variation „*comparisons between study data and the HCD bounds are less useful, and consequentially, less emphasis should be placed on using HCD to contextualize a particular study's results* “.

To analyse between and within study variation, the ratio between the two estimated variance components was computed. If the between study variance is higher than the within study variance, the ratio will be greater than one, and it will be smaller than one, if the within study variance is higher than the between study variance. In other words, one can test the null hypothesis H0: between study variance/within study variance = 1. This can be done by application of a confidence interval. If the corresponding 95% confidence interval of the ratio does not cover one (the H0), it can be concluded that the two variance components differ significantly from each other (Fig. 7 in chapter 3.6).

2.5. Computational details

Data management, plotting and statistical modelling was done using R (R Core Team, 2022). Data management and plotting of the data was done based on the tidyverse packages (Wickham et al., 2019). Random and mixed effects models were fitted using lme4 (Bates et al., 2015), and type three ANOVA was done based on lmerTest (Kuznetsova et al., 2017). The CIs depicted in Fig. 7A and B in chapter 3.6 are based on a parametric bootstrap and were calculated using `lme4::bootMer()`. The QQ-plots in the supplementary materials were obtained using the hnp package (Moral et al., 2017). The R code regarding the variance components analysis is given in supplementary material S7 and S7a-e. Data quality was assessed using the R package dataquieR (Schmidt et al., 2021) and ggQC (Grey, 2018).

3. Results

3.1. Data distributions

As a rough overview, cell-level negative control values across studies and animals were agglomerated while stratifying for organ and laboratory (Supplementary S4). To obtain an overview of the data, there was no further stratification by other factors such as strain, electrophoresis time or vehicle type (cf. Chapter 3.4). For the most abundant organ liver, median TI values across all laboratories ranged from 0.050 % (B) to 3.050 % (E), whereas for stomach, the median TI values amounted to 4.022 % (C) and 8.719 % (B). Single laboratories also provided data for blood (laboratory A, median TI: 1.106 %; 0.05 and 0.95 quantiles: 0.031 and 6.981, respectively), lung (laboratory A, median TI: 2.870 %, 0.05 and 0.95 quantiles: 0.126 and 15.765, respectively), and duodenum (laboratory B, median TI: 4.109 %, 0.05 and 0.95 quantiles: 0.027 and 26.466, respectively). Control charts for data quality were generated (see supplementary S3) and can provide further initial impressions of the laboratories and organs over time.

As TI data on the single-cell level cannot take values lower than zero, a normal distribution is not expected, and a more right skewed data distribution is observed. To deal with right skewed data, a log-transformation can be applied, but only after addition of a small constant. Transformation is depicted in Fig. 1A and shows a peak of zero-values for some laboratory and organ combinations. Both abundance of zero-values and logarithmic transformations are discussed in Section 3.2. In general, logarithmic transformation obviously helped in obtaining a more normal or bell-shaped data distribution, compared to the more right skewed raw data. Please note that in the random and mixed effects models applied in this study, it is always assumed that the random effects that describe both the randomization structure and the residuals (random noise that remains unexplained by the model) follow a normal distribution (see Section 2.4 and supplementary materials). The right skewness of the empirical distribution of the NC data thus indicated that the assumption of normality for the random effects was violated.

Interestingly, a bimodal distribution on cell level was noted for liver data of laboratory E after transformation (Fig. 1A). Due to a high number of zero values, in some cases there were peaks at 0.001 at the left of the data distributions, which is caused by the added small constant of 0.001. It was obvious that the recommended transformation (OECD TG 489) does not fully yield symmetrically distributed data, as discussed in detail in Section 3.2. Please note that for most, but not all (see 3.6), statistical analyses the constant of 0.001 was added on the single-cell level, in accordance with OECD TG 489.

The main experimental unit in the *in vivo* comet assay is the animal (Lovell and Omori, 2008), and not the cell. Due to the hierarchical experimental design, the guideline recommends the aggregation of observations on animal level by the calculation of the medians of the cell level values of each slide and subsequently the arithmetic means of the slide medians. The empirical distributions of the used “real world” data set on the animal level are depicted in Fig. 1B. By respectively summarizing the data, no zero-peaks remained and for some organs, the data appeared more symmetrical (e.g., liver data of laboratory A). Notably, bimodality of the single-cell liver data of laboratory E disappeared on the animal level. In contrast, for liver data of laboratory D, tendencies towards bimodality occurred on the animal level, which were not present on cell level. Investigations showed that the study year, and related

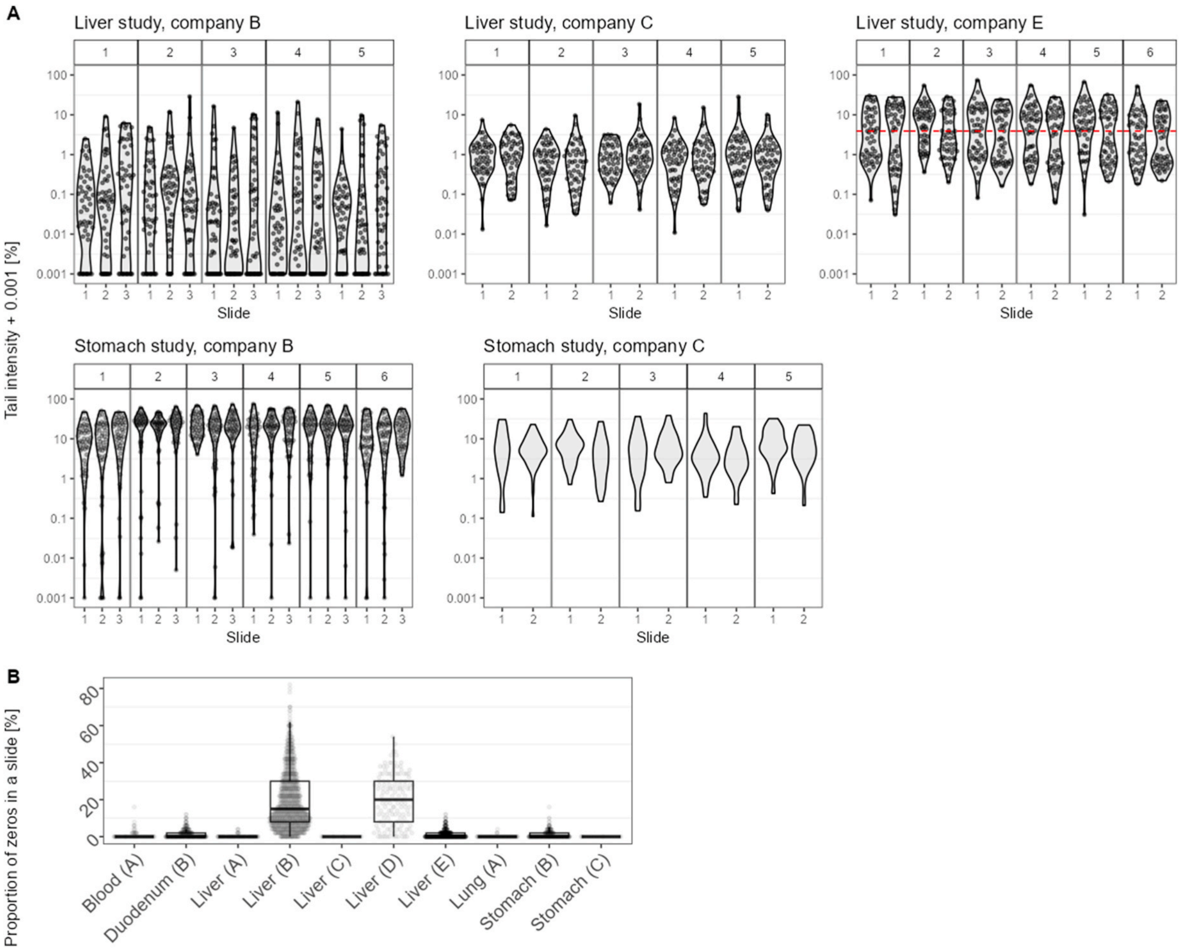


Fig. 2. (**A**) Violin plots of single-cell negative control data of five representative studies with five or six animals each for liver (top row) or stomach (bottom row) of laboratory B (left), C (middle) and E (right). The data per animal (the 5–6 panels within each sub-graphic) is subdivided into 2 or 3 available slides. For laboratory E, the red horizontal line represents the median of all negative control liver cell values + 0.001 to help to see the bimodality. (**B**) Regular boxplots of proportions of zero-values in all analysed slides stratified by organ and laboratory (A–E). The measurements of all studies for a laboratory and organ are summarized in one boxplot.

unknown factors, might be the cause of the bimodality (Fig. S8).

