technische universität
dortmund

fakultät
statistik

# "Integrative Statistical Methods for Analyzing Biomedical Data: Applications in Health and Disease"

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät Statistik der Technischen Universität Dortmund


Vorgelegt von

**Timur Tuğ**


Dortmund, April 2025

**Amtierender Dekan:**
Prof. Dr. Philipp Doebler

**Gutachter:**

Prof. Dr. Katja Ickstadt (Technische Universität Dortmund)

Prof. Dr. Jörg Rahnenführer (Technische Universität Dortmund)

Prof. Dr. Anke Hüls (Emory University Atlanta)

**Tag der Prüfung:**
14. Mai 2025

# *Abstract*

In a series of four complementary studies, we apply innovative integrative statistical methods to diverse biomedical datasets to address both fundamental research questions and practical challenges in health and disease. Two of these investigations focus on the *in vivo* alkaline comet assay - a pivotal tool for assessing DNA damage as a marker of genotoxicity. In the first comet assay study (Article 1), we examine the impact of different centrality measures on the evaluation of tail intensity data. Using both original experimental data and simulation frameworks, we demonstrate that even subtle variations in summarizing techniques - whether using medians, arithmetic means, or geometric means - can lead to markedly different statistical conclusions and dose–response interpretations. These findings emphasize the critical need for careful methodological selection in genotoxicity assessments. In a subsequent comet assay work (Article 2), we compile and analyze extensive historical control data from multiple laboratories. This investigation addresses key statistical issues, including inter-laboratory variability and the handling of zero-valued measurements, and discusses whether the findings from the first paper are similar in the centrality statistical measures and regulatory interpretations.

In the third study (Article 3), we introduce a novel multi-omics approach to better understand Alzheimer's disease (AD). By integrating genome-wide DNA methylation profiles with high-resolution metabolomics data from prefrontal cortex tissue samples, we develop innovative single-, joint- and multi-omics profile scores using Machine Learning and advanced regression techniques. This integrative analysis with multi-omics profile scores significantly improves the prediction of AD neuropathology compared to single-omics profile scores derived from DNA methylation or metabolomics alone. It also uncovers pivotal biological pathways, such as lipid metabolism and signal transduction, that are potentially involved in AD neuropathology. These findings underscore the potential of combining multiple omics layers to elucidate complex molecular interactions underlying neurodegenerative disorders.

Our fourth investigation (Article 4) applies hierarchical modeling to veterinary epidemiology, specifically targeting respiratory diseases in piglet production. We thereby compare frequentist and Bayesian hierarchical regression models to assess the influence of various environmental and management factors - including floor condition, water flow rates, stocking density, and indoor climate conditions - on respiratory health outcomes in pigs. By accounting for the multi-level structure inherent in farm data (spanning individual animals, pens, compartments, and farms), we demonstrate that Bayesian approaches with informative priors can effectively

overcome challenges posed by small sample sizes and high inter-cluster variability. This ultimately provides more robust estimates and practical insights for disease management in livestock production.

Collectively, the four projects of my cumulative thesis illustrate how tailored, integrative statistical methodologies can enhance our understanding of complex biological systems. These methodologies enhance analytical precision and facilitate robust conclusions across diverse fields of application, including the regulatory evaluation of chemical safety, the elucidation of neurodegenerative disease mechanisms, and the optimization of animal health in agricultural contexts. The work emphasizes that the choice of statistical methods is not merely a technical detail but a pivotal factor that can substantially alter study outcomes and subsequent interpretations in both clinical and applied research environments. While the first two manuscripts are published, the third and fourth work are submitted and attached in its current version.

# Acknowledgments

Firstly, I would like to express my gratitude to all those who have provided me with support during the recent years, which have been marked by both positive and negative periods. If I were to enumerate everyone here individually, this post would be twice as long. In lieu of a comprehensive enumeration, I extend my profound gratitude to all those who have contributed to this journey.

I would like to extend my deepest gratitude to my supervisor, Katja, for the profoundly enriching experiences we shared in our professional endeavours and for the numerous sporting excursions we embarked on together. Her consistent support and encouragement have been instrumental in my professional growth and development.

Additionally, I would like to express my profound gratitude to Jörg, the speaker of the Research Training Group 2624 "Statistical Methods for High-Dimensional Data in Toxicology," who made it possible for me to delve into the vast realm of statistics in the field of toxicology. Additionally, I would like to express my sincere gratitude for the unwavering support and positive mindset he has consistently demonstrated throughout my university years.

I would also like to express my deepest gratitude to Anke for her invaluable support, guidance, and encouragement throughout my scientific journey. Her mentorship has been instrumental in broadening my perspective and fostering my growth in the field. I am particularly grateful for the opportunity she has given me to work on our joint project at Emory University in Atlanta, Georgia, United States, where we have had the chance to spend more time together.

I would also like to express my profound gratitude to Bernd, whose guidance and support in the realm of non-clinical statistics was instrumental in my transition into the scientific realm and the completion of this dissertation.

During my final year at the university, I had the opportunity to become acquainted with Markus. I am profoundly grateful to him for his invaluable advice, insightful concepts, and meticulous feedback on this work.

Finally, I would like to express my gratitude to Guido for providing me with the opportunity to engage in scientific research as a student assistant. His support has been instrumental in broadening my intellectual horizons and introducing me to novel concepts and employment prospects.

Last but not least, I want to say a heartfelt thank you to my family and friends who have been there for me every step of the way.

Your love and support have been the wind beneath my wings, and I can't thank you enough. Your encouragement and unwavering love helped me persevere through the tough times.

*"Do… or do not. There is no try."*

– Yoda, Star Wars.

# *List of Publications*

This cumulative thesis is based on the following four manuscripts:

Article 1: 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. Regulatory Toxicology and Pharmacology, 104808. https://doi.org/10.1016/j.yrtph.2020.104808

Contribution of the author:

The author of this thesis significantly contributed to this research paper by leading the statistical analysis and its methodological development. His contributions included:

- **Designing and conducting data analyses**, focusing on the impact of different centrality measures (median vs. geometric mean) in the interpretation of *in vivo* alkaline comet assay results.
- **Developing and applying statistical models**, including linear models and trend tests, to assess the effects of different analytical approaches.
- **Performing simulation studies** to evaluate statistical power under various experimental conditions, such as varying the number of animals, slides, and cells per slide.
- **Interpreting the results**, demonstrating that the median is a more robust centrality measure than the geometric mean in genotoxicity studies.
- **Writing the first draft of the manuscript**, structuring the findings, and presenting the methodological and experimental insights.
- **Engaging in scientific discussions and revisions** in collaboration with Bernd-Wolfgang Igl and under the supervision of Katja Ickstadt.

His contributions were **critical to the methodological depth and scientific impact** of the study.

Article 2: Tug, T., Duda, J. C., Menssen, M., Bruce, S. W., Bringezu, F., Dammann, M., Frötschl, R., Harm, V., Ickstadt, K., Igl, B. W., Jarzombek, M., Kellner, R., Lott, J., Pfuhler, S., Plappert-Helbig, U., Rahnenführer, J., Schulz, M., Vaas, L., Vasquez, M., Ziegler, V., Ziemann, C. (2024). *In vivo* alkaline comet assay: Statistical considerations on historical negative and positive control data. Regulatory Toxicology and Pharmacology, 105583. https://doi.org/10.1016/j.yrtph.2024.105583

Contribution of the author:

Timur Tug played a fundamental role in this study, expanding upon previous research by applying and refining statistical methodologies on real-world data. His key contributions included:

- **Designing and conducting data analyses**, focusing on statistical considerations for historical negative and positive control data in the *in vivo* alkaline comet assay.
- **Developing and applying statistical models**, including variance components analysis and mixed-effects models, to assess sources of variability across laboratories and studies.
- **Performing comprehensive statistical evaluations**, including comparisons of different summarizing strategies (median, arithmetic mean, geometric mean), handling of zero values, and the influence of data distribution on statistical conclusions.
- **Interpreting the results**, demonstrating how inter-laboratory variability, choice of statistical methods, and summarization techniques affect the assessment of DNA damage and the validity of historical control data.
- **Writing the first draft of the manuscript**, structuring the findings, and presenting statistical recommendations for improving the robustness of comet assay evaluations.
- **Engaging in scientific discussions and revisions** in collaboration with Julia C. Duda and under the supervision of Christina Ziemann and Bernd-Wolfgang Igl.

Through his work, he significantly enhanced the statistical foundation of the study, providing crucial insights into the handling and interpretation of historical control data, ultimately contributing to more reliable and standardized approaches in genotoxicity assessments.

Article 3: Tug, T., Liang, D., Teschke, S., Tan, Y., Gearing, M., Levey, A. I., Lah, J. J., Wingo, A. P., Wingo, T. S., Lau, M., Ickstadt, K., Hüls, A. (2025). Development and application of brain tissue-based multi-omics profile scores for Alzheimer's disease.

Under revision (09.04.25) in Alzheimer's & Dementia®: The Journal of the Alzheimer's Association (The manuscript is attached in its current version).

Contribution of the author:

The author of the current thesis played a crucial role in this study by leading the development and application of brain tissue-based multi-omics profile scores for Alzheimer's disease. His contributions included:

- **Designing and conducting data analyses,** integrating genome-wide DNA methylation and high-resolution metabolomics data from 157 frontal cortex samples to improve the understanding of Alzheimer's pathology.
- **Developing novel statistical models,** utilizing Machine Learning techniques such as Random Forest, Elastic Net, Boosting, and ordinal logistic regression to construct and evaluate single- and multi-omics profile scores (*PS*).
- **Performing pathway analysis, identifying lipid metabolism and signal transduction** as key biological pathways associated with Alzheimer's disease by analysing enriched CpG sites and metabolomic features.
- **Interpreting the results,** demonstrating that combining DNA methylation and metabolomics enhances predictive accuracy for AD pathology beyond single-omics approaches.
- **Writing the first draft of the manuscript,** structuring findings and presenting statistical methodologies, results, and implications for future research.
- **Engaging in scientific discussions and revisions** in collaboration with multiple researchers from Emory University and the Research Training Group 2624 at TU Dortmund University, under the supervision of Anke Hüls.

Through his contributions, he significantly advanced the methodological framework for multi-omics analyses in Alzheimer's disease research, providing valuable insights into epigenetic and metabolic interactions that may inform future therapeutic strategies.

Article 4: Tug, T., Mers, F., Schäkel, F., Höltig, D., Kreienbrock, L., & Ickstadt, K. (2025) Hierarchical modelling of risk factors with and without prior information – regression model evaluation for respiratory diseases in piglet production from daily practice data.

Submitted (The manuscript is attached in its current version).

Contribution of the author:

The author played a central role in this study by designing, implementing, and evaluating hierarchical regression models to identify risk factors for respiratory diseases in piglet production using real-world data from veterinary practice. His contributions included:

- **Leading the data preparation process**, including structuring complex hierarchical data from 30 piglet-producing farms encompassing 450 animals across multiple nested levels (farm, compartment, pen, individual).
- **Developing and comparing multiple statistical modelling approaches**, including hierarchical frequentist and Bayesian logistic regression models with informative and non-informative priors. This and the following ideas and analyses were jointly developed and carried out together with Fiona Mers.
- **Utilizing advanced model selection techniques** based on AIC/AICC, marginal and conditional $R^2$, intraclass correlation (ICC), and Bayesian model comparison metrics such as elpd and Bayes $R^2$ to assess predictive accuracy and robustness.
- **Implementing multiple imputation strategies** using predictive mean matching to handle missing data across categorical and continuous variables while respecting the hierarchical structure of the dataset.
- **Identifying key environmental and management-related risk factors** -including floor condition, stocking density, and water flow rate - as significant covariates of piglet cough prevalence.
- **Writing the first draft of the manuscript**, interpreting model outcomes in a veterinary epidemiological context, and providing statistical insights to guide preventive measures in livestock health.
- **Collaborating closely with veterinary and statistical researchers** at the University of Veterinary Medicine Hannover and TU Dortmund University under the supervision of Prof. Katja Ickstadt and Prof. Lothar Kreienbrock.

Through his contributions, the author significantly advanced the statistical modelling framework for veterinary epidemiological studies, especially in the context of small-sample, complex data structures, demonstrating the utility of hierarchical Bayesian methods in improving animal health surveillance and decision-making in agricultural practice.

**Further publications during my PhD:**

1. Hoffman, S.-S., Lane, A.-N., Gaskins, A.-J., Ebelt, S., **Tug, T.,** Tran, V., Jones, D.-P., Liang, D., Hüls, A. (2024). Development of a metabolomic risk score for exposure to traffic-related air pollution: A multi-cohort study. Environmental Research, 120172. https://doi.org/10.1016/j.envres.2024.120172

Contribution of the author:
Timur Tug assisted the first author with statistical analysis and programming. The metabolic profiles of individuals with varying levels of exposure to TRAP were statistically analyzed to identify specific metabolites associated with exposure. Identified metabolites were then weighted and integrated into a single risk score (metRS) that quantifies individual exposure.

# *Contents*

# Part I

# Introduction

# 1 Motivation

In biomedical research, particularly in toxicology, epidemiology, and public health, the collaboration between domain experts and statisticians is indispensable. While statistical models and computational tools are central to the analysis and interpretation of complex datasets, their utility depends critically on how well they are aligned with the scientific questions and experimental designs defined by subject-matter experts. As such, the relationship between applied scientists and statisticians must be seen as a continuous dialogue rather than a one-time transaction [1], [2], [3].

Toxicologists, for example, are concerned with understanding the biological effects of chemical exposures and assessing risk thresholds for adverse health outcomes. Their experimental designs often involve repeated measures, dose-response relationships, historical controls, and biological interpretation. Similarly, clinicians deal with population-level data, observational designs, and confounding structures shaped by social, environmental, and genetic factors. While these disciplines are data-rich and problem-driven, the correct statistical approach is rarely straightforward. This underscores the need for joint planning, transparent interpretation, and model evaluation between biomedical researchers and statisticians [4]. The success of such interdisciplinary collaborations depends on mutual understanding and trust. Statisticians bring rigor in study design, model building, and uncertainty quantification, but must also be sensitive to the specific hypotheses, data limitations, and decision-making contexts in the respective field. Conversely, domain experts must articulate their research questions clearly and be willing to engage with abstract statistical reasoning. Misalignment often arises when statistical models are applied without an understanding of biological constraints, or when statistical results are misinterpreted due to a lack of methodological knowledge [5], [6].

In the age of data-intensive sciences, interdisciplinary collaboration is not only valuable, but essential. The integration of expertise from different disciplines enables researchers to address complex biomedical questions more effectively, while also improving the reproducibility, transparency, and interpretability of results. Particularly in fields like toxicology, epidemiology, and systems biology, the complexity of data requires joint efforts that go beyond traditional disciplinary boundaries. One illustrative example of this principle is the RIPOSTE framework, which was developed to promote early and structured dialogue between laboratory scientists and statisticians. It emphasizes the importance of clarifying

specific hypotheses, aligning experimental goals with statistical strategies, and ensuring that methodological aspects - such as sample size determination, control group selection, and analysis plans - are defined collaboratively from the outset. The framework serves as a structured prompt to ensure that all critical components of a study are addressed before data collection begins, thereby improving the design, conduct, and analysis of laboratory experiments and ultimately enhancing reproducibility [7]. A second compelling example comes from the field of multi-omics research, where the integration of genomics, transcriptomics, epigenomics, proteomics, and metabolomics data has created powerful new opportunities to unravel disease mechanisms. The success of such integration critically depends on interdisciplinary collaboration [8]. Statisticians are responsible for selecting appropriate analytical frameworks, ensuring robustness in model building, and managing the high-dimensional nature of omics data. At the same time, clinicians and life scientists provide essential biological insight to ensure that the models are grounded in realistic physiological assumptions. Without coordinated input from both sides, studies may suffer from low statistical power, irrelevant or misinterpreted biomarkers, or fundamentally flawed conclusions. Therefore, co-designing experiments and maintaining a continuous exchange throughout the entire research process is indispensable for translating complex data into meaningful biomedical knowledge.