Fig. 2A shows laboratory specific differences based on liver and stomach negative control data from representative studies on cell level. Comparing the laboratories across the five panels (see Fig. 2A), clear differences in the TI distributions can be noted. This includes mean TI and amount of zero-values. For example, there are more zero-values within the laboratory B data set, and generally the stomach TI values are larger than those of the liver. For laboratory E the bimodal distribution of the single-cell measurements is even present within a single slide. A statistical overview of the original, negative control single-cell TI data by organ and laboratory is given in Fig. S9.

3.2. *Impact of zero handling*

Zero-values are a central challenge for statistical analyses and not rare in comet assay experiments. Zero-values on the cell level can occur in comet assays especially on slides of negative control animals, due to diverse technical reasons (Collins et al., 2014). The presence of a peak in the left part of the histograms given in Figs. S6 and 1A (chapter 3.1) reflects the amount of zero-values in the raw data sets, measured by the respective laboratory for the respective organ. For liver samples of laboratories B and D half of the analysed slides contain zero-values above 15% or 20% of total cells analysed, respectively (Fig. 2B), whereas for laboratories A and E only small amounts of zero-values were observed, and for laboratory C there were no zero-values at all for liver and stomach samples.

Zero-values can considerably complicate statistical analyses both on the descriptive and the inferential level. To describe comet assay data, summary measures are usually used to sum up the cell and slide level values in order to have a single value representing genotoxicity in a single animal. If the geometric mean is preferred, a single zero-value in the cells of a slide leads to a final zero-value for the entire slide. For example, if 50 TI values $x_1, \ldots, x_{50}$ are measured on a slide and all but one are real positive values and one is zero, then the geometric mean is

$$\sqrt[50]{x_1 \bullet \ldots \bullet x_{50}} = \sqrt[50]{0} = 0.$$

Zero-values do not only lead to unreasonable, descriptive summary measures, but also complicate statistical inference. For common statistical tests or models such as a *t*-test or ANOVA, symmetrical data within each treatment group are assumed. However, cell level negative control comet assay data are typically right skewed (Fig. S6). For right skewed data, log-transformation can be applied on the raw data to achieve better symmetry (cf. Section 2.3). However, the logarithm is only defined for positive values. For a zero-value x = 0, the logarithm is not defined.

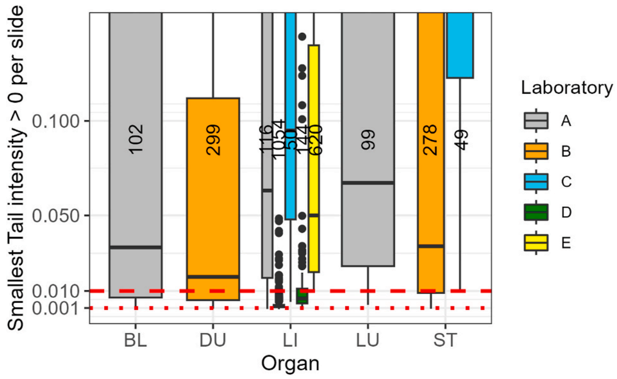


Fig. 3. *Smallest non-zero TI values* per *slide across laboratories and organs. The red dotted line indicates the OECD TG 489-proposed constant of 0.001 to be added to all data to enable processing steps such as a logarithmic transformation. The dashed line represents a heuristic alternative constant of 0.01 for comparison purposes. Blood* (BL), liver (LI), lung (LU), duodenum (DU), and stomach (ST).

A simple strategy to deal with zero-values is to add a small constant to *all* values, as suggested in OECD TG 489. The reasoning behind adding a constant to all and not just the zero-values is that by adding a small constant (0.001) to all values, variability of the data on the raw scale remains the same, as the entire data set is only shifted, and logarithmic transformation and statistical inference can be applied. Whether this strategy is reasonable for our data will be evaluated below.

But even with this strategy, challenges remain. After respective processing of the raw data on cell level, the processed values are desired to be symmetrically distributed (because only negative control data is considered). Symmetrically distributed cell measurements per slide can be adequately represented by a single value using a summary measure. The summarizing step on slide level is finally followed by a summarizing step on the animal level, such that as a result, a single value per animal can be used. To further evaluate the strategy, it was elucidated with the collected data sets, how the presence of many zero-values on the cell level affects the distribution of processed cell values and the resulting summary measures.

OECD TG 489 suggests using the median of the single-cell values per slide and the arithmetic mean to summarize the resulting slide medians on animal level (Uno et al., 2015; Tug et al., 2020). In this case, a peak of zeros in the raw data persists after adding a small constant to all data, as the peak is simply shifted to the value of the small constant, and it remains, if in the next step a transformation such as the logarithm is applied. As consequence, *any* summary measure might not represent the data adequately and can under- or overestimate the target quantity (TI), because due to the nature of a summary measure, only a single value is calculated. We illustrate that on real data of a single slide. Fig. S10 shows data of a single slide of a liver sample with 50 cell measurements. The red line indicates the arithmetic mean of (a) the raw and (b) the pre-processed data. It clearly demonstrates that the raw peak of zero-values remained after processing with its new location at $\log(0 + 0.001) = -6.91$. Hence, if many zero-values are present in raw cell data, pre-processing of the data will not lead to symmetrically distributed data. Consequently, in the summarizing step, the arithmetic mean does neither capture the center of the non-zero-values, nor does it contain information about the peak at zero. It is only a compromise not well representing the real data structure.

As for the mean, also the median can be influenced by an excess of zeros in the data. With an increasing amount of zero-values, the median eventually becomes zero. This results in a lack of representation of the values at the upper end of the distribution. Hence, a *single* value as a summary measure cannot capture the data structure appropriately, if the data consists of *two* main parts, i.e., zero-values and the remaining non-zero-values. The slide used for illustration purposes was carefully selected, for containing a relatively high amount of zero-values (Supplementary S10). However, high amounts of zero-values (>20%) can still appear regularly in some laboratories (Fig. 2A, chapter 3.1) and, therefore, should not be confused with an artificial phenomenon.

As given above, OECD TG 489 suggests adding a small constant of 0.001 to all measured cell-level values to avoid zero-values and their statistical consequences. It was, therefore, interesting whether this constant can also be deduced from the collected data sets. Obviously, the constant should generally be small as otherwise it might shift the entire data to tail intensities that indicate damage, where originally there is no damage at all. It could be argued that the smaller the constant the better, to keep the artificial shift in the data as non-influential as possible, while still enabling desired logarithmic transformations. Figure 3 shows that almost all smallest (non-zero) cell-level measurements, stratified by organ and laboratories, were above the OECD TG 489-proposed constant of 0.001 (red dotted line). From all non-negative cell-level measurements in the negative control data set (except laboratory C) only 0.14% were smaller than 0.001, with 0.000372057% as the smallest non-negative value. However, already 8.3% of the values are below 0.01, clearly indicating that increasing the constant to 0.01 seems unreasonable. The proposed constant is therefore, based on the present data set, indeed small enough, as only a neglectable amount of non-zero-values were slightly smaller than 0.001.

One might consider an even smaller constant such as 0.0001 or 10^{-5}, as it would change the data even less and, therefore, reduce a downside. But this is not true for the geometric mean or if log-transformed values are used for further statistical analyses that are typically mean-based. The smaller the constant that replaces a zero-value, the smaller the geometric mean of all values or, equivalently, the mean of log-transformed values. Note that mathematically, the mean of the logarithmic values is the same as the logarithm of the geometric mean. This is visualized in Figs. S10a and S10b, as the mean of the log-transformed values (red line), keeps decreasing if the small constant is further decreased. Hence, an increasingly small constant would not help by alternating the data less, but instead might heuristically pull the often-used log-transformed data or the geometric mean towards zero, which can lead to false-positive results. The constant depends on the data scale and a smaller constant is thus not necessarily better. In summary, our data set confirms the OECD TG 489-proposed constant of 0.001 by evaluating a large data set. As practical consequence laboratories should not round their data to two decimal places (like laboratory E) but use at least three decimal places. When using two decimal places, the smallest possible non-negative value must be set to 0.01. Non-negative values below this, like 0.004, would be rounded to zero and would then be shifted to 0.001 after adding the recommended constant. This unnecessarily changes the raw data and can easily be avoided by setting the number of decimal places to at least three.

3.3. Impact of summarizing strategies

In general, in the comet assay, tail intensities are determined on cell-level and then summarized per slide and animal afterwards. The current OECD TG 489 (2016) recommends analysing at least 150 cells per animal and organ, which can be done, e.g., by using 3 slides with 50 cells each. It is suggested to then take the median per slide and the arithmetic mean of the medians per animal. However, other summarizing statistical measures might be sensible in certain situations (see Wiklund and Agurell, 2003; Bright et al., 2011). To examine the dependency of the test result on the chosen summarizing strategy, we compared the following five measures on slide level, i.e., arithmetic mean (ArithM), median (Med), geometric mean (GeoM), trimmed arithmetic mean (remove upper 10 percent, apply ArithM on rest) (TrArithM), and trimmed geometric mean (remove upper 10 percent, apply GeoM on rest) (TrGeoM) and calculated their ArithM per animal. Summarized tail intensities of liver data per animal and laboratory are presented in Fig. 4. Tail intensities for further organs can be found in the Supplements (S11 – S14).