In the era of high-dimensional biomedical research, the integration of heterogeneous data types has become a cornerstone for advancing our understanding of human health and disease. The explosion of high-throughput technologies - ranging from next-generation sequencing and microarrays to mass spectrometry and advanced imaging - has enabled the generation of vast datasets across multiple biological domains. These include genomics, transcriptomics, epigenomics, proteomics, metabolomics, and clinical data, each offering a distinct perspective on the molecular and cellular landscape of biological systems. However, analyzing these data sources in isolation fails to capture the complexity and interdependence of biological processes. As such, integrative statistical methods are essential for combining information across multiple data layers to derive more comprehensive, biologically meaningful, and clinically actionable insights [8], [9], [10], [11], [12]. Integrative approaches offer several advantages: they enhance statistical power by borrowing strength across datasets, improve the robustness of findings, and enable the discovery of cross-modal biomarkers and disease subtypes that remain hidden in single-omics analyses [8], [9]. By leveraging complementary information from different biological layers - such as genomics,

transcriptomics, epigenomics, and metabolomics - researchers can gain a more holistic and biologically meaningful understanding of disease mechanisms [12]. Additionally, methods such as similarity network fusion [10] and regularized clustering frameworks [11] have demonstrated how computational strategies can effectively integrate heterogeneous data while preserving biological structure.

In this cumulative thesis, we apply integrative and structured statistical methods to a diverse set of research questions in biomedical and veterinary sciences. Our goal is to demonstrate how tailored statistical modeling - based on both data characteristics and subject-matter knowledge - can enhance the reliability, interpretability, and clinical relevance of research findings across different domains. The integration of prior biological knowledge and structured statistical models has significantly advanced the precision and interpretability of disease modeling. These principles are central to this work, where we use integrative multi-omics modeling to construct tissue-specific profile scores for neurodegenerative disease and apply hierarchical models to improve inference in veterinary epidemiological settings. Across all four studies presented here, we illustrate how statistical frameworks that combine methodological rigor with biological insight can yield clinically meaningful results from complex, high-dimensional data.

In the first article, we investigate the effect of statistical summarization strategies in the *in vivo* alkaline comet assay, a widely used method in genotoxicology for detecting DNA strand breaks. The data originates from five experimental studies conducted at Bayer AG, each with 30 animals across five treatment groups (6 animals per group) and 50 cells per slide. We compare the use of the median versus the geometric mean as per-slide summary measures for the primary endpoint "tail intensity" and assess their impact on statistical inference. This work directly addresses a key challenge in genotoxicology: skewed distributions and sensitivity to extreme values, which can bias standard statistical estimates. The median offers a robust measure of central tendency by resisting the influence of extreme values, whereas the geometric mean, though not inherently robust, down-weights large positive outliers and is therefore more appropriate than the arithmetic mean for right-skewed distributions. Both are essential to mitigate these issues and ensure reliable conclusions [13], [14]. We demonstrate that the choice of summary measures can substantially alter the detection of genotoxic effects and even reverse study outcomes - highlighting the need for methodological transparency and justification in statistical analysis. The findings align with recent calls for improving statistical practices in regulatory toxicology, particularly when

working with noisy, cell-based assay data [15], [16].

In the second article, we expand the scope by analyzing historical control data from over 200 *in vivo* alkaline comet assay studies collected from five laboratories. This dataset comprises more than 2,000 rats and includes organ-level data (liver, blood, stomach, duodenum, lung), detailed slide protocols, and measurements at the single-cell level. Using mixed-effects models and variance component analyses, we investigate how laboratory effects, summarization techniques, and data structure affect study outcomes. This study exemplifies the need for integrative statistical approaches in toxicogenomics, where data complexity, batch effects, and biological variability are prominent [15], [16]. The incorporation of historical control data plays a critical role in contextualizing new findings and improving the estimation of background variability, which is increasingly recognized as essential for robust decision-making in both scientific and regulatory contexts [17], [18]. We provide methodological recommendations for harmonizing evaluation practices, demonstrating how structured modeling can strengthen inference in complex toxicological data settings.

In the third article, we shift to the domain of neurodegenerative diseases and apply integrative multi-omics modeling to Alzheimer's disease. The dataset consists of genome-wide DNA methylation (789,286 CpG sites) and high-resolution metabolomics data (35,348 features) from 157 postmortem prefrontal cortex samples. We develop single- and multi-omics profile scores to predict established neuropathological scores (ABC score, Braak stage, CERAD [19], [20], [21], [22], [23], [24]) using Machine Learning and Statistical Learning algorithms. We also perform pathway analysis to identify biological mechanisms. This work embodies the strengths and challenges of modern integrative modeling, which combines diverse layers of omics data to uncover complex biological interactions. Common challenges in multi-omics studies include the heterogeneity in data formats, a high dimensionality, and the presence of missing values [9], [11], [12]. To address these, we implement robust Machine Learning techniques and normalization procedures to construct interpretable and predictive models. Our findings support the use of Machine Learning and Statistical Learning models, in uncovering non-linear, multi-layered biological interactions [25], [26], [27], [28]. Moreover, by integrating molecular and clinical data, we contribute to advancing precision medicine, which aims to tailor disease prediction and therapy based on individual biological profiles [8], [29].

In the fourth article, we analyze data from the VASIB project, which investigates respiratory disease risk in piglet production. The dataset includes 450 animals from 30 farms and spans

four hierarchical levels (farm, compartment, pen, animal). We explore how environmental (e.g., $CO_2$, temperature, floor condition), clinical (e.g., coughing), and management factors interact across levels. Hierarchical frequentist and Bayesian logistic regression models - both with and without informative priors - are used to assess risk structures and model uncertainty. This application illustrates the advantages of hierarchical modeling for epidemiological data with nested structure, limited sample sizes, and correlated covariates. Traditional frequentistic models are often limited in their applicability in such contexts, but Bayesian models offer flexibility and better convergence properties [30]. By integrating prior knowledge and expert assumptions through informative priors, we improve model interpretability and performance. These approaches are crucial when working with observational field data, and they align with broader efforts in veterinary epidemiology to combine structured models with contextual domain expertise [31], [32], [33], [34].

Taken together, these four articles illustrate how integrative and structured statistical methods can be effectively tailored to meet the specific challenges posed by genetic toxicology, neuroscience, and veterinary epidemiology. Across all studies, we combine rigorous statistical methodology with biological and clinical insight, underscoring the value of interdisciplinary collaboration in extracting clinically and scientifically meaningful information from complex biomedical data. In summary, this work contributes to the growing field of integrative biostatistics by developing and applying statistical approaches that bridge heterogeneous sources of biomedical data. Our objective is not only to enhance inference and prediction but also to ensure interpretability, reproducibility, and biological plausibility. By designing statistical frameworks that are informed by biological priors and empirical data structures, we aim to improve the practical utility of biomedical data integration in both research and clinical contexts. Importantly, integrative collaboration goes far beyond study design and data analysis - it plays a pivotal role in science communication, policy development, and regulatory decision-making. Statisticians increasingly serve as translators between data and decision-making, transforming complex analyses into actionable insights for stakeholders such as public health institutions, regulatory authorities, and industry partners. In this sense, collaboration is not merely an academic exercise but a pathway to real-world impact.

Nevertheless, significant challenges remain. Time constraints, differing professional incentives, and disciplinary silos can hinder effective collaboration. Additionally, barriers such as restrictive data-sharing policies, privacy regulations, and proprietary technologies

continue to limit access to high-quality datasets for integrated analyses. Overcoming these challenges will require not only technical solutions but also institutional support for open science, interdisciplinary funding schemes, and collaborative governance models that foster trust and transparency across disciplines. In conclusion, the evolving relationship between statisticians and biomedical researchers underscores the necessity of sustained, interdisciplinary engagement. As the complexity and volume of biomedical data continues to grow, so, too, must our investment in building robust, communicative, and equitable scientific collaborations. The future of biomedical discovery depends not only on better data or more sophisticated models, but on the strength of the partnerships that guide their use.

The rest of this cumulative thesis is structured as follows: In Chapter 2, we describe the statistical background and methods that form the basis of the work presented in the included manuscripts. Chapter 3 summarizes the four original articles that form the core of this dissertation. We highlight the specific research questions, datasets, and modeling strategies used in each study, and show how the results contribute to both methodological innovation and applied knowledge in biomedical research. In Chapter 4, we finally synthesize the findings across all four studies and reflect on their broader implications for biostatistics, regulatory science, and health research. We discuss methodological strengths and limitations, outline directions for future research, and emphasize the importance of interdisciplinary collaboration in solving complex data-driven problems. The four original manuscripts are included in the second part of this thesis.

# 2 Statistical Methods

This chapter provides a broad methodological overview of the established methods that have been utilized or extended in the studies presented in this thesis. To allow intuitive notations, symbols are mostly but not strictly used consistently across topics. Each of the four articles is examined in more detail in its corresponding subchapter (i.e., Article 1 in Subchapter 2.1, Article 2 in 2.2 etc.), with a brief introduction to the current situation, outlining existing gaps in knowledge, current analytical challenges. The statistical methods and analyses used are also briefly explained.

## 2.1 Comparison of Median and Geometric Mean for Right-Skewed *in vivo* Alkaline Comet Assay Data

The first article presents statistical methods for analyzing *in vivo* alkaline comet assay data, focusing on how different slide-level summary measures - particularly the median and geometric mean - affect test outcomes. Using mixed models and simulation studies, we show that the choice of central tendency significantly influences statistical power and the interpretation of genotoxic effects.

The comet assay (single-cell gel electrophoresis) is an established method for detecting DNA strand breaks at the level of individual cells and is widely used in genotoxicity testing. In its alkaline version, the assay detects single and double-strand breaks as well as alkali-labile sites in DNA. According to OECD guideline No. 489, comet assay data is evaluated by calculating the median tail intensity per slide, followed by the arithmetic mean of slide medians per animal. This approach accounts for the typically right-skewed distribution of tail intensity measurements (e.g., % tail DNA), but also discards part of the data since the median relies on a limited subset of observations. For elementary biological concepts on cells, DNA and the comet assay, see [35], [36], [37].

The primary endpoint, tail intensity (TI, also referred to as "% tail DNA"), is summarized by using the median per slide and compute the arithmetic mean of these medians per animal as the final unit of analysis. This procedure is motivated by the typically right-skewed, non-negative distribution of comet assay data [38], [39]. Formally, for each slide $s$ for animal $a$ with measurements $x_{as1}, x_{as2}, \ldots, x_{asn} \in \mathbb{R}_{\geq 0}$, we calculate the median $\tilde{x}_{as}$ as per slide summary measure and then compute the final per-animal summary $\overline{\tilde{x}_a} = \frac{1}{S}\sum_{s=1}^{S} \tilde{x}_{as}$, where

$S$ is the number of slides for animal $a$. However, alternative central tendency measures to the median, such as the geometric mean, are also considered, particularly when data exhibit high skewness. In such cases, the geometric mean can provide a more representative per slide summary of the central tendency. The geometric mean per slide is given by $\hat{g}_{as} = (\prod_{i=1}^{n} x_{asi})^{1/n} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log x_{asi}\right)$, which corresponds to the exponential of the arithmetic mean of log-transformed values. The arithmetic mean of these geometric means per animal, $\overline{\hat{g}_a} = \frac{1}{S}\sum_{s=1}^{S} \hat{g}_{as}$, then is an alternative per-animal summary. Despite its appealing properties for multiplicative or right-skewed data [40]), the use of $\hat{g}_{as}$ can significantly alter hypothesis test results compared to the median $\tilde{x}_{as}$ (or the respective per-animal summaries measures), as demonstrated in this article.

We apply two primary statistical models depending on the level of aggregation: the statistical slide model and the statistical animal model. The purpose of this comparison is to evaluate whether the animal-level approach, which better reflects real-world practices, provides improved insights compared to the slide-level model:

**Statistical Slide Model.** In the statistical slide model, we treat $S$ slide-level summaries $y_{as}$ (either medians or geometric means) per animal $a$ as repeated measures. We fit a one-way linear model:

$$y_{as} = \mu_j + \epsilon_{as}, \qquad \epsilon_{as} \sim \mathcal{N}(0, \sigma^2), \qquad for\ j \in \{vehicle, low, medium, high\}$$

with fixed treatment effect $\mu_j$ and uncorrelated residuals $\epsilon_{as}$.

**Statistical Animal Model.** For the statistical animal model, we compute one summary per animal $a$, either $\overline{\tilde{x}_a}$ or $\overline{\hat{g}_a}$, and use a one-sided t-test with Satterthwaite approximation to infer the group means hypotheses

$$H_0: \mu_v = \mu_d \quad vs. \quad H_1: \mu_v < \mu_d\,,$$

where $\mu_v$ and $\mu_d$ denote the expected tail intensities in the control/vehicle ($v$) and dose group $d \in \{low, medium, high\}$, respectively [41], [42].

To handle non-monotonic dose-response patterns, particularly downturn effects at high doses, we implement the downturn-protected trend test [43], which identifies the best-fitting dose-response shape among a predefined contrast set $C = \{c_1, \ldots, c_K\}$ using a minimum p-

value approach. The contrast set $C$ consists of different hypotheses for example like $c_1$: $\mu_{vg} < \mu_{low} = \mu_{medium} < \mu_{high}$ and $c_2$: $\mu_{vg} < \mu_{low} < \mu_{medium} = \mu_{high}$ with their corresponding contrast coefficients. The contrast hypotheses are structured to detect specific patterns in the treatment effects (e.g., linear, non-linear with downturn effects), and the method controls for the family-wise error rate across multiple comparisons [43], [44], [45], [46]. To assess the effect of different centrality measures and study designs on statistical power, we simulate data under realistic conditions informed by experimental data from five *in vivo* comet assay studies in this article. We consider $A \in \{6,8\}$ animals per group, $S \in \{3,4,5\}$ slides per animal and $n \in \{30,50,70\}$ cells per slide. In the simulations, tail intensities $x_{si}$ are generated from a Gamma distribution, which fitted the real-world data set best:

$$x_{si} \sim \Gamma(\alpha, \beta), \text{with } \mathbb{E}[x] = \frac{\alpha}{\beta}, Var[x] = \frac{\alpha}{\beta^2},$$

parameterized using empirical means and variances from control data (see the Appendix in Article 1). To simulate dose effects, we apply multiplicative factors $f_j \in \{1.3,1.5,1.8\}$ to the mean of the vehicle group to generate low, medium, and high doses, respectively.

For more realistic scenarios we introduce up to 10% "noisy" values per slide: a random proportion $q \sim Unif(0,0.1) \times n$ of values per slide are drawn from a gamma distribution with 5-fold increased variance and a constant shift of 10, mimicking extreme cell-level damage. Each scenario is replicated 10,000 times, and we estimate statistical power as the proportion of rejections of $H_0$ at level α=0.05 (type I error). In addition to the gamma distribution, we also conducted the analyses with the lognormal distribution, which can also provide a good fit to real data [47].