For negative control animals OECD TG 489 suggests that the average negative control TI should not exceed 6% for rat liver. In our data set this requirement was fulfilled for most laboratories and summarizing strategies. However, the computation of ArithM per slide tended to result in remarkably large average tail intensities per animal and consequently also per treatment group. This is underlined in Fig. 4A, where arithmetic slide means are given in green and medians in blue color, indicating clear differences, particularly based on data from laboratories B and E. In the positive control group, a slightly different result was observed (Fig. 4B), with all summarizing strategies leading to a similar outcome on a lab-specific level, i.e., differing only marginally within the different laboratories.

3.4. Effect of meta parameter

During the data collection process, several methodological meta parameters such as analysis system or duration of electrophoresis were recorded for each study (Table S1). The aim of meta data collection was to give insights into relevant settings and to identify parameters that are very influential on measured cell damage and hence statistical analyses.

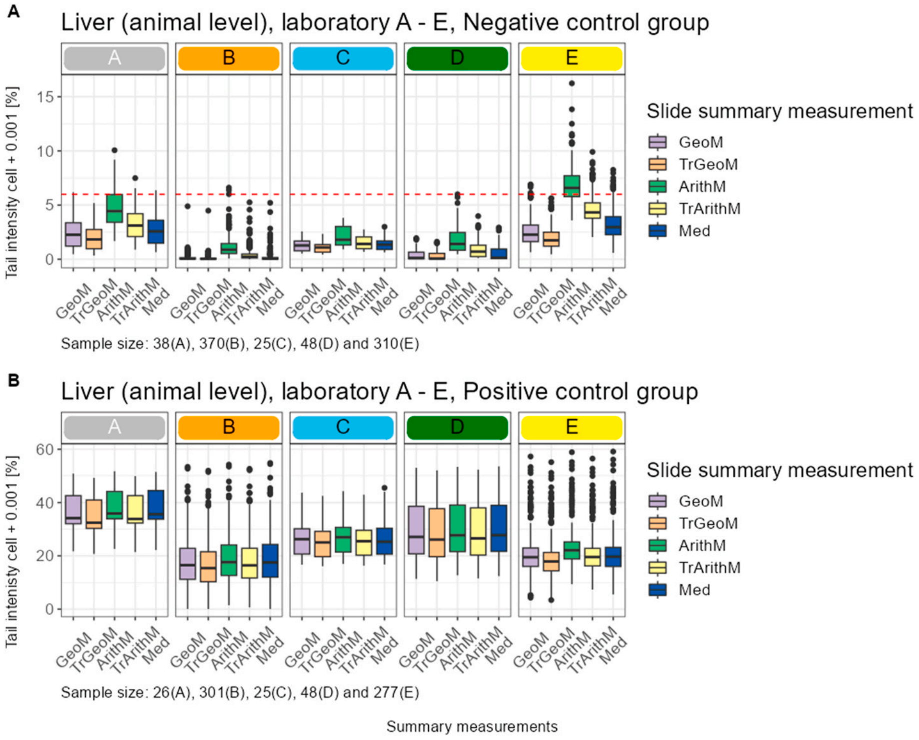


Fig. 4. (**A**): Negative control group data and (**B**) positive control group data from the different laboratories for liver tissue using five different statistical slide summary measures (GeoM: geometric mean, purple; TrGeoM: trimmed (10%) geometric mean, orange; ArithM: arithmetic mean, green, TrArithM: trimmed (10%) arithmetic mean, yellow; Med: median, blue). For animal level the arithmetic mean from all respective slide summaries was calculated. The dashed red line in the upper figure represents the 6% upper TI limit for negative control liver tissue, given by OECD TG 489.

In the following, first a descriptive overview of experimental parameters is given, followed by a discussion of the interpretative limits to these observations.

The used meta data collection sheet listing all meta parameters that were provided can be found in Table S1. For brevity reasons, we focused on a subset of these settings that were categorized in Fig. 5B. The choice for this subset was based on internal discussions and literature searches (Collins et al., 2014).

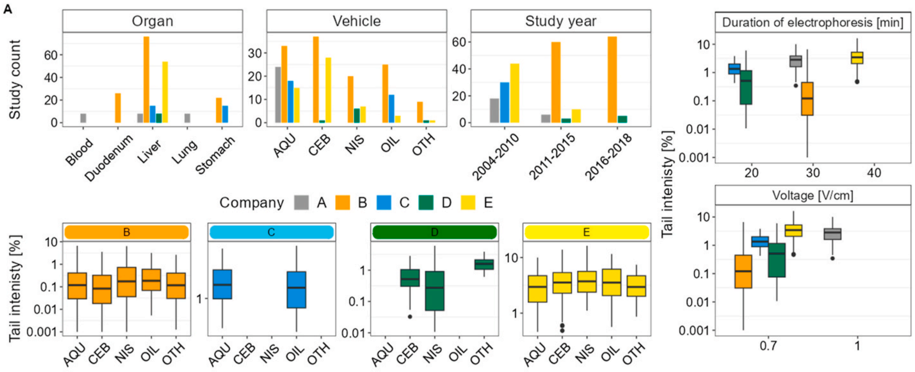
Regarding technical settings and the analysis system, each laboratory maintained a single protocol (Table S15). For example, all data generated by laboratory A used an electrophoresis duration of 30 min. In contrast, the used vehicle varied within the laboratories (Fig. 4A, upper left), depending on the respective test item. We point out that the agarose concentration was the same (0.5%) across all laboratories, which is why its effect cannot be analysed using the present data. The vehicle types provided by the laboratories were categorizes into aqueous, cellulose-based (CEB), non-ionic surfactants (NIS), oil and other (cf. Section 2.1). No vehicle type was used by all laboratories. The most common vehicle was water, used by all laboratories except laboratory E, which mostly used CEB as vehicle. Ethyl methanesulfonate (EMS) was used as sole positive control in all laboratories. However, the EMS dose varied between 125 and 300 mg/kg (not shown in Fig. 5). For details on the effect of the positive control concentration, we refer to Section 3.5. Furthermore, sample characteristics such as species, sex, and organ were collected. As described in Section 2.2, only data of male rats remain. Regarding the analysed organs, liver was the most studied organ, followed by stomach. Laboratory B furthermore conducted 26 duodenum studies, and blood and lung were analysed by laboratory A with 8 studies each. All included studies were conducted between 2004 and 2018, but laboratories B and D did not provide studies performed before 2011.

Theoretically, several parameters can influence the results of the comet assay *in vivo* (see Section 4). To evaluate which settings in which way and to what extent affect the measured DNA damage, the different settings must be distributed in the available data such that the effect is statistically *identifiable*. However, the data for this work were provided retrospectively by the different laboratories, without a chance to assign settings prior the conduction of the studies for respective statistical analyses. Consequently, within a laboratory the same technical settings were used for all studies, while certain settings differed between laboratories. Therefore, the single settings are not statistically identifiable, and the influence of the settings could not be estimated. For example, it was impossible to conclude whether the larger animal-level value for laboratory E (yellow boxplot), as compared to the other laboratories (Fig. 5A, right panel), was due to the longer electrophoresis time, the voltage of 0.7, or the analysis system (not shown). Only factors that were different within one laboratory could be examined. One factor that varied within the single laboratories was the type of vehicle. Fig. 5A (lower left) shows the liver negative control values on animal level across vehicles and laboratories (with at least two different vehicles). The data suggested that the effect of the vehicle is not robust across laboratories. For example, lower TI values for NIS, compared to CEB, as observed for laboratory D were not supported by data from laboratories B and E. Additionally, in laboratory E, the vehicle did not appear to have any effect. The analysis of the laboratory settings offers, in principle, an overview on experimental decisions. However, due to the structure and not the quality of the data, it was statistically not possible to deduce effects of the experimental parameters on the measured tail intensities across laboratories.

3.5. Comparison of negative and positive control data

An important measure for the quality of the experiment is a sufficient distance (dynamic range) between the NC and PC group. Therefore, it was initially evaluated for all studies if both NC and PC data were provided, and studies with either NC or PC missing has to be excluded from the analysis, resulting in a total of 254 studies. To ensure better comparability, the type, dose, and application mode (oral in all studies) of the PC substance were accounted for through stratification. Ethyl methanesulfonate (EMS) was used in 249 studies at 125, 200, 250, or 300 mg/kg; in the other 5 studies, no information on the used positive control was provided. All but one laboratory used the same EMS dose throughout their studies (Fig. S16). Only Laboratory E provided a study with all four available EMS concentrations. Subsequent statistical analyses were focused on liver tissue, as respective data were available from all laboratories.