## 2.2 Advanced Statistical Approaches for Analyzing Historical Control Data in *In vivo* Alkaline Comet Assays

Our first article highlighted for the *in vivo* comet assay how the choice of slide-level summary statistics - particularly the median versus the geometric mean - can substantially influence test decisions and statistical power, especially in the presence of skewed or noisy data [48]. In the present study, we validate and extend these findings for the *in vivo* comet assay using a substantially larger real-world dataset of historical control data from multiple laboratories. We compiled the dataset by collecting raw single-cell comet assay data from

over 200 *in vivo* studies conducted across five laboratories, including detailed metadata on study design, tissue processing, and assay conditions. We investigate the statistical properties of *in vivo* alkaline comet assay data derived from a large historical control dataset (HCD). This investigation focuses on aspects jointly selected by toxicologists and statisticians, including distributional characteristics, summary strategies, zero-value handling, variance decomposition, and the relationship between negative and positive control groups. The methods build upon findings from the first article, which emphasized the importance of appropriate aggregation strategies in detecting genotoxic effects. We proceed in four consecutive steps.

First, we stress that empirical distributions of comet assay data at the cellular level typically exhibit strong right-skewness, often accompanied by a high proportion of zero-values. This non-normality poses significant challenges for standard statistical methods, which rely on the assumption of symmetry or normality. To mitigate these issues, we recommend a logarithmic transformation following the addition of a small constant:

$$x^*_{easi} = x_{easi} + \eta\,, \quad with\, \eta = 0.001,$$
$$y_{easi} = \ln x^*_{easi}\,,$$

where $x_{easi}$ denotes the tail intensity value for the $i$-th cell of slide $s$ of animal $a$ in study $e$ ($y_{easi}$ the logarithmic value of the TI). The shift ensures that all values are positive and stabilizes variance while preserving the scale of the original data, consistent with OECD TG 489 recommendations [18]. The subsequent logarithmic transformation typically leads to an approximately normal distribution.

Second, we critically analyze data aggregation strategies, closely aligned with the simulation design of Article 1. Comet assay data follow a three-level hierarchical structure: cells nested within slides, which are nested within animals. According to OECD TG 489, the median per slide $\tilde{x}_{eas}$ is calculated first, followed by the arithmetic mean per animal $\bar{x}_{ea} = \frac{1}{S}\sum_{s=1}^{S} \tilde{x}_{eas}$, where S denotes the number of slides. In addition to the median, we also evaluate four alternative summary measures: Arithmetic mean, geometric mean, trimmed arithmetic mean, and trimmed geometric mean. For the trimmed means, the largest and smallest 5% of values are excluded prior to averaging. As with the median, we compute the per-animal summaries by first summarizing the data per slide and then averaging across all slides per animal. We calculate summarized tail intensities of each organ data per animal and

laboratory and compare this five slide summary measurements for the vehicle and positive control groups.

Third, for the OECD Test Guideline 489 summary, we perform a variance components analysis to quantify the proportion of variability attributed to study-level differences. To be concrete, we model the shifted and log-transformed per animal-level summarized TI values (i.e., the averaged medians) using a linear random-effects model

$$y_{ea} = \mu + u_e + \varepsilon_{ea}, \qquad u_e \sim \mathcal{N}(0, \sigma_u^2), \qquad \varepsilon_{ea} \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

where $y_{ea}$ is the observed per-animal summary from step 1 and 2 (here, only for the arithmetic mean of the slide medians) for animal $a$ in study $e$, $\mu$ is the overall mean, $u_a$ is the random effect for study $e$, and $\varepsilon_{ea}$ is the residual error term.

The total variance is then decomposed as

$$\mathrm{Var}(y_{ea}) = \sigma_\mathrm{u}^2 + \sigma_\varepsilon^2.$$

This enables us to evaluate the relative contribution of between-study ($\sigma_u^2$) and within-study variability ($\sigma_\varepsilon^2$) to the total variance structure [39].

Fourth and lastly, we quantitatively assess the separation between negative (NC) and positive control (PC) groups, which is crucial for validating assay sensitivity. We compute the difference ($\Delta$) and ratio (R) between control groups:

$$\Delta = \overline{y_{PC}} - \overline{y_{NC}}, \qquad R = \frac{\overline{y_{NC}}}{\overline{y_{PC}}},$$

where $\overline{y_{PC}}$ and $\overline{y_{NC}}$ are the animal arithmetic mean above the slide medians of the positive and negative controls, respectively. This is in line with the guideline OECD TG 489 [18]. We apply these statistical methods to analyze key toxicological hypotheses and thus answer research questions about the comet assay.

## 2.3 Development of Single-, Joint-, and Multi-Omics Profile Scores for Quantifying Molecular Associations with Alzheimer's Disease Neuropathology

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by the accumulation of amyloid-β plaques and neurofibrillary tangles in the brain. Despite extensive research, no curative treatments are currently available, and understanding the multifactorial molecular underpinnings of AD remains a major challenge [49]. Recent advances in high-throughput omics technologies, including epigenomics and metabolomics, provide new opportunities to explore the biological mechanisms driving disease onset and progression [50], [51]. DNA methylation (DNAm), a key epigenetic mechanism, has been linked to inflammation, aging, oxidative stress, and synaptic dysfunction - all of which are implicated in AD pathophysiology [52]. In parallel, metabolomics captures small-molecule profiles that reflect alterations in cellular pathways, including lipid metabolism, amino acid turnover, and mitochondrial function, which are commonly disrupted in AD [53], [54], [55]. While prior studies focus on single-omics approaches, we integrate DNAm and metabolomics data (metabolome) to identify shared and distinct molecular signatures associated with AD neuropathology.

We analyze prefrontal cortex samples from 157 donors from the Emory Goizueta Alzheimer's Disease Research Center brain bank. Neuropathologic outcomes include Braak staging for neurofibrillary tangle severity [22], the CERAD score for amyloid plaque density [24], and the composite ABC score combining Braak stage, CERAD, and Thal amyloid phase to assess overall AD pathology [20]. DNAm is measured using Illumina Infinium MethylationEPIC BeadChips (EPIC ; Illumina Inc., CA, USA), yielding 789,286 CpG sites after preprocessing. Metabolomic profiling is performed via liquid chromatography–high resolution mass spectrometry (LC-HRMS) using both HILIC and C18 chromatography modes, resulting in 35,348 high-quality features. Covariates (Z) include age at death, sex, race, educational attainment, postmortem interval (PMI), and neighborhood deprivation measured via the Area Deprivation Index [56].

We develop omics-based profile scores (*PS*) to predict AD neuropathology using single-, joint- and multi-omics strategies. These scores serve as aggregate covariates that summarize the relevant molecular features associated with the ordinal outcome $Y$, e.g. the ABC score. Following the framework established by [57], [58], we compute weighted profile scores

based on internally estimated feature weights in the absence of reliable external information. A profile score is defined as a linear combination of selected, standardized features:

$$PS = \sum_{j=1}^{K} w_j \cdot x_j \,,$$

where $w_j$ denotes the weight of the $j^{th}$ feature and/or CpG-site, $x_j$ is the actual value of the $j^{th}$ feature and/or CpG-site, and $K$ is the number of selected features.

After merging and standardizing the omics datasets, we randomly split the data into 50% training and 50% test sets. Importantly, the feature selection and weight estimation procedures were performed solely on the training set to ensure an unbiased evaluation. To enhance the robustness and reproducibility of our findings, this entire process—including data splitting, feature selection, and weight calculation—was repeated across 10 random seeds, following established protocols from previous studies [57], [58] . The resulting profile scores were then validated independently in the test datasets.

We construct (i) single-omics scores for DNAm and metabolomics separately, (ii) joint score models that include interaction terms, and (iii) a combined multi-omics score. Each score is evaluated using an ordinal logistic regression model:

(i) Single-omics *PS* model (e.g., DNAm or metabolomics):

$$logit(\mathbb{P}(Y \leq k)) = \theta_k + \beta \cdot PS_{DNAm}\,(or\ PS_{metabolome}) + \gamma^T Z \ ,$$

(ii) Multi-omics *PS* model (integrating DNAm and metabolomics CpG-sites/features into one combined data set):

$$logit(\mathbb{P}(Y \leq k)) = \theta_k + \beta \cdot PS_{DNAm+meta} \quad + \gamma^T Z \ ,$$

(iii) Joint *PS* model (including both single-omics PS with and without interaction term):

$$logit(\mathbb{P}(Y \leq k)) = \theta_k + \beta_1 \cdot PS_{DNAm} + \beta_2 \cdot PS_{metabolome}\,[+\beta_3 \cdot (PS_{DNAm} \cdot PS_{metabolome})] + \gamma^T Z,$$

where $Y$ is the ordinal dependent variable with categories 1,2,…,T (e.g., ABC score), is the $k = 1, \ldots, T-1$ ordinal category, $\theta_k$ are the threshold parameters, $\beta, \beta_1, \beta_2, \beta_3$, are the *PS* coefficients, $Z$ is the vector of covariates, and $\gamma$ their respective coefficients.

We apply five methods to select features and estimate weights:

- **Pruning & Thresholding (PT):** We perform hierarchical clustering using the flashClust, ClassDiscovery, and cluster packages to reduce redundancy among features. We then apply ordinal logistic regression to select variables significantly associated with the outcome ($p < 0.05$), following procedures similar to those described in [59].

- **Elastic Net (EN):** We implement regularized ordinal regression using the ordinalNet package [60], which effectively handles multicollinearity and enables sparse variable selection in high-dimensional settings.

- **Boosting (BO):** We apply an ordinal regression adaptation of gradient boosting [61], which incrementally builds the prediction model and improves performance for ordered categorical outcomes.

- **Random Forest (RF):** We use the ranger package [62] to compute variable importance scores, which we then use as weights in the profile score calculation.

- **Window Approach (WA):** We calculate cross-leverage scores [63] across defined genomic or metabolic windows to identify highly influential variables, leveraging the correlation structure within regions.

Missing values in the metabolomics data are imputed using missRanger without predictive mean matching [64], a Random Forest-based imputation approach that is often applied in metabolomics research [65], [66], [67].

Model performance is evaluated via the partial McFadden's pseudo-$R^2$[68], which isolates the contribution of the *PS* beyond baseline covariates:

$$R^2_{partial} = \frac{ln\, L_{restricted} - ln\, L_{full}}{ln\, L_{restricted}} = 1 - \frac{ln\ \ full}{ln\, L_{restricted}},$$

where $ln\ L_{full}$ is the log-likelihood of the fitted model, including the *PS* and covariates and $ln\ L_{restricted}$ is the log-likelihood of the restricted model, which includes covariates only. We further explore underlying biology by conducting KEGG pathway enrichment on selected CpG sites and metabolites using missMethyl and metapone, respectively [69], [70].

## 2.4 Model Comparisons of Frequentist and Bayesian Hierarchical Logistic Regression

Respiratory diseases in pig production are a major concern in veterinary public health and animal welfare. Despite ongoing advances in biosecurity and diagnostics, farms with recurring respiratory problems in piglets pose a complex challenge due to interacting environmental, managerial, and animal-level risk factors. The goal of this study is to identify and quantify those risk factors that significantly contribute to the occurrence of coughing in weaner pigs - a practical and routinely used proxy for respiratory health [71].

To address this, we evaluate a multilevel dataset collected from 30 pig-producing farms in Germany with known histories of respiratory health issues. These farms were selected as part of the VASIB project. Data were collected across four hierarchical levels: farm, compartment, pen, and individual animal, resulting in a nested structure with $n = 450$ animals distributed across 300 pens, 130 compartments, and 72 pigsties. Our research question is twofold:

(1) Which environmental, management, or animal-based variables significantly affect coughing probability?

(2) Which statistical modelling framework - frequentist or Bayesian, hierarchical or flat - provides the most robust, interpretable, and generalizable inference?

Data acquisition followed a standardized protocol. Before each veterinary inspection, participating farms completed a pre-tested questionnaire covering structural, hygienic, and management variables. These were verified and completed during onsite clinical examinations by veterinarians. All data were consolidated in a relational SQL database, checked for consistency, and prepared for analysis. Missing data were handled using multilevel multiple imputation methods [72], [73], ensuring preservation of the data's clustering structure. To address missing data in continuous variables, we applied predictive mean matching (PMM), which preserves the distributional characteristics of the original data by drawing observed values from similar cases [72]. PMM is a semi-parametric method that combines regression modeling with observed donor values. Specifically, for each case with missing data, a predictive value is first estimated using linear regression. Then, instead of filling in this estimated value directly, the algorithm finds a set of "donor" observations with similar predicted values and randomly selects one of their observed values for imputation. This approach retains the original distribution and ensures that imputed values are realistic

and consistent with the observed data space [74]. For ordinal categorical variables, we used a proportional odds logistic regression (POLR) model, which accounts for the inherent ordering of categories and provides consistent imputations under the assumption of proportional odds [73], [75]. In POLR, the log-odds of being in a category or below are modeled as a linear function of covariates. The key assumption is that the effect of covariates is the same across all thresholds (the proportional odds assumption). This ensures that ordinal structure is respected during imputation, unlike methods that treat categories as purely nominal. Both procedures were implemented using the mice package in R [74].

The primary outcome variable is binary:

$$y_{ijkl} = \begin{cases} 1, & if\ pig\ l\ in\ pen\ k, compartment\ j, farm\ i\ coughs, \\ 0, & otherwise \end{cases}$$

($l = 1,\dots,n_{ijk}\ pigs,\ k = 1,\dots,n_{ij}$ pens, $j = 1,\dots,n_i$ compartments and $i = 1,\dots,n$ farms). Covariates include continuous variables (e.g., temperature, $CO_2$, stocking density) and categorical factors (e.g., floor condition, skin lesions).

We model the probability of coughing using generalized linear models with a logit link. Let $\pi_{ijkl} = \mathbb{P}(y_{ijkl} = 1)$ denote the probability that pig $l$ in pen $k$ of compartment $j$ in farm $i$ coughs and $\mathrm{x}_{ijkl}$ the corresponding covariate value. The general model structure is:

$$logit(\pi_{ijkl}) = \beta_0 + \sum_{p=1}^{P} \beta_{\mathrm{p}} \mathrm{x}_{ijkl} + u_i + v_{ij} + w_{ijk}.$$

Here $P$ is the number of covariates, $\beta_0$ is the fixed global intercept, $\beta_p$ represents fixed effects for covariates $x_{ijkl}$, and $u_{\mathrm{i}} \sim N(0, \sigma_{\mathrm{u}}^2)$, $v_{\mathrm{ij}} \sim N(0, \sigma_{\mathrm{v}}^2)$ and $w_{ijk} \sim N(0, \sigma_w^2)$ are random intercepts for farms, pens and compartments, respectively and this model best reflects the hierarchical structure of the data.

This hierarchical logistic regression accounts for within-cluster dependencies at multiple levels and avoids artificially narrow confidence intervals or overconfident predictions [76], [77].

We compare four model classes:

(i) Frequentist hierarchical logistic regression (maximum likelihood)

(ii) Bayesian non-hierarchical model with non-informative priors

(iii) Bayesian hierarchical models with non-informative priors

### (iv) Bayesian hierarchical models with highly informative priors

We estimated the Bayesian regression models using the Hamiltonian Monte Carlo (HMC) algorithm as implemented in the brms package [78], which serves as a user-friendly interface to the probabilistic programming language Stan. HMC is a state-of-the-art method for drawing samples from complex posterior distributions and is particularly efficient in high-dimensional parameter spaces, as it avoids the slow mixing and high autocorrelation that often affect traditional Markov Chain Monte Carlo (MCMC) methods like Gibbs sampling or Metropolis-Hastings [30]. To ensure robust estimation, we ran four parallel chains with 5,000 iterations each. The first 2,500 iterations per chain were used for warm-up (also called burn-in), during which the algorithm adapts its internal parameters (e.g., step size, mass matrix) to efficiently explore the posterior distribution. The remaining 10,000 post-warm-up samples (4 chains × 2,500 samples) were then used for inference, i.e., estimating parameters and quantifying uncertainty. Convergence of the chains was monitored using the $\hat{R}$ statistic (target value: < 1.01) and by visually inspecting trace plots [79], [80].