Although the distances between negative and positive control groups varied, they differed clearly in most of the studies, when evaluating the 4 laboratories with a constant EMS dose (Fig. S17). In laboratory E different EMS dosages were used in the provided studies, but, regardless



B

Technical settings	Administration settings	Sample characteristics	Other
Unwinding time	Vehicle type	Species	Analysis system
Amperage	Type of positive control	Sex	Year of performance
Voltage	Dose of positive control	Organ	
Duration of electrophoresis			

Fig. 5. ***(A)*** *Upper left: Distribution of study counts based on organ, vehicle, and study conduction year across laboratories. Lower left: Negative control tail intensities for liver tissue on animal level, stratified by laboratory (with at least two vehicles) and vehicles. Right: Non-identifiability of parameter effects for negative control tail intensities for liver tissue on animal level. Since each laboratory only has one setting, the effect of the factor cannot be distinguished as it is mixed up with the overall laboratory effect. Aqueous (AQU),* cellulose-based (CEB), non-ionic surfactants (NIS), oil (OIL), and others (OTH) ***(B)*** *Subset of meta parameters provided by laboratories for each study.*

of the dose, clear differences between negative and positive control groups were noted (Fig. 6A). Thus, all laboratories demonstrated clear gaps between the negative and positive control groups, but its extent varied greatly. For laboratory E, at 125 and 300 mg/kg EMS, the first studies (first years) showed a slightly different behavior, compared to the other studies.

Therefore, we first determined the interquartile range (IQR) for the positive control groups based on the medians on slide level to see if there were differences between the respective studies and the rest of the studies. For all studies (inconspicuous and conspicuous) the IQR values of the PC data lie in a similar range (up to 8 % TI) except for one study (17 % TI). Thus, the year does not influence the IQR of the PC studies and the first conspicuous studies are in the same range as the rest of the studies. We further analysed whether the PC mean values at animal level of the conspicuous studies differ clearly from the PC values of the inconspicuous studies. And if here the year matters. For most studies from all years, the mean values were within a range of 8–30 % TI for the EMS-treated animals. Five mean PC values were larger than 38% TI, with 4 corresponding studies conducted in 2006 and 1 study in 2010. Other variables (e.g., vehicle group) were not further investigated.

For the user, the type and size of the distance between the two groups, NC and PC, matters. Previously, animal-level graphs were generated, and the differences were assessed visually. But the difference and the ratio of the two groups should better be determined at study level (arithmetic mean (study-level) of the arithmetic mean (animal-level) of the median values (slide-level)). Another stricter approach is to see what the difference or ratio is between the smallest animal value of the positive control group and the largest animal value of the negative control group.

In both scenarios, i.e. difference and ratio, a clear separation of the two groups could be observed. In the case of the difference, the positive control values were clearly positive for all laboratories (mostly above 10%), and for the ratio, the values are all clearly below 1 (Figs. S18 and S19). Overall, in no study the arithmetic mean value of the median tail intensities in the positive control group was smaller than that in the negative control group (analogous for the ratio).

However, if one looks at the animal level and compares the largest value of TI of the NC group with the smallest value of TI in the PC group, a few studies with an overlap between the two groups were present, i.e., the value in the negative control group was larger than in the positive control group (Fig. S20) For most of those studies, a detailed look revealed that mislabelling of animals lead to an overlap between NC and PC. Only for one study the data remained critical as the overlap could not be explained through closer examination.

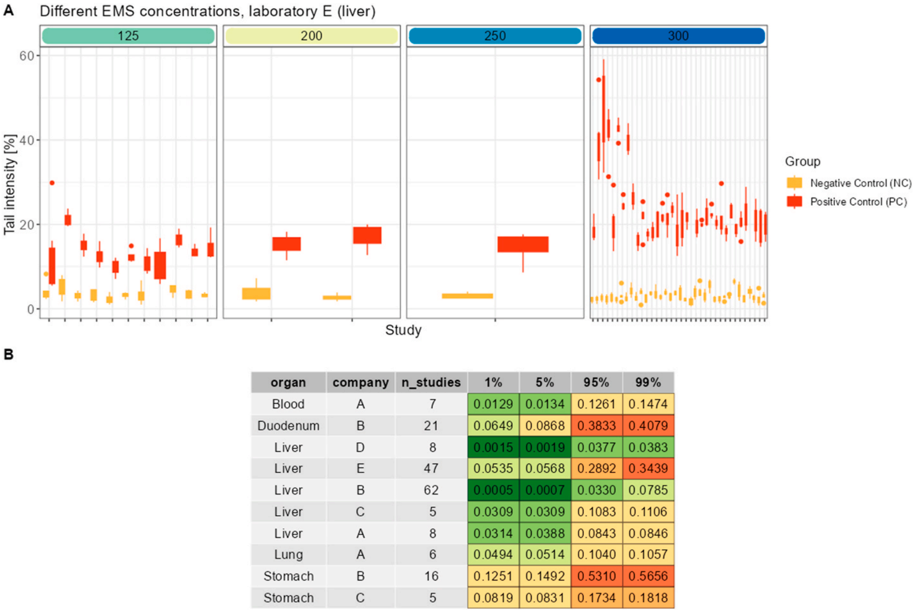
Finally, different empirical quantiles of the per study ratio between the two groups were determined for each laboratory separately. The values vary extremely for the same laboratory (Fig. 6B). Increasing values are alarming, as, e.g., a value of 0.3 means that the value for the NC is already 30% of that of the PC.

3.6. Variance components analysis

A variance components analysis was performed to quantify the sources of variability in the data. The estimated variance components for the between and within study variance as well as their corresponding 95% CIs are given in Fig. 7A. The ratios of the estimated between study variance and the estimated within study (residual) variance and corresponding 95% CIs are given in Fig. 7B. If this ratio equals one, both estimates for the variance components are the same. A ratio below one means that the between study variance is smaller compared to the within study variance (and vice versa).

From the five laboratories, only laboratories An und E fulfilled the suggestion that the estimated within study variance should be higher than the estimated between study variance (Dertinger et al., 2023). In contrast, in laboratories B, C, and D the estimated between study variance is the dominating part of the total variability. Note the low number of historical studies for laboratories A (n = 8), C (n = 5) and D (n = 8). Resulting uncertainty is reflected by corresponding large intervals in Fig. 7A. The OECD TG 489 suggests that preferably 20, but at least 10, historical studies should be available as pool for historical controls. We therefore focus on results obtained from laboratories B (n = 62) and E (n = 53). CIs.

The sufficient amount of available information for laboratories B and E resulted in substantially narrower CIs for the variance components as well as for their ratios compared to the CIs for the other three laboratories. For each of the two laboratories the proportion of variance



organ	company	n_studies	1%	5%	95%	99%
Blood	A	7	0.0129	0.0134	0.1261	0.1474
Duodenum	B	21	0.0649	0.0868	0.3833	0.4079
Liver	D	8	0.0015	0.0019	0.0377	0.0383
Liver	E	47	0.0535	0.0568	0.2892	0.3439
Liver	B	62	0.0005	0.0007	0.0330	0.0785
Liver	C	5	0.0309	0.0309	0.1083	0.1106
Liver	A	8	0.0314	0.0388	0.0843	0.0846
Lung	A	6	0.0494	0.0514	0.1040	0.1057
Stomach	B	16	0.1251	0.1492	0.5310	0.5656
Stomach	C	5	0.0819	0.0831	0.1734	0.1818

Fig. 6. *(**A**) Comparison of negative control (NC) and positive control (PC) liver data derived from studies of laboratory E with different EMS concentration. For each study NC and PC on animal level were compared regarding the gap between both groups. (**B**) Overview of the quantiles of the different laboratories and organs used. The quantiles (1%, 5%, 95%, 99%) of the ratio (NC/PC) can provide information about the quality of the studies within a laboratory. If the value approaches 1, the values of both groups are less distinguishable (red shading), and if the value approaches 0 (green shading) the NC and PC groups are better distinguishable.*