A central feature of Bayesian modelling is the use of prior distributions, which encode existing knowledge or assumptions about the parameters before observing the data. For example, if we are estimating the effect of stocking density on coughing, a non-informative prior (e.g., a wide normal distribution with large variance) expresses that we have no strong expectations about the effect's direction or magnitude. Conversely, a highly informative prior can be used when expert knowledge or previous studies suggest a plausible range for a parameter - for instance, assuming that higher stocking density likely increases respiratory issues. Formally, if $\beta$ denotes a regression coefficient, we might specify:

- Non-informative prior:
  $\beta \sim \mathcal{N}(0, 10^2)$, allowing a wide range of plausible values,
- Weakly informative prior:
  $\beta \sim \mathcal{N}(0, 2.5^2)$, narrowing the range but still flexible,
- Informative prior:
  $\beta \sim \mathcal{N}(1.0, 0.5^2)$, concentrating the prior around a known or expected effect.

These priors are then updated through Bayes' Theorem by combining them with the data via the likelihood function, resulting in the posterior distribution -the final estimate of what we believe about each parameter after seeing the data. Priors were assigned to all model

parameters; in models using non-informative priors, wide distributions were specified to reflect minimal prior knowledge, whereas in models with highly informative priors, prior distributions were explicitly tailored based on expert knowledge and existing literature. In some Bayesian models, particularly for variance parameters, we considered specifying an Inverse-Gamma distribution as a prior, which is commonly used to ensure that variance estimates remain strictly positive and to express prior beliefs about the likely scale of variability. By applying this Bayesian workflow [30], [79], [81], we were able to quantify the full uncertainty around each parameter and generate posterior credible intervals, which are the Bayesian equivalent of confidence intervals but with a more direct probabilistic interpretation: for example, there is a 95% probability that the true effect lies within the 95% credible interval.

To compare model performance, we used several evaluation metrics:

- Marginal $R_m^2$: proportion of variance explained by fixed effects
- Conditional $R_c^2$: total variance explained by fixed and random effects
- Intraclass Correlation Coefficient (ICC): measures clustering strength at each level
- Akaike Information Criterion corrected ($AIC_C$): model fit penalized for complexity (for frequentist models)
- Leave-One-Out Cross Validation (LOO-CV): out-of-sample predictive accuracy (for Bayesian models)
- Bayes $R^2$: posterior predictive performance, analogous to classical $R^2$

Higher conditional $R^2$ and ICC values indicate that a significant portion of variability is due to cluster-level effects, justifying the use of hierarchical models. The Bayesian LOO approach estimates the expected log predictive density (elpd) and its standard error. Models with higher elpd values are preferred for prediction.

All analyses were conducted in R version 4.0.5. [82]. Beyond the aforementioned packages, we used performance measures for model diagnostics and tidybayes for posterior processing and visualization [83], [84].

# 3 Summary of the Articles

## 3.1 Article 1: Statistical analysis of *in vivo* alkaline comet assay data - Comparison of median and geometric mean as centrality measures

The first article addresses the challenge of selecting suitable statistical measures to analyze data derived from comet assays, specifically comparing the use of medians and geometric means for summarizing slide data in *in vivo* experiments. The comet assay, also known as single-cell gel electrophoresis, is a well-established technique for detecting DNA strand breaks and assessing potential genotoxicity. Its wide use results from its sensitivity, rapidity, cost-effectiveness, and applicability across various tissues, and regulatory guidelines, such as OECD guideline TG 489, recommend it as part of standard toxicological evaluations.

In this study, we investigate how choosing different statistical measures - median versus geometric mean - to represent central tendency affects statistical outcomes and the subsequent interpretation of experimental results. Currently, the OECD guideline [18] recommends summarizing comet assay data per slide using medians and then averaging these medians per animal. However, alternative methods like the geometric mean are permissible with scientific justification. To illustrate potential issues arising from different summarization approaches, we provide an illustrative example using artificially simulated data. The experiment includes a vehicle control group and multiple dose groups, with slide data summarized using median, arithmetic mean, and geometric mean. We find significant differences in the dose-response relationships depending on the chosen summary measure, potentially influencing biological interpretations.

Additionally, we use real-world experimental data as the foundation for our simulation studies. This data originates from five independent studies conducted by Bayer AG, employing standardized experimental designs involving male rats with liver as the target tissue. Each experiment includes control and treatment groups at various dosages, six animals per group, and multiple slides per animal. The primary endpoint measured is tail intensity (% tail DNA), characterized by a strongly right-skewed distribution. We apply two statistical modeling approaches: a "statistical slide model," treating each slide as repeated measures per animal, and a "statistical animal model," summarizing slides into one value per animal. We compare dose groups of the test item to the corresponding negative control, and

in addition, we perform down-turn protected trend tests to address potential non-monotonic dose-response relationships.

We supplement our experimental datasets and their results with simulation studies, generating the corresponding data based on these realistic scenarios and distributions. We employ gamma and lognormal distributions to replicate the skewness and variability observed in actual comet assay data. Systematic variations in animal numbers, slides per animal, and cells per slide enable comprehensive power analyses. Despite extensive analyses, our study has several limitations. The datasets primarily originate from specialized experiments, potentially restricting the generalizability of our findings. Additionally, we generate simulated data based on specific assumptions and distributions, which might not fully represent all realistic experimental scenarios. Furthermore, variance heteroscedasticity in the data complicates the application of down-turn protected trend tests, necessitating alternative statistical approaches. Therefore, we recommend further studies utilizing larger, more diverse datasets and advanced methods to comprehensively validate our results and enhance methodological robustness.

Our key findings demonstrate that choosing between median and geometric mean profoundly impacts statistical power and the resulting conclusions of comet assays. Specifically, median-based analyses often yield higher statistical power compared to geometric means, notably in scenarios with moderate differences among treatment groups. Power analyses show substantial gains in detection capability with increasing numbers of animals, slides, and cells analyzed. Moreover, we emphasize that while both statistical measures maintain appropriate significance levels under null hypotheses, the differences observed under alternative hypotheses can lead to conflicting interpretations regarding genotoxicity.

We conclude that selecting the measure of central tendency per slide significantly influences comet assay outcomes and interpretations. Therefore, careful consideration and scientific justification are crucial when deviating from guideline-recommended statistical measures. Additionally, we advocate further research using extensive data sets, additional summarization techniques, and more complex modeling to validate these findings and promote methodological consistency across diverse comet assay studies.

## 3.2 Article 2: *In vivo* alkaline comet assay: Statistical considerations on historical negative and positive control data

In the second article, we explore critical statistical aspects for analyzing historical control data (HCD) in the context of the *in vivo* alkaline comet assay. This assay is widely used to assess genotoxic potential by detecting DNA damage, including single- and double-strand breaks, as well as lesions resulting from incomplete DNA excision repair. We highlight that, although OECD Test Guideline 489 [18] emphasizes the importance of statistical analyses and historical control data, it does not provide detailed methodological guidance. Building on findings from the first paper, our aim is to use real-world data to confirm or refute prior results and to investigate additional unresolved statistical questions related to the comet assay. Therefore, the German-speaking Society for Environmental Mutation Research (GUM) working group conducts extensive analyses of real-world comet assay data from over 200 studies performed across five laboratories. We observe significant inter-laboratory variability, challenging the usefulness of absolute quality thresholds for tail intensity (TI), the primary measure of DNA damage. We also highlight the substantial issue posed by zero-values on slides, noting that occurrences exceeding 50% significantly influence statistical summarization and subsequent analyses at both slide and animal levels. Our study evaluates various summarizing strategies in detail, comparing medians, arithmetic means, geometric means, and trimmed means. The results demonstrate that the choice of summary measures significantly impacts outcomes, particularly for negative control data. Arithmetic means yield notably higher TI values, whereas medians and geometric means produce lower, more conservative estimates, consequently affecting interpretations of genotoxic potential.

We further address the handling of zero-values, a frequent challenge in comet assays, particularly in negative controls. Our findings support the OECD TG 489 recommendation to add a small constant (0.001) to cell-level values, facilitating logarithmic transformations essential for achieving approximate symmetry in data distribution. Extensive real-world data analyses confirm the appropriateness of this constant, prompting our recommendation that laboratories maintain data precision to at least three decimal places to minimize distortion. Additionally, we investigate meta-parameters such as electrophoresis duration, staining procedures, and vehicle types to assess their potential influence on TI values. Due to variability across laboratories, definitive conclusions about individual parameters remain challenging, although vehicle type within laboratories demonstrates some effect, suggesting laboratory-specific optimization is essential. We conduct comparative analyses of negative

control (NC) and positive control (PC) data, predominantly using Ethyl methane sulfonate (EMS) as the positive control substance. Our analyses clearly differentiate NC from PC groups, validating the reliability of historical control data across varying EMS concentrations. We emphasize maintaining a distinct dynamic range between these controls as crucial for assay validation and accurate genotoxic assessments. A comprehensive variance component analysis explores variability sources within and between studies. We apply mixed-effects models to decompose variance, concluding significant differences in within-study versus between-study variability among laboratories. In some laboratories, within-study variability predominates, enhancing robustness for historical comparisons, whereas others exhibit higher between-study variability, potentially undermining historical comparisons' validity.

The applied statistical methods depend significantly on the distributional assumptions of the data, which often do not hold true for comet assay data due to their inherent right-skewness and the frequent occurrence of zero-values. The use of summarizing measures such as arithmetic or geometric means may lead to a loss of information, especially if a high percentage of zero-values exists. Additionally, the large observed variability between laboratories limits the generalizability of specific statistical thresholds across different experimental settings. Furthermore, results strongly rely on historical control data quality, thus inaccuracies or incomplete datasets from past experiments may reduce the reliability of comparisons and conclusions. Lastly, due to the fixed experimental settings within each laboratory, it is impossible to statistically identify the isolated effects of specific meta-parameters like electrophoresis conditions or staining methods on DNA damage results We stress the importance of laboratories demonstrating consistent proficiency in experimental conditions to ensure low basal DNA damage levels and reliable positive control responses. Advanced statistical methodologies, including zero-inflation models, are recommended when zero-value occurrences are high.

In conclusion, our paper provides essential insights into the management, summarization, and statistical analysis of historical comet assay data. It underscores the need for careful selection of statistical methods, robust experimental protocols, and meticulous data handling to ensure valid interpretations and regulatory compliance. We recommend detailed reporting of experimental conditions, validation of summary measures, and vigilant control of zero-value occurrences to enhance the assay's reliability and interpretability.

## 3.3 Article 3: Development and Application of Brain Tissue-Based Multi-Omics Profile Scores for Alzheimer's Disease

The third article investigates how integrating multi-omics data, specifically genome-wide DNA methylation (DNAm) and high-resolution metabolomics, can enhance insights into Alzheimer's Disease (AD) neuropathology. Recognizing the complexity of AD, we aim to determine whether combined omics layers more effectively capture biological mechanisms compared to single-omics approaches.

We perform a comprehensive multi-omics analysis on 157 frontal cortex samples from brain donors exhibiting varying stages of AD pathology, evaluated using established neuropathologic scores such as Braak staging, CERAD scoring, and the ABC score (Amyloid, Braak, CERAD) [19], [20], [21], [22], [23], [24]. We develop novel profile scores (*PS*) integrating DNAm and metabolomic data to predict AD neuropathology. These *PS* are computed using advanced statistical and Machine Learning methods, including Pruning & Thresholding (PT), Elastic Net (EN), Boosting (BO), Random Forest (RF), and the Windows Approach (WA). Our single-omics analyses show that DNAm-based *PS* ($PS_{\text{DNAm}}$) achieve stronger predictive performance (median $R^2 = 0.11$) compared to metabolomics-based *PS* ($PS_{\text{Metabolome}}$; median $R^2 = 0.04$). The median $R^2$ values are calculated across ten iterations. When combining DNAm and metabolomics data into a multi-omics *PS*, predictive accuracy modestly improves (median $R^2 = 0.15$). Notably, the DNAm and metabolomics *PS* moderately correlate (Pearson correlation of 0.25), highlighting complementary yet distinct biological contributions to AD neuropathology. Through pathway enrichment analyses, we identify several biological pathways significantly associated with AD across both omics layers. Key shared pathways include lipid metabolism and signal transduction, underscoring their critical roles in AD pathology. Specifically, lipid-related pathways such as alpha-linolenic acid metabolism and linoleic acid metabolism show significant enrichment, aligning with previous studies linking lipid dysregulation to neurodegeneration [85], [86]. Additionally, our DNAm analysis uniquely identifies multiple AD-relevant pathways, including glycerophospholipid metabolism, lipid metabolism, arachidonic acid metabolism, Ras signaling, complement system activation, and thermogenesis. These pathways influence neuroinflammation, oxidative stress, and neuronal health core elements of AD pathology [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97]. Metabolomics-specific analyses further highlight pathways such as arginine and proline metabolism, sphingolipid metabolism, steroid hormone biosynthesis, and cholesterol metabolism, previously

implicated in AD through their involvement in inflammation, synaptic integrity, and amyloid-beta dynamics.

Our advanced analytical approaches enable effective feature selection even within high-dimensional data, enhancing the robustness and interpretability of our results. Among the tested methods, PT and RF demonstrate superior effectiveness for DNAm and metabolomics, respectively, measured by $R^2$. Despite the strengths of our study - including unique integration of multi-omics data from brain tissues and sophisticated analytical techniques - limitations exist. Our sample, derived from a specialized brain bank predominantly containing AD cases, may limit generalizability. Additionally, analyzing postmortem brain tissue provides only a static view, potentially overlooking dynamic molecular changes occurring throughout AD progression.

In conclusion, our research demonstrates the potential of multi-omics integration to enhance predictive accuracy and clarify biological mechanisms underlying AD neuropathology. Identifying key molecular pathways offers promising targets for therapeutic intervention, emphasizing the importance of continued multi-omics research in addressing Alzheimer's disease.

## 3.4 Article 4: Hierarchical modelling of risk factors with and without prior information – the process of regression model evaluation for an example of respiratory diseases in piglet production from daily practice data

The fourth article explores statistical modeling approaches for assessing respiratory health in piglet production using data from the VASIB project, focusing on 30 German farms experiencing persistent respiratory issues in weaner pigs. The objective is to identify risk factors influencing pig coughing, a primary clinical sign serving as an indicator for respiratory disease in veterinary practice. We collect data across multiple hierarchical levels - farm, compartment, pen, and individual animal - covering environmental conditions, management practices, and animal health indicators. The central challenge addressed is the complexity of modeling health outcomes given clustered and incomplete real-world data. Traditional regression models often struggle under these conditions, particularly with small sample sizes and high intra-cluster correlations. To address this complexity, we explore and compare frequentist and Bayesian hierarchical logistic regression models, including variants using informative and non-informative priors. Initially, hierarchical frequentist models (HFMs) quantify clustering contributions to outcome variability. Our results reveal that about 35 - 46 % of cough incidence variation is due to effects at the pen and compartment levels. Models incorporating additional fixed effects such as environmental and animal-level variables identify stocking density, animal contamination, and flooring conditions as influential factors, though statistical significance varies. We subsequently apply Bayesian models to address convergence issues and incorporate prior knowledge. We evaluate five Bayesian models: one non-hierarchical and four hierarchical (two with non-informative and two with informative priors). Hierarchical Bayesian models consistently outperform non-hierarchical alternatives. The model with highly informative priors (BM5) achieves the best predictive accuracy, measured via leave-one-out cross-validation (LOO), and the highest Bayes $R^2$ ($\approx$0.41). Across models, worn flooring significantly increases the odds of coughing (OR $\approx$ 4 - 6), as does high stocking density, emphasizing the importance of environmental hygiene and space management in preventing respiratory diseases. Additionally, water flow rate and temperature emerge as either protective or aggravating factors depending on the selected model.

Hierarchical modelling provides essential advantages for evaluating respiratory disease risk

factors in piglet production. Frequentist regression models encounter convergence issues due to small sample sizes and complex data structures, whereas Bayesian models effectively handle these challenges. Bayesian hierarchical models particularly show superior performance, offering more reliable results by incorporating prior information and adequately representing the nested data structure. Model choice significantly influences which risk factors, such as floor condition or stocking density, are identified as statistically significant. A balanced approach considering both statistical robustness and practical interpretability remains crucial for model selection in veterinary epidemiology. Specifically, BM5 exhibits the highest explanatory power and the most realistic effect estimates. One advantage compared to the train test scenario in Article 3 is the use of leave-one-out cross-validation, as the computational effort here is much lower and manageable. The analysis demonstrates that both frequentist and Bayesian approaches consistently identify key factors influencing coughing in pigs - most notably floor condition, stocking density, and water flow - with worn floors and reduced water flow significantly increasing risk, while a larger pen size shows a non-significant protective effect (OR = 0.79, [95% CI: 0.51,1.24]), ultimately indicating that, despite continuous veterinary supervision, some hygiene measures are either insufficiently implemented or practically difficult to realize, thereby underscoring the need for alternative strategies.

In conclusion, hierarchical Bayesian modeling with informative priors proves to be a powerful tool for analyzing veterinary health data in routine practice. Our findings underscore the critical importance of improving flooring conditions, managing stocking densities, and optimizing environmental parameters to reduce respiratory disease incidence in pig production.

# 4 Discussion and Outlook

This cumulative PhD thesis presents four partly peer-reviewed manuscripts that collectively contribute to the methodological development and practical application of statistical modeling in biomedical and veterinary sciences. Each paper addresses a distinct challenge - ranging from regulatory toxicology to epidemiological modeling and multi-omics integration - yet all share a focus on developing robust, interpretable, and context-sensitive analytical strategies. Throughout the thesis, we emphasize how thoughtful model design and statistical methodology directly enhance decision-making in contexts characterized by data complexity, limited sample sizes, and real-world constraints.

In the first manuscript, we investigate the influence of different centrality measures in the evaluation of the *in vivo* alkaline comet assay [48], a widely applied method in regulatory genotoxicity testing. We compare the median and geometric mean as slide-level summaries for the key endpoint “tail intensity” and demonstrate - through empirical analysis and simulation studies - that this choice significantly affects statistical significance and toxicological conclusions. Given that OECD guideline 489 [18] recommends the median but permits alternative approaches if scientifically justified, these findings hold regulatory relevance. We show that geometric means, while more sensitive in right-skewed distributions, may lead to diverging interpretations compared to medians. We apply a downturn-protected trend test to better address non-monotonic dose-response patterns, such as those arising from cytotoxicity at high doses. This work highlights the importance of critically evaluating statistical assumptions and summary strategies in regulatory settings. We also begin incorporating mixed models to appropriately reflect the nested structure of the data (e.g., slides within animals), capturing within-subject variability via random effects. Furthermore, we develop simulation-based validation frameworks to test the robustness of trend tests under varying distributional assumptions, sample sizes, and skewness levels. The gamma-distributed simulation framework introduced in this study provides a solid basis for future methodological evaluations. Ongoing work also explores alternative summary statistics such as trimmed or weighted means. As demonstrated, the choice between median and geometric mean can substantially alter test outcomes, with direct implications for toxicological interpretation and regulatory decision-making.

In the second manuscript, developed in collaboration with the German-speaking Society for Environmental Mutation Research (GUM), we extend the analysis of the *in vivo* comet assay

by incorporating historical control data (HCD) from over 200 genotoxicity studies across five laboratories [98]. We evaluate statistical procedures for comparing negative and positive controls, identify substantial inter-laboratory variability, and critically assess the use of absolute thresholds in study evaluation. We examine the impact of various summary statistics - including arithmetic mean, geometric mean, and median - and highlight the role of zero inflation and variance components in shaping study conclusions. By drawing on a comprehensive real-world dataset, we provide empirical evidence in support of refined statistical guidance beyond what is currently specified in OECD TG 489 [18]. One of the key findings reveals that the way single-cell measurements are summarized plays a critical role in interpretation, and that arbitrary thresholds - if applied without context-sensitive variance modeling - can obscure true biological signals. Future research directions include the development of Bayesian hierarchical frameworks for dynamic control limit estimation and the use of shrinkage estimators to improve cross-laboratory harmonization. Following our productive collaboration in the first two studies, we have initiated further projects, such as the development of a Shiny application for comet assay data and extended comparisons of analysis systems on identical slides.

The third manuscript shifts focus to translational neuroscience and presents the development of brain tissue–based multi-omics profile scores (*PS*) for Alzheimer's disease (AD). In this study, we integrate genome-wide DNA methylation and high-resolution metabolomics data from 157 postmortem frontal cortex samples using several Machine Learning algorithms - Random Forest, Elastic Net, Pruning and Thresholding, Boosting, and the so-called Windows Approach - to construct *PS* predictive of neuropathological AD staging (e.g., Braak, CERAD, and ABC scores) [80]. Two of these algorithms, Boosting and the Windows Approach, were developed by other doctoral researchers within the RTG 2624 "Statistical Methods for High-Dimensional Data in Toxicology." These methods perform comparably to Random Forest but with substantially lower runtime, making them promising candidates for future applications. A bachelor's thesis is currently optimizing the parameters of the Windows Approach, while an additional Sketching method derived from these techniques already shows improved results in preliminary testing. Our analysis confirms that multi-omics *PS* outperform single-omics scores in terms of predictive accuracy and biological relevance. Pathway analyses consistently reveal enrichment in lipid metabolism and signal transduction processes, offering mechanistic insight into AD pathophysiology. To address challenges posed by high-dimensional covariates, limited sample sizes, and ordinal

outcomes, we implement rigorous cross-validation strategies and penalized regression models. Future directions include applying transfer learning across cohorts, validating *PS* and selected features (e.g., CpG sites) in external datasets, and integrating further omics layers such as transcriptomics or proteomics. The selection of hyperparameters for the various Machine Learning algorithms remains an open question and is currently under investigation as part of a student project.

In the fourth and final manuscript, we apply advanced regression modeling strategies to investigate environmental and management-related risk factors for respiratory diseases in piglet production. Using data from the VASIB project - which includes 30 piglet-producing farms and over 400 animals - we compare frequentist and Bayesian logistic regression models, including hierarchical models with informative priors, to model coughing symptoms in piglets [81]. The hierarchical structure of the dataset (farm, compartment, pen, animal) poses significant challenges for model convergence and variable selection. Our Bayesian models, particularly those incorporating prior information, provide improved interpretability and robustness - especially when handling missing data and correlated covariates. The analysis reveals that both frequentist and Bayesian approaches identify key factors influencing coughing in pigs, including stocking density, floor condition, and water flow. We consistently identify floor condition as a major risk factor, while the influence of other variables - such as water flow rate, temperature, and stocking density - varies depending on the model used. An increase in pen size shows a protective effect (OR = 0.79), although it does not reach statistical significance. Pigs housed on worn floors experience more than five times higher odds of coughing compared to those on new floors, emphasizing the critical role of environmental hygiene. Higher water flow rates also demonstrate a protective influence (OR = 0.57), underscoring the importance of adequate hydration. Despite continuous veterinary supervision in the studied farms, these findings suggest that some known hygiene measures remain either under-implemented or practically difficult to realize. This study highlights the value of flexible statistical methods for epidemiological analyses based on routine veterinary data. To advance this field, we propose future research to explore dynamic models for longitudinal herd health monitoring, apply empirical Bayes techniques for informed prior specification, and extend analyses to multivariate outcomes such as co-occurring symptoms or composite animal welfare indices. Since the current dataset comes from a specific German region with a high burden of respiratory disease, it remains uncertain whether similar data exist in other national or international contexts to support broader

generalization.

Taken together, these four studies demonstrate the potential of modern statistical modeling to improve inference, prediction, and interpretability in fields as diverse as regulatory toxicology, neuroscience, and veterinary public health. A common thread throughout this thesis is the recognition that rigid, one-size-fits-all approaches are insufficient for the complexity of real-world biomedical data. Instead, we advocate for context-aware, data-driven, and flexible methodological strategies. Across all manuscripts, we show how the quality of statistical inference critically depends on aligning assumptions and models with data structure - whether by choosing appropriate summary statistics, applying hierarchical modeling, or integrating high-dimensional multi-omics data.

Looking ahead, we see great potential in synthesizing elements from all four studies into unified methodological frameworks. For example, combining Bayesian modeling techniques from the toxicology and veterinary work with the machine learning–based variable selection approaches from the Alzheimer's study could yield powerful hybrid tools for translational research. Likewise, the simulation-based model validation strategies developed in the comet assay work could guide the development of robust evaluation pipelines in omics-based prediction models.

In conclusion, this thesis makes a significant contribution to the methodological foundation required to meet the growing complexity of biomedical and veterinary data analysis. It creates a pathway for future interdisciplinary progress in health research. Across all four projects, our collaboration between applied researchers and statisticians proves highly productive and insightful. Through close interdisciplinary exchange, we develop a deeper mutual understanding of each other's perspectives, challenges, and needs. By bridging methodological innovation and practical application, we succeed in connecting distinct scientific disciplines and fostering lasting synergies. While it takes time and effort to convince both experienced professionals and early-career scientists that statistics is not an obstacle but a powerful ally, the process ultimately enables a shift in mindset. By demonstrating how thoughtful statistical tools clarify data, reveal meaningful patterns, and support better decisions, we lay the foundation for more open, collaborative, and data-driven scientific inquiry.

# Bibliography


[1] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis, 'What does research reproducibility mean?', *Sci Transl Med*, vol. 8, no. 341, pp. 341ps12-341ps12, Jun. 2016, doi: 10.1126/scitranslmed.aaf5027.

[2] C. G. Begley and J. P. A. Ioannidis, 'Reproducibility in Science', *Circ Res*, vol. 116, no. 1, pp. 116–126, Jan. 2015, doi: 10.1161/CIRCRESAHA.114.303819.

[3] Esosa Enoyoze and Goddidit Esiro Enoyoze, 'Statistical applications in the biomedical sciences: A review', *International Journal of Science and Research Archive*, vol. 12, no. 2, pp. 1594–1601, Aug. 2024, doi: 10.30574/ijsra.2024.12.2.1433.

[4] S. Greenland *et al.*, 'Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations', *Eur J Epidemiol*, vol. 31, no. 4, pp. 337–350, Apr. 2016, doi: 10.1007/s10654-016-0149-3.

[5] S. Holmes and W. Huber, *Modern Statistics for Modern Biology*. Cambridge University Press, 2019.

[6] R. L. Wasserstein and N. A. Lazar, 'The ASA Statement on *p* -Values: Context, Process, and Purpose', *Am Stat*, vol. 70, no. 2, pp. 129–133, Apr. 2016, doi: 10.1080/00031305.2016.1154108.

[7] N. G. D. Masca *et al.*, 'Science Forum: RIPOSTE: a framework for improving the design and analysis of laboratory-based research', *Elife*, vol. 4, p. e05519, May 2015, doi: 10.7554/eLife.05519.

[8] Y. Hasin, M. Seldin, and A. Lusis, 'Multi-omics approaches to disease', *Genome Biol*, vol. 18, no. 1, p. 83, 2017, doi: 10.1186/s13059-017-1215-1.

[9] N. Rappoport and R. Shamir, 'Multi-omic and multi-view clustering algorithms: review and cancer benchmark', *Nucleic Acids Res*, vol. 46, no. 20, pp. 10546–10562, Nov. 2018, doi: 10.1093/nar/gky889.

[10] B. Wang *et al.*, 'Similarity network fusion for aggregating data types on a genomic scale', *Nat Methods*, vol. 11, no. 3, pp. 333–337, Mar. 2014, doi: 10.1038/nmeth.2810.

[11] A. R. Baião *et al.*, 'A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches', Jan. 2025, doi: 10.48550/arXiv.2501.17729.

[12] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, 'Multi-omics Data Integration, Interpretation, and Its Application', *Bioinform Biol Insights*, vol. 14, p. 117793221989905, Jan. 2020, doi: 10.1177/1177932219899051.

[13] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. Wiley, 2009. doi: 10.1002/9780470434697.

[14] J. A. C. Sterne, 'Sifting the evidence---what's wrong with significance tests? Another comment on the role of statistical methods', *BMJ*, vol. 322, no. 7280, pp. 226–231, Jan. 2001, doi: 10.1136/bmj.322.7280.226.

[15] S. Brendler-Schwaab, A. Hartmann, S. Pfuhler, and G. Speit, 'The in vivo comet assay: use and status in genotoxicity testing', *Mutagenesis*, vol. 20, no. 4, pp. 245–254, Jul. 2005, doi: 10.1093/mutage/gei033.

[16] D. P. Lovell and T. Omori, 'Statistical issues in the use of the comet assay', *Mutagenesis*, vol. 23, no. 3, pp. 171–182, Feb. 2008, doi: 10.1093/mutage/gen015.

[17] J. Menz *et al.*, 'Genotoxicity assessment: opportunities, challenges and perspectives for quantitative evaluations of dose–response data', *Arch Toxicol*, vol. 97, no. 9, pp. 2303–2328, Sep. 2023, doi: 10.1007/s00204-023-03553-w.

[18] OECD, *Test No. 489: In Vivo Mammalian Alkaline Comet Assay*. Paris: OECD, 2016. doi: 10.1787/9789264264885-en.

[19] L. M. Besser *et al.*, 'The revised national Alzheimer's coordinating center's neuropathology form-available data and new analyses', *J Neuropathol Exp Neurol*, vol. 77, no. 8, pp. 717–726, Aug. 2018, doi: 10.1093/jnen/nly049.