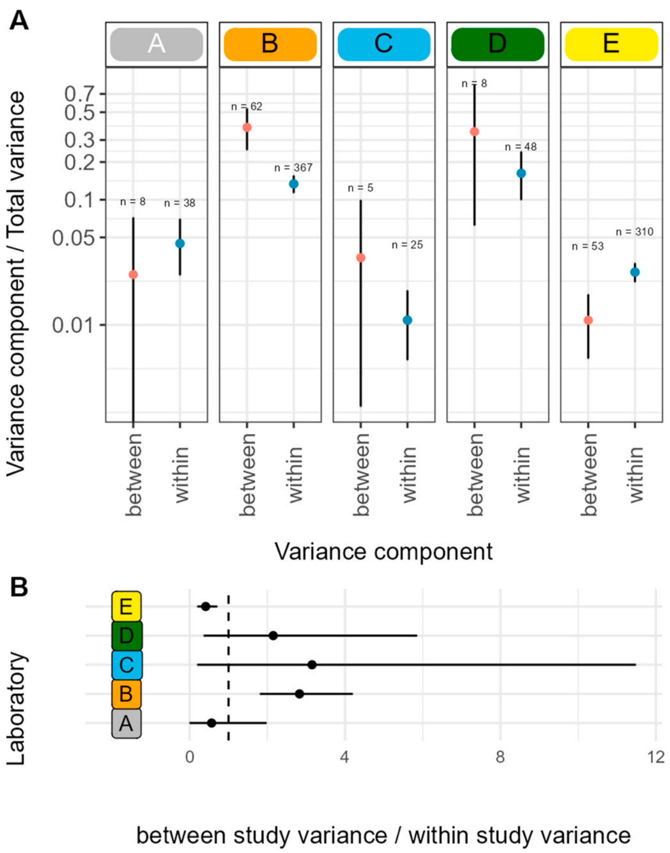


Fig. 7. (**A**) Estimated variance components. Red dots: Between study variance; Blue dots: Within study (residual) variance black bars: 95% pointwise confidence intervals, n: number of factor levels available for the estimation of the corresponding variance component. The different laboratories are given in the columns as upper cases (A–E). (**B**) Ratio of the estimated between study variance and the estimated within study (residual) variance with 95% confidence intervals. Black dots represent the estimated ratios, solid horizontal lines the 95% confidence intervals, and the black vertical dashed line the ratio 1, corresponding to the null hypothesis H0: between study variance/within study variance = 1. If the CIs include one (the dashed line), the corresponding ratio does not differ significantly from one.

components differed from each other, since their corresponding 95% CIs did not contain the one (vertical dashed line in Fig. 7B). For laboratory E, the between study variance was significantly smaller than the within study (residual) variance, whereas for laboratory B, the between study variance was significantly larger than the within study variance.

Because of these findings, the interpretation of simple point estimates ignoring their uncertainty can be heavily misleading, since the differences between the estimated variance components for laboratory A, C, and D may be caused by random variation. For both laboratories that used different strains for their experiments (B and E), the strain had no significant impact on the tail intensities measured (see supplementary material S7 and S7a-e).

4. Discussion

As a part of the 6th International Workshop on Genotoxicity Testing (IWGT), an expert working group on the comet assay evaluated critical topics related to the use of the *in vivo* comet assay in regulatory genotoxicity testing (Speit et al., 2015). The working group identified critical parameters that should be carefully controlled and described in detail in every published study report (see also Moller, 2020). *In vivo* comet assay results are more reliable if they were obtained in laboratories that have demonstrated proficiency. This includes demonstration of adequately low damage in the vehicle controls and an adequate response to a positive control for each tissue being examined. Adequate interpretation of the test data may require an extensive historical data base for the evaluated organ and scoring more cells and/or repeating experiments could help in reaching a clear conclusion in case of inconclusive data (van der Leede et al., 2014). Using valid statistical approaches, suiting the specifics of the datasets, is key to analyse experimental data. To achieve this, both toxicological and statistical expertise is needed to provide valid recommendations. Therefore, the working group "Statistics" of the GUM was established comprising both genetic toxicologists and statisticians from academia, authority, and industry. So far, the working group has focused on the analysis of historical control data of the *in vivo* micronucleus test (Igl et al., 2019). Now this work is followed by considerations on the *in vivo* alkaline comet assay to provide recommendations on statistical methods to make the best use of historical control data for valid interpretation of compound test results.

Right skewness and differences between laboratories and organs. The empirical distribution of the raw (cell level) comet assay data is an important aspect for accurate statistical analyses. Data distribution on the cell level for negative/vehicle control comet assay data is typically right skewed with well-known organ-dependent differences and inter-laboratory variation that were also present in our "real-world" data set (Fig. S6, Table S9; Lovell et al., 2020). For the two most abundant organs, (liver and stomach) the laboratory- and organ-specific median values (TI) were within the JaCVAM control limits (Uno et al., 2015), where arithmetic mean (TI, animal level) damage is stated to be preferably within 1–8% for liver (compare to our data: Fig. 4A) and within 1–30 or 1–20 % for stomach (compare to our data: Fig. S11). For blood, duodenum and lung tissue, no guideline intervals are yet available. Although the summary measure at slide level is not explicitly stated in the paper by Uno et al., (2015), we assume that the study was conducted in accordance with the guideline and that the median was used such that comparability is given. Therefore, our data set might enhance future studies of less frequently studied organs. However, our results clearly showed limitation of fixed reference values due to large inter-laboratory variability. In general, we noted that the cell-level data distribution can vary substantially between different organs and also between laboratories, and that desirable symmetrical distributions for further statistical analyses could often not be achieved by simple transformations such as adding a small constant and then performing logarithmic transformation. For one laboratory, a bimodal distribution was present that was notable even on the cell-level values of a single slide. Based on the present "real world" data set, the suggested (OECD TG 489) addition of a small constant of +0.001 could be confirmed to be adequate and useful. We therefore advise laboratories performing the *in vivo* alkaline comet assay to save and process their data with at least 3 decimal places.

Zero-values and violated statistical assumptions. Another important observation in the present data set was the varying proportion of zero-values. For one laboratory, more than half of the negative/vehicle control liver samples had more than 20% zero-values. Many zero-values prevent the appropriateness of distributional assumptions for further statistical analyses, such as normally distributed residuals. This issue persisted also after simple log10 transformations. The amount of zero-values can be influenced by experimental settings. It is known that there are certain critical experimental parameters like DNA-unwinding and electrophoresis time as well as electrophoresis conditions (temperature, current, voltage) that influence TI (Plappert-Helbig and Guérard, 2015). Moreover, the staining procedure of DNA and different comet assay image analysis systems can have an impact on TI values (Plappert-Helbig and Guérard, 2015). Therefore, there is a large variability in the amount of zero-values in different laboratories, which reflects the different experimental setups. One well known experimental setting to reduce the amount of zero-values is the electrophoresis time (Plappert-Helbig and Guérard, 2015). Increasing the electrophoresis time allows non-damaged DNA to move slightly, such that there are no or almost no zero-values. In our data set, the effect of the laboratory practice becomes evident as laboratories A and C have almost no zeros in their slides. However, it is important to note that increasing the electrophoresis time is only conducive to a certain degree. In general, it helps to reduce the amount of zero-values. But, if the electrophoresis time is very long, basal DNA damage in the negative controls might become less distinguishable from that in the treatment groups or even the positive control group. This can result in reduced statistical power for detection of genotoxic effects.

One approach to handle many zeros on the cell level is to use advanced statistical methods, e.g., zero-inflation models or Hurdle models (Rose et al., 2006). These methods are more complex and might therefore not be a feasible option for some practitioners due to a lack of statistical expertise. Therefore, after pointing out the limitations of simple statistical methods as summary measures when dealing with many zero-values and to improve the trustworthiness of statistical analyses, we encourage the experimental setups to be considerate of zero-values. For regulatory purposes, the comet assay should, in addition, only be performed by laboratories that have demonstrated proficiency. We point out that a certain amount of zero-values is acceptable, as its effect on statistical analyses is diminished after summing up the data to the animal level. However, a high number of zero-values on cell level, say >50% within a single slide, are problematic. Even after summing up single-cell data, using the median per slide, zero-values can remain and can lead to potentially false positive results by artificially lowering the negative control measurements.

Impact of summarizing strategies. Before the OECD TG 489 was issued in 2014 there were no general guidelines for statistical evaluation of individual cell data, and, therefore, laboratories used their own strategies for data handling. In addition to the well-known summary statistics such as arithmetic mean and median, also more “exotic” ones, such as the geometric mean or various trimmed means, were used.

It frequently happens that the methodology by which data is aggregated per slide and/or animal is not specified precisely or only mentioned on animal level, see e.g., the JaCVAM paper by Uno et al. (2015). In addition, the OECD TG 489 (2016) refers to individual observations to be the “endpoint” (such as the measured TI on cell level). Then, the estimated effect is defined as a difference or ratio of an average estimate of the endpoint observed in the negative control and the treatment groups. Furthermore, the literature lacks a detailed justification why different types of data aggregation are recommended. Usually, there is no distinction between the two summarizing levels, i.e., slide and animal level. However, based on a simulation study using a small data set Tug et al. (2020) concluded that the chosen summary statistics such as the mean or median has an immense impact on the final statistical test result and the outcome of the study. According to Tug et al. (2020), the difference between the summary statistics seems to become more and more negligible with increasing dose, but an extreme difference might be found at small doses or in the negative/vehicle control group. A similar effect was found in the present evaluation for all organs and laboratories.