[20] T. J. Montine *et al.*, 'National Institute on aging-Alzheimer's association guidelines for the neuropathologic assessment of Alzheimer's disease: A practical approach', *Acta*

*Neuropathol*, vol. 123, no. 1, pp. 1–11, Jan. 2012, doi: 10.1007/s00401-011-0910-3.
[21] M. A. Deture and D. W. Dickson, 'The neuropathological diagnosis of Alzheimer's disease', *Mol Neurodegener*, vol. 14, no. 1, p. 32, Aug. 2019, doi: 10.1186/s13024-019-0333-5.
[22] H. Braak and E. Braak, 'Neuropathological stageing of Alzheimer-related changes', *Acta Neuropathol*, vol. 82, no. 4, pp. 239–259, Sep. 1991, doi: 10.1186/s13024-019-0333-5.
[23] D. R. Thal, U. Rüb, M. Orantes, and H. Braak, 'Phases of Aβ-deposition in the human brain and its relevance for the development of AD', *Neurology*, vol. 58, no. 12, pp. 1791–1800, Jun. 2002, doi: 10.1212/WNL.58.12.1791.
[24] S. S. Mirra *et al.*, 'The Consortium to Establish a Registry for Alzheimer's Disease (CERAD)', *Neurology*, vol. 41, no. 4, pp. 479–479, Apr. 1991, doi: 10.1212/WNL.41.4.479.
[25] N. D. Nguyen and D. Wang, 'Multiview learning for understanding functional multiomics', *PLoS Comput Biol*, vol. 16, no. 4, pp. e1007677-, Apr. 2020, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1007677
[26] G. P. Way and C. S. Greene, 'Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders.', *Pac Symp Biocomput*, vol. 23, pp. 80–91, 2018, Accessed: Jun. 03, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5728678/
[27] A. Alkhateeb and L. Rueda, Eds., *Machine Learning Methods for Multi-Omics Data Integration*. Cham: Springer International Publishing, 2024. doi: 10.1007/978-3-031-36502-7.
[28] L. Zhang *et al.*, 'Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma', *Front Genet*, vol. 9, Oct. 2018, doi: 10.3389/fgene.2018.00477.
[29] R. Chen and M. Snyder, 'Promise of personalized omics to precision medicine', *WIREs Systems Biology and Medicine*, vol. 5, no. 1, pp. 73–82, Jan. 2013, doi: 10.1002/wsbm.1198.
[30] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, Third. in Chapman & Hall/CRC Texts in Statistical Science Series. Boca Raton, Florida: CRC, 2013. [Online]. Available: https://stat.columbia.edu/~gelman/book/
[31] I. R. Dohoo, W. Martin, and H. E. Stryhn, *Veterinary epidemiologic research*. University of Prince Edward Island, 2014. Accessed: Jun. 03, 2025. [Online]. Available: https://projects.upei.ca/ver/
[32] M. P. Ward and F. I. Lewis, 'Bayesian Graphical modelling: Applications in veterinary epidemiology', *Prev Vet Med*, vol. 110, no. 1, pp. 1–3, May 2013, doi: 10.1016/j.prevetmed.2013.02.007.
[33] A. Biggeri, E. Dreassi, D. Catelan, L. Rinaldi, C. Lagazio, and G. Cringoli, 'Disease mapping in veterinary epidemiology: a Bayesian geostatistical approach', *Stat Methods Med Res*, vol. 15, no. 4, pp. 337–352, Aug. 2006, doi: 10.1191/0962280206sm455oa.
[34] G. Kratzer *et al.*, 'Bayesian Network Modeling Applied to Feline Calicivirus Infection Among Cats in Switzerland', *Front Vet Sci*, vol. 7, Feb. 2020, doi: 10.3389/fvets.2020.00073.
[35] N. P. Singh, M. T. McCoy, R. R. Tice, and E. L. Schneider, 'A simple technique for quantitation of low levels of DNA damage in individual cells', *Exp Cell Res*, vol. 175, no. 1, pp. 184–191, 1988, doi: 10.1016/0014-4827(88)90265-0.
[36] G. Speit *et al.*, '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)', *Mutat Res Genet Toxicol Environ Mutagen*, vol. 783, pp. 6–12, 2015, doi: 10.1016/j.mrgentox.2014.09.006.
[37] O. Ostling and K. J. Johanson, 'Microelectrophoretic study of radiation-induced DNA damages in individual mammalian cells', *Biochem Biophys Res Commun*, vol. 123, no. 1, pp. 291 – 298, 1984, doi: 10.1016/0006-291X(84)90411-X.
[38] P. E. Verd and A. Rottmann, 'Statistical models to analyze genotoxicological experiments with the comet assay data', *Ann Biom Biostat*, vol. 2, no. 3, p. 1025, 2015.

[39] J. Bright *et al.*, 'Recommendations on the statistical analysis of the Comet assay', *Pharm Stat*, vol. 10, no. 6, pp. 485–493, Nov. 2011, doi: 10.1002/pst.530.
[40] 'Skewness and Kurtosis', in *Statistics for Scientists and Engineers*, John Wiley & Sons, Ltd, 2015, ch. 4, pp. 89–110. doi: https://doi.org/10.1002/9781119047063.ch4.
[41] W. W. Stroup, G. A. Milliken, E. A. Claassen, and R. D. W. Wolfinger, *SAS for Mixed Models: Introduction and Basic Applications*. SAS Institute, 2018. Accessed: Jun. 03, 2025. [Online]. Available: https://books.google.de/books?id=e1rPvwEACAAJ
[42] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*, Reprint. New York: Springer Verlag, 2009. doi: https://doi.org/10.1007/978-1-4419-0300-6.
[43] F. Bretz and L. A. Hothorn, 'Testing dose-response relationships with a priori unknown, possibly nonmonotone shapes', *J Biopharm Stat*, vol. 11, no. 3, pp. 193 – 207, 2001, doi: 10.1081/BIP-100107657.
[44] L. A. Hothorn, *Statistics in Toxicology Using R*. Taylor & Francis Group, LLC, 2016. doi: https://doi.org/10.1201/b19659.
[45] L. A. Hothorn, 'A robust statistical procedure for evaluating genotoxicity data', *Environmetrics*, vol. 15, no. 6, pp. 635 – 641, 2004, doi: 10.1002/env.649.
[46] F. Bretz and L. A. Hothorn, 'Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays', *Alternatives to Laboratory Animals*, vol. 31, no. SUPPL. 1, pp. 81 – 96, 2003, doi: 10.1177/026119290303101s06.
[47] S. J. Wiklund and E. Agurell, 'Aspects of design and statistical analysis in the Comet assay', *Mutagenesis*, vol. 18, no. 2, pp. 167–175, 2003, doi: https://doi.org/10.1093/mutage/18.2.167.
[48] T. Tug, K. Ickstadt, M. Kunz, A. Sutter, and B.-W. Igl, 'Statistical analysis of in vivo alkaline comet assay data - Comparison of median and geometric mean as centrality measures', *Regulatory Toxicology and Pharmacology*, vol. 118, p. 104808, Dec. 2020, doi: 10.1016/j.yrtph.2020.104808.
[49] Alzheimer`s Association, '2022 Alzheimer's disease facts and figures', *Alzheimer's & Dementia*, vol. 18, no. 4, pp. 700–789, Apr. 2022, doi: https://doi.org/10.1002/alz.12638.
[50] J. M. Long and D. M. Holtzman, 'Alzheimer Disease: An Update on Pathobiology and Treatment Strategies', *Cell*, vol. 179, no. 2, pp. 312–339, Oct. 2019, doi: 10.1016/j.cell.2019.09.001.
[51] D. J. Selkoe, 'Alzheimer's Disease Is a Synaptic Failure', *Science (1979)*, vol. 298, no. 5594, pp. 789–791, Oct. 2002, doi: 10.1126/science.1074069.
[52] H. Mathys *et al.*, 'Single-cell transcriptomic analysis of Alzheimer's disease', *Nature*, vol. 570, no. 7761, pp. 332–337, Jun. 2019, doi: 10.1038/s41586-019-1195-2.
[53] X. Dai and L. Shen, 'Advances and Trends in Omics Technology Development', *Front Med (Lausanne)*, vol. 9, Jul. 2022, doi: 10.3389/fmed.2022.911861.
[54] J. M. Wilkins and E. Trushina, 'Application of metabolomics in Alzheimer's disease', *Front Neurol*, vol. 8, no. JAN, Jan. 2018, doi: 10.3389/fneur.2017.00719.
[55] Y. Yuan, G. Zhao, and Y. Zhao, 'Dysregulation of energy metabolism in Alzheimer's disease', *J Neurol*, vol. 272, no. 1, p. 2, Jan. 2025, doi: 10.1007/s00415-024-12800-8.
[56] A. J. H. Kind and W. R. Buckingham, 'Making Neighborhood-Disadvantage Metrics Accessible — The Neighborhood Atlas', *New England Journal of Medicine*, vol. 378, no. 26, pp. 2456–2458, Jun. 2018, doi: 10.1056/NEJMp1802313.
[57] A. Hüls, U. Krämer, C. Carlsten, T. Schikowski, K. Ickstadt, and H. Schwender, 'Comparison of weighting approaches for genetic risk scores in gene-environment interaction studies', *BMC Genet*, vol. 18, no. 1, Dec. 2017, doi: 10.1186/s12863-017-0586-3.
[58] A. Hüls, K. Ickstadt, T. Schikowski, and U. Krämer, 'Detection of gene-environment interactions in the presence of linkage disequilibrium and noise by using genetic risk scores with internal weights from elastic net regression', *BMC Genet*, vol. 18, no. 1, Jun. 2017, doi: 10.1186/s12863-017-0519-1.
[59] J. Chen *et al.*, 'Pruning and thresholding approach for methylation risk scores in multi-ancestry populations', *Epigenetics*, vol. 18, no. 1, Dec. 2023, doi: 10.1080/15592294.2023.2187172.

[60] M. J. Wurm, P. J. Rathouz, and B. M. Hanlon, 'Regularized Ordinal Regression and the ordinalNet R Package', *J Stat Softw*, vol. 99, no. 6, pp. 1–42, 2021, doi: 10.18637/jss.v099.i06.
[61] M. Lau, T. Schikowski, and H. Schwender, 'logicDT: a procedure for identifying response-associated interactions between binary predictors', *Mach Learn*, vol. 113, no. 2, pp. 933–992, Feb. 2024, doi: 10.1007/s10994-023-06488-6.
[62] M. N. Wright and A. Ziegler, 'ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R', *J Stat Softw*, vol. 77, no. 1, pp. 1–17, 2017, doi: 10.18637/jss.v077.i01.
[63] S. Teschke, K. Ickstadt, and A. Munteanu, 'Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores', *Biometrical Journal*, vol. 66, no. 8, Dec. 2024, doi: 10.1002/bimj.70014.
[64] M. Mayer, 'missRanger: Fast Imputation of Missing Values', 2024. doi: https://doi.org/10.32614/CRAN.package.missRanger.
[65] E. G. Armitage, J. Godzien, V. Alonso-Herranz, Á. López-Gonzálvez, and C. Barbas, 'Missing value imputation strategies for metabolomics data', *Electrophoresis*, vol. 36, no. 24, pp. 3050–3060, Dec. 2015, doi: 10.1002/elps.201500352.
[66] D. J. Stekhoven and P. Bühlmann, 'MissForest—non-parametric missing value imputation for mixed-type data', *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.
[67] M. Kokla, J. Virtanen, M. Kolehmainen, J. Paananen, and K. Hanhineva, 'Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study', *BMC Bioinformatics*, vol. 20, no. 1, p. 492, Dec. 2019, doi: 10.1186/s12859-019-3110-0.
[68] D. McFadden, 'Conditional logit analysis of qualitative choice behavior', in *Zarembka, P., Ed., Frontiers in Econometrics*, no. 105–142, New York: Academic Press, 1974, ch. 4, pp. 105–142.
[69] B. Phipson, J. Maksimovic, and A. Oshlack, 'missMethyl: an R package for analysing methylation data from Illuminas HumanMethylation450 platform', *Bioinformatics*, vol. 32, no. 2, pp. 286–288, Jan. 2016, doi: https://doi.org/10.1093/bioinformatics/btv560.
[70] L. Tian and T. Yu, 'metapone: Conducts pathway test of metabolomics data using a weighted permutation test', 2024, doi: 10.18129/B9.bioc.metapone.
[71] J. Pessoa *et al.*, 'Environmental Risk Factors Influence the Frequency of Coughing and Sneezing Episodes in Finisher Pigs on a Farm Free of Respiratory Disease', *Animals*, vol. 12, no. 8, p. 982, Apr. 2022, doi: 10.3390/ani12080982.
[72] D. B. Rubin, 'Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations', *Journal of Business & Economic Statistics*, vol. 4, no. 1, p. 87, Jan. 1986, doi: 10.2307/1391390.
[73] S. Grund, O. Lüdtke, and A. Robitzsch, 'Handling Missing Data in Cross-Classified Multilevel Analyses: An Evaluation of Different Multiple Imputation Approaches', *Journal of Educational and Behavioral Statistics*, vol. 48, no. 4, pp. 454–489, Aug. 2023, doi: 10.3102/10769986231151224.
[74] S. van Buuren and K. Groothuis-Oudshoorn, '**mice** : Multivariate Imputation by Chained Equations in *R*', *J Stat Softw*, vol. 45, no. 3, 2011, doi: 10.18637/jss.v045.i03.
[75] I. R. White, R. Daniel, and P. Royston, 'Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables', *Comput Stat Data Anal*, vol. 54, no. 10, pp. 2267–2275, Oct. 2010, doi: 10.1016/j.csda.2010.04.005.
[76] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.
[77] P. C. Austin, 'Estimating Multilevel Logistic Regression Models When the Number of Clusters is Low: A Comparison of Different Statistical Software Procedures', *Int J Biostat*, vol. 6, no. 1, Jan. 2010, doi: 10.2202/1557-4679.1195.
[78] P.-C. Bürkner, 'brms: An R Package for Bayesian Multilevel Models Using Stan', *J Stat Softw*, vol. 80, no. 1, pp. 1–28, 2017, doi: 10.18637/jss.v080.i01.
[79] P.-C. Bürkner, 'Advanced Bayesian Multilevel Modeling with the R Package brms', *R J*,