Effect of meta parameter. One of the aims of this work was to identify relevant effects of experimental settings on negative control data. The provided data set is large and offered broad insights into the used comet assay protocols, but, due to many differences across laboratories and little variation of settings within laboratories, it was impossible to identify experimental settings and corresponding effects common to all laboratories. However, one parameter that varied also within laboratories was the chosen vehicle. For example, whereas for one laboratory the use of non-ionic surfactants compared to a cellulose-based vehicle appeared to lower the measured DNA damage in negative control liver cells, for another laboratory, it was the other way around.

These results highlight the large inter-laboratory variability, which is in line with observations of Ersson et al. (2013) and Lovell et al. (2020), hence the challenge to harmonize comet assay experiments across laboratories remains. The large inter-laboratory variability underlines that the use of fixed regulatory limits for all laboratories is critical, if the calculation of such thresholds does not account for this source of variability. However, if one accounts for inter-laboratory variability, the limits might be very wide. Hence, laboratory-specific thresholds, for example based on historical control data, should be considered and preferred. We refer to Menssen (2023) and Kluxen et al. (2021) for comprehensive overviews about the use of historical control data and the work of Menssen and Schaarschmidt (2022) on prediction intervals that are based on random effects models which can be used to set laboratory specific historical control limits.

Relation of negative and positive control data. A sufficiently high dynamic range (ratio between the largest and smallest value) of the study, as demonstrated by a high negative to positive control ratio, is considered one important factor to demonstrate proficiency of a test facility. The investigated data set demonstrated that almost all studies showed appropriately high differences between negative/vehicle and positive control data. Therefore, a sufficiently high dynamic range was not considered a frequent problem. However, when increasing the quantile of the ratio between the two control groups, the ratio increases (see Fig. 6B, chapter 3.5). Values for the PC and NC groups required for the ratio were calculated at the animal level in a guideline-compliant manner (arithmetic mean of slide medians). Small values for the ratio that are desired mean a big difference between NC and PC group. Empirical quantiles are used to look at the distributions of the ratios of each laboratory and organs combination to classify large values. At a quantile level of 95%, ratios are, e.g., >0.3 for gastrointestinal (GI) tract tissues (DU/ST) of laboratory B, i.e., that 5% of the ratios is greater than or equal to this value. The main value of historical control data is in monitoring both study quality with respect to reliability and robustness of the assay, proficiency of the test facility, and correct interpretation of the test compound results. This is especially important for borderline results, i.e., criteria for a positive result are fulfilled, but the measurements for treatment groups are still inside the range of historical negative controls. In this respect the ratio between control groups (NC and PC) represents a measure of the dynamic range and sensitivity of the assay within a test facility. Therefore, the ratio value is also a parameter for the level of confidence in statements like “the result is well within the range of the historical negative controls”. It is noteworthy to mention that the concentration of the positive control substance needs to be evaluated separately. In Fig. 6A it is demonstrated for EMS as highly potent genotoxin in one test facility that the low concentration shows a positive response but does not provide a sufficient ratio value in some cases. As seen in Fig. S18, the effect of vehicle on the negative to positive control ratio was of no relevance, in contrast to the laboratory effect.

Variance components analysis (VCA). The aggregation of observations on animal level was performed to cure violations of model assumptions and is in line with the recommendations given in the OECD TG 489 test guideline. However, aggregation of data always results in a loss of information. Hence, the estimates for the within study (residual) variance shown in Fig. 7A and B represent the only measure of intra-study variation, which, without aggregation, could have been decomposed into three variance components (animal, slide and residual). Nevertheless, the models used were in line with the recommendations of Bright et al. (2011) regarding the modelling of comet assay data and were also applied by others (Dertinger et al., 2023). We would also like to refer to a more complex, Bayesian hierachical modelling approach of comet assay data by Ghebretinsae et al. (2013).

We showed that the aggregation of tail intensities on animal level and their log10 transformation could cure the violation of model assumptions to a certain extent. Nevertheless, for some laboratories, log10 transformation resulted in slight right skewness or bimodality and thus in slight violations of model assumptions. Hence, it is questionable, if the modelling of log-transformed aggregated observations, as proposed in the test guideline, always leads to reliable and reproducible conclusions.

A possible way to overcome this problem might be the application of Box-Cox-transformations based on random (or mixed) effects models that are fit to the unaggregated observations on cell level. This would enable the researcher to analyse a current trial according to the experimental design and hence to address all sources of variability that are present in the data. Box-Cox transformation is classically used on linear (fixed) effects models. This approach is implemented in standard software like R (function boxcox (.) from R package MASS) or PROC TRANSREG in SAS, but it is not directly applicable in the context of random (or mixed) effects models, since it is based on the likelihood of the model. Methodology for Box-Cox transformation, based on linear random (or mixed) effects models, as applied for the variance components analysis shown above, is described in Gurka (2006). But, to the authors knowledge, this approach is currently not implemented in standard software like R and SAS and is, therefore, not easily applicable for toxicologists. From this point of view, the implementation of the approach of Gurka (2006) and its application to comet assay data is promising.

In their recent publication, Dertinger et al. (2023) stated: “When inter-study variation is the major source of variability, comparisons between study data and the HCD bounds are less useful, and consequentially, less emphasis should be placed on using HCD to contextualize a particular study’s results.” Based on CIs for the proportion of inter- and intra-study variance, two (B and E) out of the five laboratories showed significant differences between the variance components. For laboratory B, intra-study variance is significantly larger than inter-study variance and vice-versa for laboratory E. Following the suggestions by Dertinger et al. (2023) for our data, HCD bounds seem to be less useful for laboratory B to contextualize study data, as the variation between studies is significantly larger than the variation within a study. For laboratory E, HCD are likely useful to evaluate study responses and a HCD-derived interval. For the remaining three laboratories, no statement can be made due to too relatively few historical studies. These laboratories have less historical studies than recommended by the OECD TG 489 and corresponding (non-significant) results depend on a relatively high degree of uncertainty. Based on the results of the five laboratories, no clear tendency to whether inter- or intra-study variability is dominating across laboratories can be observed. We recommend using CIs as additional uncertainty consideration in VCAs and workflows proposed by Dertinger et al. (2023) to help prevent too hasty conclusions on inter- and intra-study variability. We note that this requires advanced statistical training, e.g. on bootstrapping approaches. However, high variation in our data set and in data analysed by Dertinger et al. (2023) suggest that adding uncertainty considerations through CIs when comparing inter- and intra-study variability is a useful extension to evaluate HCD quality.

5. Conclusion

The *in vivo* alkaline comet assay becomes increasingly important in regulatory genetic toxicology testing, considering, e.g., ICH S2 (R1) or the *Scientific opinion on genotoxicity testing strategies applicable to food and feed safety assessment* of the European Food Safety Authority (EFSA Scientific Committee, 2011). In recent years, updates of OECD technical guidance document for genotoxicity testing also draw attention towards the use of adequate statistics to be key for valid analyses and interpretation of toxicological test data. Therefore, we set the focus on addressing different statistical questions, to provide a better and more understandable statistical evaluation as a tool for genetic toxicologists. From the various statistical analyses performed, the following conclusions were drawn:

The large inter-laboratory difference in effect size measured makes it impossible to define absolute control limits to evaluate test quality. The amount of zero-values on single-cell level should be closely monitored and laboratories should avoid large amounts of zero-values by optimizing experimental settings. However, it is also acknowledged that over-optimizing experimental conditions to completely avoid any zero-value will most probably not improve the quality of the results.

From a statistical perspective, relative amounts of >50% zero-values on a single slide are considered problematic and question acceptability of the slide, as even the robust median would yield a zero-value as slide summary.

When using the geometric mean to summarize tail intensities on cell level, the investigator should be aware that a single zero-value would reduce the geometric mean to zero.

OECD TG 489 suggests adding 0.001 to all TI values prior to log or square root transformation, if necessary. The adding of a constant to the observed tail intensities is only sensible, if zeros occur in the data set, since it is a heuristic method to enable log transformation in this case. If there are no zeros in the data set, the adding of a constant is unnecessary. Especially for negative control data, it is recommended to use at least 3 decimal places when saving single-cell-level values.

In part considerable differences in summarized negative control TI values can be found between certain statistical summary measures like median, arithmetic mean, and geometric mean. These differences are not eminent in positive control TI values. This effect is likely similar for many other biological test systems, as the variance of the relative effect size is high, when representing mainly the biological background of effect, and becomes smaller with increasing biological insult, i.e., increasing dose.