vol. 10, no. 1, pp. 395–411, 2018, doi: 10.32614/RJ-2018-017.
[80] P.-C. Bürkner, 'Bayesian Item Response Modeling in R with brms and Stan', *J Stat Softw*, vol. 100, no. 5, pp. 1–54, 2021, doi: 10.18637/jss.v100.i05.
[81] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman, 'Visualization in Bayesian workflow', *J. R. Stat. Soc. A*, vol. 182, no. 2, pp. 389–402, 2019, doi: 10.1111/rssa.12378.
[82] R Core Team, 'R: A Language and Environment for Statistical Computing', 2021, *Vienna, Austria*. [Online]. Available: https://www.R-project.org/
[83] M. Kay, 'tidybayes: Tidy Data and Geoms for Bayesian Models', 2024. doi: 10.5281/zenodo.1308151.
[84] D. Lüdecke, M. S. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski, 'performance: An R Package for Assessment, Comparison and Testing of Statistical Models', *J Open Source Softw*, vol. 6, no. 60, p. 3139, 2021, doi: 10.21105/joss.03139.
[85] N. R. W. Cleland, S. I. Al-Juboori, E. Dobrinskikh, and K. D. Bruce, 'Altered substrate metabolism in neurodegenerative disease: new insights from metabolic imaging', *J Neuroinflammation*, vol. 18, no. 1, p. 248, Oct. 2021, doi: 10.1186/s12974-021-02305-w.
[86] I. Korczowska-Łącka *et al.*, 'Selected Biomarkers of Oxidative Stress and Energy Metabolism Disorders in Neurological Diseases', *Mol Neurobiol*, vol. 60, no. 7, pp. 4132–4149, Jul. 2023, doi: 10.1007/s12035-023-03329-4.
[87] D. A. Butterfield *et al.*, 'Redox Proteomics in Selected Neurodegenerative Disorders: From its Infancy to Future Applications', *Antioxid Redox Signal*, vol. 17, no. 11, pp. 1610–1655, Dec. 2012, doi: 10.1089/ars.2011.4109.
[88] F. Yin, 'Lipid metabolism and Alzheimer's disease: clinical evidence, mechanistic link and therapeutic promise', *FEBS J*, vol. 290, no. 6, pp. 1420–1453, Mar. 2023, doi: 10.1111/febs.16344.
[89] N. Fabelo, V. Martín, R. Marín, D. Moreno, I. Ferrer, and M. Díaz, 'Altered lipid composition in cortical lipid rafts occurs at early stages of sporadic Alzheimer's disease and facilitates APP/BACE1 interactions', *Neurobiol Aging*, vol. 35, no. 8, pp. 1801–1812, Aug. 2014, doi: 10.1016/j.neurobiolaging.2014.02.005.
[90] J.-L. Olivier, 'Arachidonic acid in Alzheimer's disease', *J Neurol Neuromedicine*, vol. 1, no. 9, pp. 1–6, Dec. 2016, doi: 10.29245/2572.942X/2016/9.1086.
[91] M. Grimm *et al.*, 'Effect of Different Phospholipids on α-Secretase Activity in the Non-Amyloidogenic Pathway of Alzheimer's Disease', *Int J Mol Sci*, vol. 14, no. 3, pp. 5879–5898, Mar. 2013, doi: 10.3390/ijms14035879.
[92] S. Kumari, R. Dhapola, and D. H. Reddy, 'Apoptosis in Alzheimer's disease: insight into the signaling pathways and therapeutic avenues', *Apoptosis*, vol. 28, no. 7–8, pp. 943–957, Aug. 2023, doi: 10.1007/s10495-023-01848-y.
[93] A. J. Tenner, 'Complement-Mediated Events in Alzheimer's Disease: Mechanisms and Potential Therapeutic Targets', *The Journal of Immunology*, vol. 204, no. 2, pp. 306–315, Jan. 2020, doi: 10.4049/jimmunol.1901068.
[94] J. R. Clarke, F. C. Ribeiro, R. L. Frozza, F. G. De Felice, and M. V. Lourenco, 'Metabolic Dysfunction in Alzheimer's Disease: From Basic Neurobiology to Clinical Approaches', *Journal of Alzheimer's Disease*, vol. 64, no. s1, pp. S405–S426, Jun. 2018, doi: 10.3233/JAD-179911.
[95] M. J. Kan *et al.*, 'Arginine Deprivation and Immune Suppression in a Mouse Model of Alzheimer's Disease', *The Journal of Neuroscience*, vol. 35, no. 15, pp. 5969–5982, Apr. 2015, doi: 10.1523/JNEUROSCI.4668-14.2015.
[96] C. J. Pike, J. C. Carroll, E. R. Rosario, and A. M. Barron, 'Protective actions of sex steroid hormones in Alzheimer's disease', *Front Neuroendocrinol*, vol. 30, no. 2, pp. 239–258, Jul. 2009, doi: 10.1016/j.yfrne.2009.04.015.
[97] T.-Y. Chang, Y. Yamauchi, M. T. Hasan, and C. Chang, 'Cellular cholesterol homeostasis and Alzheimer's disease', *J Lipid Res*, vol. 58, no. 12, pp. 2239–2254, Dec. 2017, doi: 10.1194/jlr.R075630.
[98] T. Tug *et al.*, 'In vivo alkaline comet assay: Statistical considerations on historical negative and positive control data', *Regulatory Toxicology and Pharmacology*, vol. 148, p. 105583,

Mar. 2024, doi: 10.1016/j.yrtph.2024.105583.

# Part II

# Publications

# *Article 1*

Regulatory Toxicology and Pharmacology 118 (2020) 104808

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

# Statistical analysis of in vivo alkaline comet assay data - Comparison of median and geometric mean as centrality measures

Timur Tug [a,*], Katja Ickstadt [a], Michael Kunz [b], Andreas Sutter [c], Bernd-Wolfgang Igl [d]

[a] Faculty of Statistics, TU Dortmund University, 44221, Dortmund, Germany
[b] Bayer AG, Research & Early Development Statistics, 13342, Berlin, Germany
[c] Bayer AG, Pharmaceuticals, Investigational Toxicology, Müllerstr. 178, 13353 Berlin / Bayer AG, Oncology Development - Program Management, 13342, Berlin, Germany
[d] Bayer AG, Research and Clinical Sciences Statistics, 13342 Berlin, Germany / Boehringer Ingelheim Pharma GmbH & Co. KG, Biostatistics + Data Sciences Corp., 88397, Biberach an der Riss, Germany

ARTICLE INFO

*Keywords:*
Comet assay
Statistical analysis
Linear model
Down-turn protected trend test
Tail intensity
Slide summary

ABSTRACT

The comet assay is one of the standard tests for evaluating the genotoxic potential of a test item able to detect DNA strand breaks in cells or isolated nuclei from various tissues. The in vivo alkaline comet assay is part of the standard test battery, given in option 2 of the ICH guidance S2 (R1) and a follow-up test in the EFSA framework on genotoxicity testing. The current OECD guideline for the testing of chemicals No. 489 directly affects the statistical analysis of comet data as it suggests using the median per slide and the mean of all medians per animal. However, alternative approaches can be used if scientifically justified. In this work, we demonstrated that the selection of different centrality measures to describe an average value per slide may lead to fundamentally different statistical test results and contradicting interpretations. Our focus was on geometric means and medians per slide for the primary endpoint "tail intensity". We compared both strategies using original and simulated data in different experimental settings incl. a varying number of animals, slides and cells per slide. In general, it turned out that the chosen centrality measure has an immense impact on the final statistical test result.

## 1. Introduction

The comet assay (also single cell gel electrophoresis or microgel electrophoresis) is a technique for the direct visualization of DNA strand breaks in individual cells and was first developed by Östling and Johanson in 1984 (Östling and Johanson, 1984). The comet assay in its alkaline version is a common genotoxicity test for detecting DNA single and double strand breaks resulting, for example, from direct interactions with DNA, alkali labile sites or as a consequence of transient DNA strand breaks resulting from DNA excision repair. It is emerging as the standard method for testing DNA damage (Wiklund and Agurell, 2003), as it is cheap, fast, easy and more sensitive than the unscheduled DNA synthesis (UDS) test for detecting carcinogens (Kirkland and Speit, 2008). Moreover, it is theoretically applicable to any tissue as it is not dependent on cell proliferation. For more details on the test procedures, see the literature (Collins, 2004; Singh et al., 1988; Fairbairn et al., 1995; Tice et al., 2000; Burlinson et al., 2007; Guérard et al., 2014).

In this work, effects of different summarizing strategies on the results of the comet assay were analysed. A related question has already been investigated for the parameters tail length and tail moment (see e.g. Wiklund and Agurell, 2003), but only in very rare cases for the tail intensity (see e.g. Duez, 2003). This parameter was studied here, since it has become increasingly important in comet assay analyses in recent years (Burlinson et al., 2007; Bright et al., 2011) and was consequently recommended as the primary parameter in OECD guideline TG 489. The guideline further suggests using the median tail intensity per slide and the arithmetic mean of all medians per animal (see section 2.3. "Statistical strategies"). However, alternative approaches can be used if scientifically justified.

This work aims to intensify the discussion about the proper choice of an adequate summary measure per slide, i.e. about which centrality measure provides the most suitable characterization of the treatment

*Abbreviations:* EFSA, European Food Safety Authority; **OECD**, Organisation for Economic Co-operation and Development; **ICH**, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; **SAS**, Statistical Analysis System.

* Corresponding author.
*E-mail address:* tug@statistik.tu-dortmund.de (T. Tug).

https://doi.org/10.1016/j.yrtph.2020.104808
Received 14 June 2020; Received in revised form 18 September 2020; Accepted 23 October 2020
Available online 28 October 2020

investigated (OECD, 2016).

Therefore, we studied a relatively small data set of five different experiments using the tail intensity ("% tail DNA") as the primary parameter and extended our considerations inspired, in particular, by Bright et al. and Wiklund & Agurell (Bright et al., 2011; Wiklund and Agurell, 2003). For each experiment, the differences between vehicle and dose groups were analysed with the aid of different centrality measures (median and geometric mean per slide and the arithmetic mean of these medians and geometric means per animal) (see section 2.3. "Statistical strategies"). In general, the fundamental statistical test results should not differ, i.e. they should not depend on the chosen centrality measure. However, sometimes even small differences between different location parameters can translate into big differences in terms of the number of significant results and, therefore, may lead to a completely different interpretation of results.

Results of the real data analysis were complemented by simulations. Based on a simple linear model for tail intensity with a fixed treatment effect and with suitable distributional assumptions, different numbers of animals, slides and cells per slide were simulated, and the statistical power of comparisons to control was calculated (see section 2.5. "Method for evaluating the power curve") and visualized using power curves.

In order to emphasize the effect of different slide summaries on the overall result of an experiment a motivating example was given in the subsequent section 2.1. involving the geometric mean, arithmetic mean and median per slide calculated on one and the same data set.

## 2. Materials and methods

### 2.1. Motivating example

An initial example will describe the possible effect of different slide summaries on the final test result. To this end, let us consider an artificial experiment based on simulated data consisting of four treatment groups in total, i.e. a vehicle and three dose groups of the test item, 6 animals per group and 3 slides with 50 cells each. The primary parameter is the tail intensity per cell and corresponding observations were firstly summarized using a median, arithmetic mean and geometric mean value per slide and secondly slide summaries were averaged per animal (see Fig. 1).

It became obvious that dose-response shapes differ extremely between different slide summaries and this concerns i) the form of such relationships themselves and ii) the general magnitude of calculated values per animal. Consequently, estimated treatment effects and their statistical significance differed substantially depending on whether e.g. a median, a geometric mean or an arithmetic mean was chosen to summarize the observed tail intensities per slide.

Here we exemplarily considered only one arbitrary data set. That means, the graphical representations, effects, statistical significances and interpretations given in Fig. 1 belong to one and the same data set and differences were solely related to the chosen summarizing measures per slide.

### 2.2. Experimental design

The present data were generated in experiments with a typical design of five different groups, i.e. a vehicle group, a positive control group and three different dose groups (low, medium and high), where each group consisted of six animals. The experimental units were male rats with the liver as the target organ. For each animal, three slides with 50 liver cells each were prepared (OECD, 2016) in order to determine the primary parameter "tail intensity" ("% tail DNA") per cell.

The used data set originated from five experiments carried out by Bayer AG and were made available for investigation. Our experiments were performed according to the OECD guideline for the testing of chemicals No. 489 (In Vivo Alkaline Comet Assay; July 29, 2016) (OECD, 2016). The experimental data were used descriptively and mainly served as a starting point for specifying the simulation framework. Then, we varied basic ingredients and used 6 or 8 animals; 3, 4 or 5 slides and 30, 50 or 70 cells per slide in our simulation studies. Note that Wiklund & Agurell (Wiklund and Agurell, 2003) have used similar simulation scenarios for analyzing "tail moment" and "tail length" as primary parameters.

### 2.3. Statistical strategies

The OECD guideline TG 489 suggests using the median per slide for the typically right skewed non-negative data of the comet assay and the

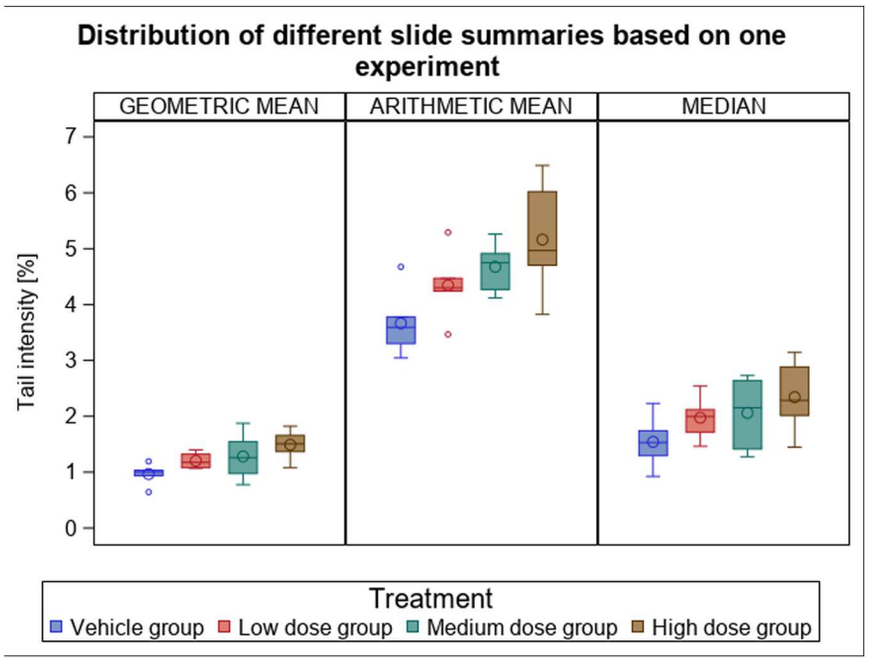



**Fig. 1.** Tail intensities of different treatment groups for the motivating example.

arithmetic mean of all slide medians per animal. The median is an extremely robust measure of centrality reflected by its breakdown point of 50%, which means it can tolerate of up to 50% of the data been shifted to extreme values before this has an impact on the median (Bright et al., 2011; Verd and Rottmann, 2015). But, and for the same reason, the median wastes a lot of information just because it uses only one (odd no. Of measurements) or two cell measurements (even no. Of measurements). However, a sensible alternative to describe right skewed data, especially, is to calculate the geometric mean (see also Shanmugam and Chattamvelli, 2015). The geometric mean is equal to the exponential of the arithmetic mean of the logarithmic transformed data, which is one common transformation for right skewed distributions (Wiklund and Agurell, 2003; OECD, 2016). In other words, the geometric mean is defined as the *n*th root of the product of all *n* positive numbers and is particularly useful for the description of a central tendency of relative numbers like growth rates and others.

Thus, in this work, both centrality measures (median and geometric mean) were of primary interest to summarize tail intensity data per slide. Note that the geometric mean puts emphasis on the tails of the distribution of the data investigated, while this is not the case for the median.

In principle, measurements on a single cell level must be summarized at least per slide; otherwise, the statistical analysis might lead to erroneous results based on autocorrelation effects (see Lovell and Omori, 2008 for more details).We used a simple 1-way linear model for the tail intensity with treatment as a single fixed effect to compare treatment and vehicle measurements. We investigated two different settings:

1) In a so-called "statistical slide model" single cell measurements per slide were summarized using a median or geometric mean yielding one value per slide. Then, $s$ (e.g. $s = 3$) slides per animal led to $s$ "repeated" values per animal which can be analysed statistically using a repeated measurements model incl. e.g. a variance component covariance structure involving uncorrelated slides.
2) A "statistical animal model" was based on the slide summaries mentioned before (see 1)), but now, these values were averaged again using a simple arithmetic mean yielding one value per animal.

In general, pairwise comparisons of dose and control data were computed using a Satterthwaite modification of a one-sided *t*-test (Stroup et al., 2018) with $\alpha = 0.05$. Therein, the hypothesis $H_0 : \mu_v = \mu_j$ vs. the alternative $H_1 : \mu_v < \mu_j$ was tested, where $v$ represented the vehicle (control) group and $j$ the respective (low, medium and high) dose group ($j \in \{l,m,h\}$). For more details and a theoretical background on the linear (mixed) model, see e.g. Verbeke and Molenberghs (2009).

Particularly in toxicology non-monotonic dose-response shapes occur, sometimes based on a downturn effect at high doses (Hothorn, 2016). Therefore, a down-turn-protected trend test as introduced by Bretz & Hothorn (Bretz and Hothorn, 2001; Bretz and Hothorn, 2008; Hothorn, 2004) was applied. It analyses different dose-response relationships and specifies the most plausible one using a minimum-p-value approach. In Appendix Table C.1, all possible shapes including a down-turn effect for three dose and one vehicle group were specified that we included in our analysis.

All calculations were done with SAS® 9.4 (TS1M6).