The data set evaluated in this publication demonstrated that the relation between negative and positive controls, although different EMS concentrations had been used, seemed to be satisfactorily distinct for all laboratories with respect to the ratio, difference and quantile analyses of all related control groups. The statistical summarization of cell-level data to a single animal value increases reliability of results by partial fulfilment of the statistical model assumptions (e.g., log-transformation). However, summarization always results in a loss of information. This confirms the importance of detailed study protocols and reports to monitor study performance and robustness.

In the variance component analysis, comparison of inter- and intra-study variability showed no clear tendency across the five laboratories, out of which 3 had too few historical studies for reliable conclusions. To properly capture such uncertainties, it is recommended to additionally calculate CIs for inter- and intra-study ratios, if expertise is available.

The present results demonstrate that some statistical questions regarding *in vivo* comet assay data are still open for future analyses and discussions, such as the optimal level of summarization of data in the analysis to allow biologically relevant test interpretation. Here we present what we believe to be an optimal trade-off between statistical fit and simplicity of understanding. For future investigations it would be very interesting to apply the statistical strategy identified in this work to further data sets, to confirm applicability of the identified recommendations.

Funding body information

The German Society for Environmental Mutation Research (GUM e. V.) supported the work by making a contribution to the open access fees.

This work was also supported by the Research Training Group “Biostatistical Methods for High-Dimensional Data in Toxicology” (RTG 2624, Project I1 and P2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation—Project Number 427806116).

CRediT authorship contribution statement

Timur Tug: Conceptualization, Data curation, Formal analysis, Investigation, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Julia C. Duda:** Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Max Menssen:** Formal analysis, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Shannon Wilson Bruce:** Resources, Writing – review

& editing. **Frank Bringezu:** Resources, Supervision, Writing – review & editing. **Martina Dammann:** Data curation, Writing – review & editing. **Roland Frötschl:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Volker Harm:** Data curation, Formal analysis. **Katja Ickstadt:** Supervision, Writing – review & editing. **Bernd-Wolfgang Igl:** Formal analysis, Supervision, Writing – original draft, Writing – review & editing. **Marco Jarzombek:** Resources, Writing – review & editing. **Rupert Kellner:** Supervision, Writing – review & editing. **Jasmin Lott:** Conceptualization, Supervision. **Stefan Pfuhler:** Resources, Supervision, Writing – review & editing. **Ulla Plappert-Helbig:** Supervision, Writing – original draft, Writing – review & editing. **Jörg Rahnenführer:** Supervision, Writing – original draft, Writing – review & editing. **Markus Schulz:** Supervision, Writing – original draft, Writing – review & editing. **Lea Vaas:** Investigation, Supervision, Writing – review & editing. **Marie Vasquez:** Writing – review & editing. **Verena Ziegler:** Resources, Supervision, Writing – original draft, Writing – review & editing. **Christina Ziemann:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

The authors are grateful to all companies and laboratories, which provided the historical control data, analysed in the study, and to all scientists and technician who were involved in data generation. We would like to thank Stephen D. Dertinger for very detailed and critical input and comments on this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2024.105583.

References

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bowen, D.E., Whitwell, J.H., Lillford, L., Henderson, D., Kidd, D., Garry, S.M., Pearce, G., Beevers, C., Kirkland, D.J., 2011. Evaluation of a multi-endpoint assay in rats, combining the bone- marrow micronucleus test, the Comet assay and the flow-cytometric peripheral blood micronucleus test. Mutat. Res. Genet. Toxicol. Environ. Mutagen 722 (1), 7–19. https://doi.org/10.1016/j.mrgentox.2011.02.009.

Bright, J., Aylott, M., Bate, S., Geys, H., Jarvis, P., Saul, J., Vonk, R., 2011. Recommendations on the statistical analysis of the Comet assay. Pharmaceut. Stat. 10 (6), 485–493. https://doi.org/10.1002/pst.530.

Brown, V.A., 2021. An introduction to linear mixed-effects modeling in R. Adv. Methods and Pract. Psycholog. Sci. 4 (1) https://doi.org/10.1177/2515245920960351, 251524592096035–251524592096035.

Burlinson, B., Tice, R.R., Speit, G., Agurell, E., Brendler-Schwaab, S.Y., Collins, A.R., Escobar, P., Honma, M., Kumaravel, T.S., Nakajima, M., Sasaki, Y.F., Thybaud, V., Uno, Y., Vasquez, M., Hartmann, A., 2007. Fourth international workgroup on genotoxicity testing: results of the in vivo comet assay workgroup. Mutat. Res. Genet. Toxicol. Environ. Mutagen 627 (1), 31–35. https://doi.org/10.1016/j.mrgentox.2006.08.011.

Chemicals Act as Amended in the Notice of 28 August 2013 (German Federal Law Gazette (FLG) I P. 3498,3991), last revised by Article 1 of the Regulation of 20 June 2014 (FLG I p. 824).

Collins, A.R., Yamani, N.E., Lorenzo, Y., Shaposhnikov, S., Brunborg, G., Azqueta, A., 2014. Controlling variation in the comet assay. Front. Genet. 5 https://doi.org/10.3389/Fgene.2014.00359.

Dertinger, S.D., Li, D., Beevers, C., Douglas, G.R., Heflich, R.H., Lovell, D.P., Roberts, D. J., Smith, R., Uno, Y., Williams, A., Witt, K.L., Zeller, A., Zhou, C., 2023. Assessing the Quality and Making Appropriate Use of Historical Negative Control Data: A Report of the International Workshop on Genotoxicity Testing (IWGT). Environmental and Molecular Mutagenesis. https://doi.org/10.1002/em.22541.

EFSA Scientific Committee, 2011. Scientific Opinion on genotoxicity testing strategies applicable to food and feed safety assessment. EFSA J. 9 (9), 2379. https://doi.org/10.2903/j.efsa.2011.2379.

Ersson, C., Moller, P., Forchhammer, L., Loft, S., Azqueta, A., Godschalk, R.W.L., van Schooten, F.-J., Jones, G.D.D., Higgins, J.A., Cooke, M.S., Mistry, V., Karbaschi, M., Phillips, D.H., Sozeri, O., Routledge, M.N., Nelson-Smith, K., Riso, P., Porrini, M., Matullo, G., Allione, A., Stępnik, M., Ferlińska, M., Teixeira, J.P., Costa, S., Corcuera, L.-A., de Cerain, A.L., Laffon, B., Valdiglesias, V., Collins, A.R., Moller, L., 2013. An ECVAG inter-laboratory validation study of the comet assay: inter-laboratory and intra-laboratory variations of DNA strand breaks and FPG-sensitive sites in human mononuclear cells. Mutagenesis 28 (3), 279–286. https://doi.org/10.1093/mutage/get001.

European Chemicals Agency, 2017. Guidance on Information Requirements and Chemical Safety Assessment : Chapter R.7a : Endpoint Specific Guidance. European Chemicals Agency. https://doi.org/10.2823/337352.

Ghebretinsae, A.H., Faes, C., Molenberghs, G., De Boeck, M., Geys, H., 2013. A Bayesian, generalized frailty model for comet assays. J. Biopharm. Stat. 23 (3), 618–636. https://doi.org/10.1080/10543406.2012.756499.

Grey, K., 2018. _ggQC: Quality Control Charts for ’ggplot’_. R Package version 0.0.31. https://CRAN.R-project.org/package=ggQC.

Gurka, M.J., 2006. Selecting the best linear mixed model under REML. Am. Statistician 60 (1), 19–26. https://doi.org/10.1198/000313006X90396.

Hartmann, A., 2003. Recommendations for conducting the in vivo alkaline Comet assay. Mutagenesis 18 (1), 45–51. https://doi.org/10.1093/mutage/18.1.45.

Heumann, C., Schomaker, M., Shalabh, 2016. Introduction to Statistics and Data Analysis. Springer International Publishing. https://doi.org/10.1007/978-3-319-46162-5.

Igl, B.W., Bitsch, A., Bringezu, F., Chang, S., Dammann, M., Frötschl, R., Harm, V., Kellner, R., Krzykalla, V., Lott, J., Nern, M., Pfuhler, S., Queisser, N., Schulz, M., Sutter, A., Vaas, L., Vonk, R., Zellner, D., Ziemann, C., 2019. The rat bone marrow micronucleus test: statistical considerations on historical negative control data. Regul. Toxicol. Pharmacol. : RTP (Regul. Toxicol. Pharmacol.) 102, 13–22. https://doi.org/10.1016/j.yrtph.2018.12.009.

Kirkland, D., Speit, G., 2008. Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. Mutat. Res. Genet. Toxicol. Environ. Mutagen 654 (2), 114–132. https://doi.org/10.1016/j.mrgentox.2008.05.002.

Kluxen, F.M., Weber, K., Strupp, C., Jensen, S.M., Hothorn, L.A., Garcin, J.-C., Hofmann, T., 2021. Using historical control data in bioassays for regulatory toxicology. Regul. Toxicol. Pharmacol. 125, 105024 https://doi.org/10.1016/j.yrtph.2021.105024.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Software 82 (13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Lovell, D.P., Omori, T., 2008. Statistical issues in the use of the comet assay. Mutagenesis 23 (3), 171–182. https://doi.org/10.1093/mutage/gen015.

Lovell, D.P., Thomas, G., Dubow, R., 1999. Issues related to the experimental design and subsequent statistical analysis of in vivo and in vitro comet studies. Teratog. Carcinog. Mutagen. 19 (2), 109–119. https://doi.org/10.1002/(SICI)1520-6866(1999)19:2%3C109::AID-TCM4%3E3.0.CO;2-5.

Lovell, D.P., Fellows, M., Saul, J., Whitwell, J., Custer, L., Dertinger, S., Escobar, P., Fiedler, R., Hemmann, U., Kenny, J., Smith, R., van der Leede, B.M., Zeller, A., 2020. Analysis of historical negative control group data from the rat in vivo micronucleus assay. Mutat. Res. Genet. Toxicol. Environ. Mutagen 849, 503086. https://doi.org/10.1016/j.mrgentox.2019.503086.

Menssen, M., 2023. The calculation of historical control limits in toxicology: do’s, don’ts and open issues from a statistical perspective. Mutat. Res. Genet. Toxicol. Environ. Mutagen 892. https://doi.org/10.1016/j.mrgentox.2023.503695.

Menssen, M., Schaarschmidt, F., 2022. Prediction intervals for all of M future observations based on linear random effects models. Stat. Neerl. 76 (3), 283–308. https://doi.org/10.1111/stan.12260.

Møller, P., Azqueta, A., Boutet-Robinet, E., Koppen, G., Bonassi, S., Milić, M., Gajski, G., Costa, S., Teixeira, J.P., Pereira, C.C., Dusinska, M., Godschalk, R., Brunborg, G., Gutzkow, K.B., Giovannelli, L., Cooke, M.S., Richling, E., Laffon, B., Valdiglesias, V., Basaran, N., Del Bo, C., Zegura, B., Novak, M., Stopper, H., Vodicka, P., Vodenkova, S., de Andrade, V.M., Sramkova, M., Gabelova, A., Collins, A.R., Langie, S.A.S., 2020. Minimum Information for Reporting on the Comet Assay (MIRCA): recommendations for describing comet assay procedures and results. Nat. Protoc. 15 (12), 3817–3826. https://doi.org/10.1038/s41596-020-0398-1.

Moral, R.A., Hinde, J., Demétrio, C.G.B., 2017. Half-normal plots and overdispersed models in R: the hnp package. J. Stat. Software 81 (10), 1–23. https://doi.org/10.18637/jss.v081.i10.

Muruzabal, D., Collins, A., Azqueta, A., 2021. The enzyme-modified comet assay: past, present and future. Food Chem. Toxicol. 147, 111865 https://doi.org/10.1016/j.fct.2020.111865.

OECD, 2016. Test No. 489. In: In Vivo Mammalian Alkaline Comet Assay. OECD. https://doi.org/10.1787/9789264264885-en.

Ostling, O., Johanson, K.J., 1984. Microelectrophoretic study of radiation-induced DNA damages in individual mammalian cells. Biochem. Biophys. Res. Commun. 123 (1), 291–298. https://doi.org/10.1016/0006-291X(84)90411-X.

Plappert-Helbig, U., Guérard, M., 2015. Inter-laboratory comparison of the in vivo comet assay including three image analysis systems. Environ. Mol. Mutagen. 56 (9), 788–793. https://doi.org/10.1002/em.21964.

Recio, L., Hobbs, C., Caspary, W., Witt, K.L., 2010. Dose-response assessment of four genotoxic chemicals in a combined mouse and rat micronucleus (MN) and Comet assay protocol. J. Toxicol. Sci. 35 (2), 149–162. https://doi.org/10.2131/jts.35.149.

Rothfuss, A., O'Donovan, M., Boeck, M.D., Brault, D., Czich, A., Custer, L., Hamada, S., Plappert-Helbig, U., Hayashi, M., Howe, J., Kraynak, A.R., jan van der, Bas-Leede, Nakajima, M., Priestley, C., Thybaud, V., Saigo, K., Sawant, S., Shi, J., Storer, R., Struwe, M., Vock, E., Galloway, S., 2010. Collaborative study on fifteen compounds in the rat-liver Comet assay integrated into 2- and 4-week repeat-dose studies. Mutat. Res. Genet. Toxicol. Environ. Mutagen 702 (1), 40–69. https://doi.org/10.1016/j.mrgentox.2010.07.006.

Sasaki, M., Dakeishi, M., Hoshi, S., Ishii, N., Murata, K., 2008. Assessment of DNA damage in Japanese nurses handling antineoplastic drugs by the comet assay. J. Occup. Health 50 (1), 7–12. https://doi.org/10.1539/joh.50.7.

Schmidt, C.O., Struckmann, S., Enzenbach, C., et al., 2021. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med. Res. Methodol. 21, 63. https://doi.org/10.1186/s12874-021-01252-7.

Searle, S.R., Casella, G., McCulloch, C.E., 2006. Variance Components. Wiley. https://doi.org/10.1002/9780470316856.

Singh, N.P., McCoy, M.T., Tice, R.R., Schneider, E.L., 1988. A simple technique for quantitation of low levels of DNA damage in individual cells. Exp. Cell Res. 175 (1), 184–191. https://doi.org/10.1016/0014-4827(88)90265-0.

Speit, G., Kojima, H., Burlinson, B., Collins, A.R., Kasper, P., Plappert-Helbig, U., Uno, Y., Vasquez, M., Beevers, C., De Boeck, M., Escobar, P.A., Kitamoto, S., Pant, K., Pfuhler, S., Tanaka, J., Levy, D.D., 2015. Critical issues with the in vivo comet assay: a report of the comet assay working group in the 6th International Workshop on Genotoxicity Testing (IWGT). Mutation research. Genetic Toxicol. Environ. Mutag. 783, 6–12. https://doi.org/10.1016/j.mrgentox.2014.09.006.

Team, R.C., 2022. R: A Language and Environment for Statistical Computing. https://www.R-project.org/.

Tice, R.R., Agurell, E., Anderson, D., Burlinson, B., Hartmann, A., Kobayashi, H., Miyamae, Y., Rojas, E., Ryu, J.-C., Sasaki, Y.F., 2000. Single cell gel/comet assay: guidelines for in vitro and in vivo genetic toxicology testing. Environ. Mol. Mutagen. 35 (3), 206–221. https://doi.org/10.1002/(SICI)1098-2280(2000)35:3<206::AID-EM8>3.0.CO;2-J.

Tug, T., Ickstadt, K., Kunz, M., Sutter, A., Igl, B.-W., 2020. Statistical analysis of in vivo alkaline comet assay data - comparison of median and geometric mean as centrality measures. Regul. Toxicol. Pharmacol. 118, 104808 https://doi.org/10.1016/j.yrtph.2020.104808.

Uno, Y., Kojima, H., Omori, T., Corvi, R., Honma, M., Schechtman, L.M., Tice, R.R., Burlinson, B., Escobar, P.A., Kraynak, A.R., Nakagawa, Y., Nakajima, M., Pant, K., Asano, N., Lovell, D., Morita, T., Ohno, Y., Hayashi, M., 2015. JaCVAM-organized international validation study of the in vivo rodent alkaline comet assay for the detection of genotoxic carcinogens: I. Summary of pre-validation study results. Mutat. Res. Genet. Toxicol. Environ. Mutagen 786, 3–13. https://doi.org/10.1016/j.mrgentox.2015.04.011. –788.

van der Leede, B.J., Doherty, A., Guérard, M., Howe, J., O'Donovan, M., Plappert-Helbig, U., Thybaud, V., 2014. Performance and data interpretation of the in vivo comet assay in pharmaceutical industry: EFPIA survey results. Mutation research. Genetic Toxicol. Environ. Mutagenesis 775–776, 81–88. https://doi.org/10.1016/j.mrgentox.2014.09.008.

Vasquez, M.Z., 2010. Combining the in vivo comet and micronucleus assays: a practical approach to genotoxicity testing and data interpretation. Mutagenesis 25 (2), 187–199. https://doi.org/10.1093/mutage/gep060.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.

Wiklund, S.J., Agurell, E., 2003. Aspects of design and statistical analysis in the Comet assay. Mutagenesis 18 (2), 167–175. https://doi.org/10.1093/mutage/18.2.167.