### 2.4. Statistical aspects for the simulation study

The simulation framework was related to a standard experimental design and parameter values were based on our present real data. More precisely, we considered a setting of a vehicle group and three dose levels of a test item (low, medium, high). Each dose group consisted of a certain (predefined) number of animals (6 or 8), slides (3, 4, or 5) and cells (30, 50, or 70). Initially, we used vehicle observations of the present real data and took the average of their means and standard deviations in order to specify vehicle parameters for the simulation procedure (see Fig. 2 and Appendix Table C.2).

Moreover, we assumed a constant coefficient of variation in each treatment group and simulated artificial treatment data. To this end, we multiplied the control mean with a factor equal to 1.3 to mimic low dose observations, with 1.5 to obtain medium dose data and with 1.8 to simulate high dose data.

The distribution of the present experimental vehicle control tail intensities was highly right-skewed (see Fig. 3) which coincided with findings in the literature (see e.g. Bright et al., 2011; Verd and Rottmann, 2015; Lovell et al., 1999). However, our database was limited and, therefore, we refrained from fitting an "optimal" distribution to the empirical data. We rather specified a flexible right skewed distribution for further statistical analyses. To this end, we used a gamma distribution to describe tail intensities per slide and estimate corresponding parameters based on the existing data set (see e.g. Shanmugam and Chattamvelli, 2015). Information on the relationship between medians and geometric means of gamma distributed data was briefly given in Appendix A. There are further alternatives, see Appendix B, for using a lognormal distribution.

According to our experience, we observed up to approximately 10% "extreme" values per slide. Accordingly, we chose a random integer $q$ between 0 and 10% of the number of analysed cells per slide and simulated "*number of cells* per *slide* - $q$" values from the gamma distribution described above. The remaining $q$ values were drawn from the same type of distribution with the same mean, but with a five times increased variance plus an offset of 10.

We than used this mixture of a "regular" and an additional "noise" part of gamma distributions to describe tail intensities on the cell level in an adequate and realistic way.

Moreover, we focused on four different shapes denoted by contrasts 1 to 4 (see Appendix Table C.1). Each simulation scenario was subsequently carried out 10,000 times.

The whole simulation process was summarized in Fig. 2.

### 2.5. Method for evaluating the power curve

An important aim of a comet assay is to answer the question whether the data indicate an increase of %tail intensities with increasing dose of the test item or not. This goal should be translated into the statistical hypotheses to be tested and the corresponding statistical power, i.e. the probability of accepting the alternative hypothesis $H_1$, when $H_1$ is true, should be large (e.g. 80% or higher). Therefore it might be reasonable to compare different strategies with the aid their statistical power. To this end, power curves were used to illustrate our findings based on a comparison of two independent samples (one dose group vs. vehicle group). We varied the treatment effect in the dose group by increasing the average dose (ranging from the average to approximately twice the average) and determined the statistical power of a comparison to its corresponding vehicle control. For all different experimental settings considered, we also varied the number of animals, slides and cells per slide. Corresponding power curves were visualized based on a penalised B-spline regression for smoothing (see Dierckx, 1993, for more details with regard to B-Spline regression, implementable, e.g., in SAS using PROC SGPANEL involving the PBSPLINE statement).

## 3. Results

### 3.1. Experimental data

The distribution of observed vehicle tail intensities of one arbitrarily chosen experiment is shown in Fig. 3. Three slides for each of the six vehicle animals are displayed next to each other, leading to 18 boxplots in the upper part of the figure. Moreover, data observed on one slide are illustrated exemplarily in more detail in the lower part of the plot from one arbitrarily chosen slide (here: slide 1 from animal 1). In all cases the observations did not appear to be normally distributed, but were highly

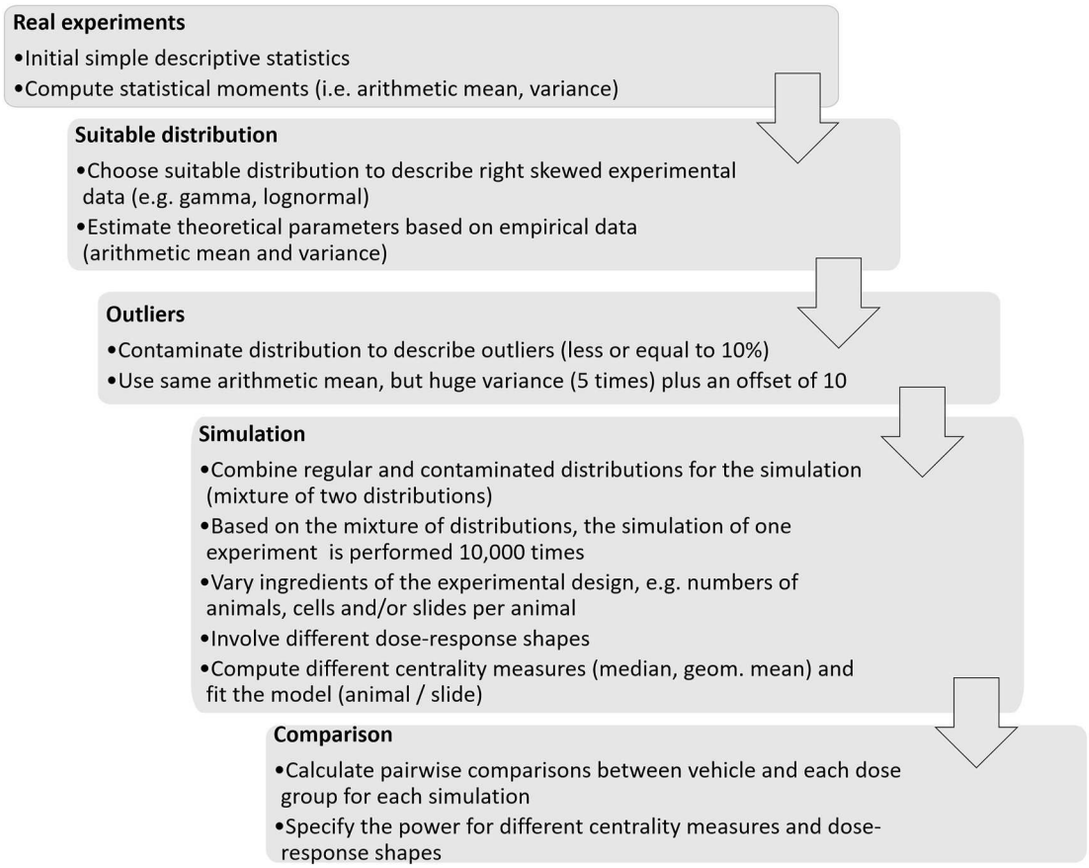



**Fig. 2.** Schematic description of the simulation process.

right-skewed. This can be explained by a majority of very small values stemming from viable cells and a minority of very high values (i.e., cells with fragmented DNA, for technical or cell-biological reasons). As a consequence, the median was clearly smaller than e.g. the usual arithmetic mean value; this phenomenon can usually be seen for vehicle as well as for dose groups.

In general, the difference between different centrality measures (e.g. arithmetic mean, geometric mean and median) increased with increasing skewness and in addition, skewness was usually dose dependent.

The current OECD guideline TG 489 (OECD, 2016) recommends using the median per slide and their arithmetic mean per animal. Obviously different centrality measures per slide led to a different description of observations and thus, these differences increased with an increasing skewness of the underlying data distribution, or in other words, with increasing dose. Content wise this required an assessment whether large or seemingly extreme tail intensity values contained important information or not. We expected that this question cannot be answered in a general way, but might also depend on the substance to be analysed.

### 3.2. Simulations

The empirical means and standard deviations given in Appendix Table C.2 were used to estimate corresponding theoretical parameters of the gamma distribution for the simulation procedure. Afterwards, dose group observations were simulated involving a predefined constant multiplied with the vehicle mean under assumption of a constant coefficient of variation. The constant depended on the dose response relationship, was larger than, equal to 1 and less than, or equal to 1.8 (high dose group) (see section 2.4. "Statistical aspects for the simulation study" and the overview given in Fig. 2).

Moreover, we varied the number of animals, slides and the number of cells in the four previously mentioned test settings and determined 10,000 iterations for each scenario. Statistical testing strategies were based on pairwise comparisons to control and on down-turn protected trend tests. In general, the statistical power is estimated by counting the number of rejected hypotheses tests.

As expected, all summary measures held the significance level of 5% fairly well under $H_0$.

Fig. 4 illustrates the statistical power [%] of various pairwise comparisons of treatment versus control groups. We varied the number of animals (6 or 8), slides (3, 4 or 5) and cells (30, 50 or 70) and compared a "low", a "medium" and a "high" dose group with vehicle measurements using a median or geometric slide summary in a "statistical animal model" (single measurement) or "statistical slide model" (repeated measurements) (see section 2.3. "Statistical strategies").

Note that Fig. 4 displays different simulation scenarios involving only 6 animals, 3 and 5 slides, 50 and 70 cells and two dose groups. An extended version of this graphic can be found in the Appendix (see Figure B2).

It was obvious, that the statistical power increased with an increasing number of animals, slides and/or cells (see Appendix Figure B2). This could be observed for all centrality measures analysed in this work.

However, it is interesting to point out that there are notable differences between the different models and slide measures. In general, the power of the geometric mean used in the "statistical animal model" as well as in the "statistical slide model" seemed to be (substantially) lower than of the median. This could be observed for all experimental settings independently of the strength of the treatment effect.

In addition to the analysis of pairwise comparisons of vehicle and dose groups, simulations for all down-turn-protected trend tests were

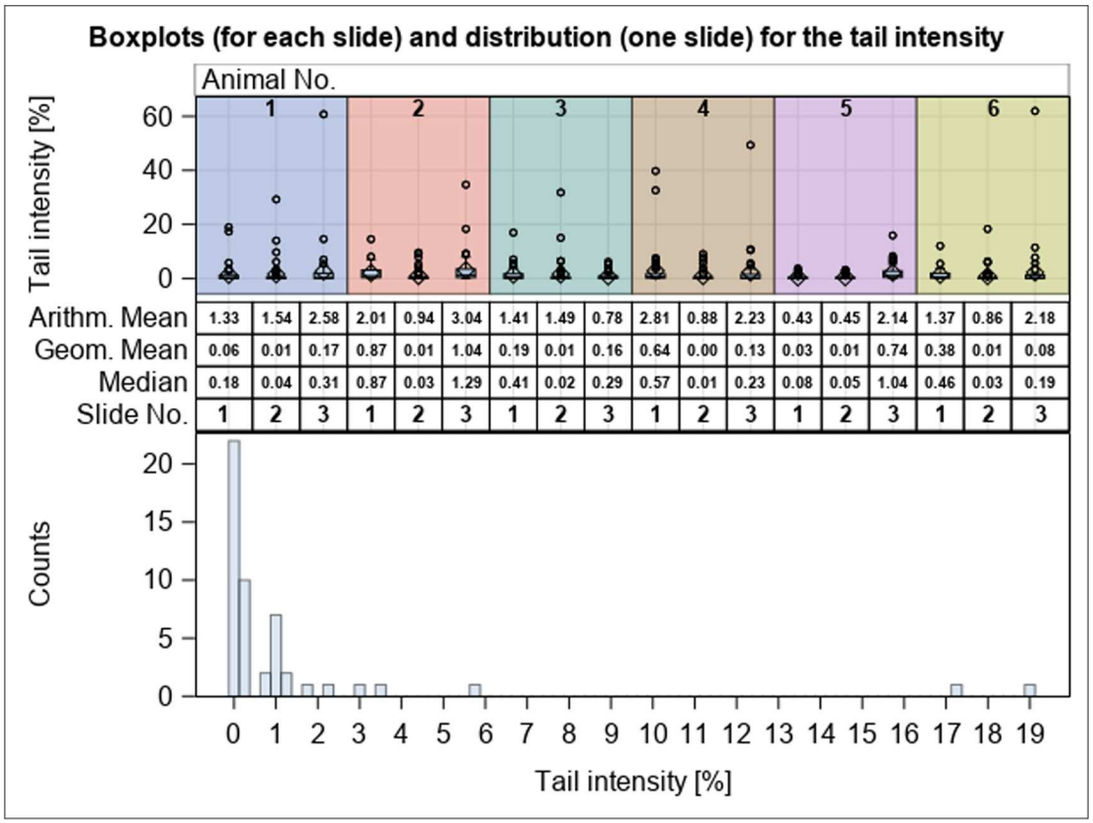



**Fig. 3.** Vehicle control measurements of one experimental data set. TOP: Boxplots and summary measures for vehicle control tail intensities per animal and slide (6 animals with 3 slides each). BOTTOM: Histogram of observed tail intensities of one slide of one arbitrarily chosen animal (here: slide 1 from animal 1).

performed (see Appendix Table C.1). Here, the significance level of 5% was well maintained in the scenario under $H_0$. In both models, i.e. the "statistical animal model" and the "statistical slide model", the power of the median was also greater than of the geometric mean. Obviously, the statements about both centrality measures were very similar to the simulated pairwise comparisons (see Appendix Table C3).

### 3.3. Power curve analysis

Additionally, power curves were studied for different centrality measures and simulation settings, i.e. the number of animals, slides and cells per slide. Fig. 5 (LEFT) was based on the mean of slide-medians of six animals using a varying number of slides and cells. Obviously, the power increased with an increasing number of slides and cells. However, differences in statistical power between different experimental settings were remarkable. For example, based on the underlying simulation framework, we detected an average increase of 50% between dose and vehicle data with a power of approx. 60%, if we took 3 slides with 50 cells each, but with a power of more than 95%, if we took 5 slides with 70 cells. In other words, using 6 animals, 5 slides and 70 cells per slide, we were able to detect an average increase of roughly 30% with an 80% power. If we took only 3 slides and 50 cells per slide, we detected an average increase of at least 65% with an 80% power.

Obviously, the effect of an increasing number of slides, cells and animals on the power to detect differences between dose and vehicle groups was independent of the chosen centrality measures for the data per slide. However, if both centrality measures were compared, substantial differences could be observed (see e.g. Fig. 5 (RIGHT)). In our simulations using gamma distributed data, the application of the geometric mean always led to a lower power compared to a similar framework using median values per slide. This fundamental behaviour could be observed for the "statistical animal model" as well as for the "statistical slide model". For lognormally distributed measurements, the corresponding figures are given in Appendix B Figure B1.

## 4. Discussion

Some of the pioneering works investigating the most appropriate statistical approach for analysing comet assay data focused on "tail length" and/or "tail moment" as the primary parameter of interest (see e.g. Wiklund and Agurell, 2003; Lovell et al., 1999). In this regard, Wiklund & Agurell (Wiklund and Agurell, 2003) used different models and statistical tests and finally proposed to compute the mean value of log-transformed tail moments (or equivalently the log of the geometric mean) as a suitable centrality measurement per slide. In addition, they suggested using an "upper" percentile (e.g. the 90th percentile) of the log distribution to describe tail length measurements per slide. A further relevant publication deals with the statistical analysis of % tail DNA measurements in mice (Hansen et al., 2014) and propose to use the median of the log-transformed data to summarize the within-sample distribution most appropriately. Duez et al. (2003) analysed the %tail DNA and Olive tail moment and demonstrated that the median might be a sensible non-parametric tool to compare samples and an efficient way to statistically demonstrate a genotoxic effect. The loss of information which goes along with an application of these non-parametric techniques might be compensated with an additional visualization of data, e.g. with the aid of boxplots (Duez et al., 2003).

However, the current guideline (OECD guideline TG 489 (OECD, 2016)) recommends tail intensity as the primary parameter to be used for genotoxicity assessment (Burlinson et al., 2007; OECD, 2016), and therefore, plays the central role in this work.

Our experimental data (vehicle control data from Bayer AG) revealed highly right-skewed distributions on the cell level, such that the arithmetic mean, geometric mean and median differed substantially per

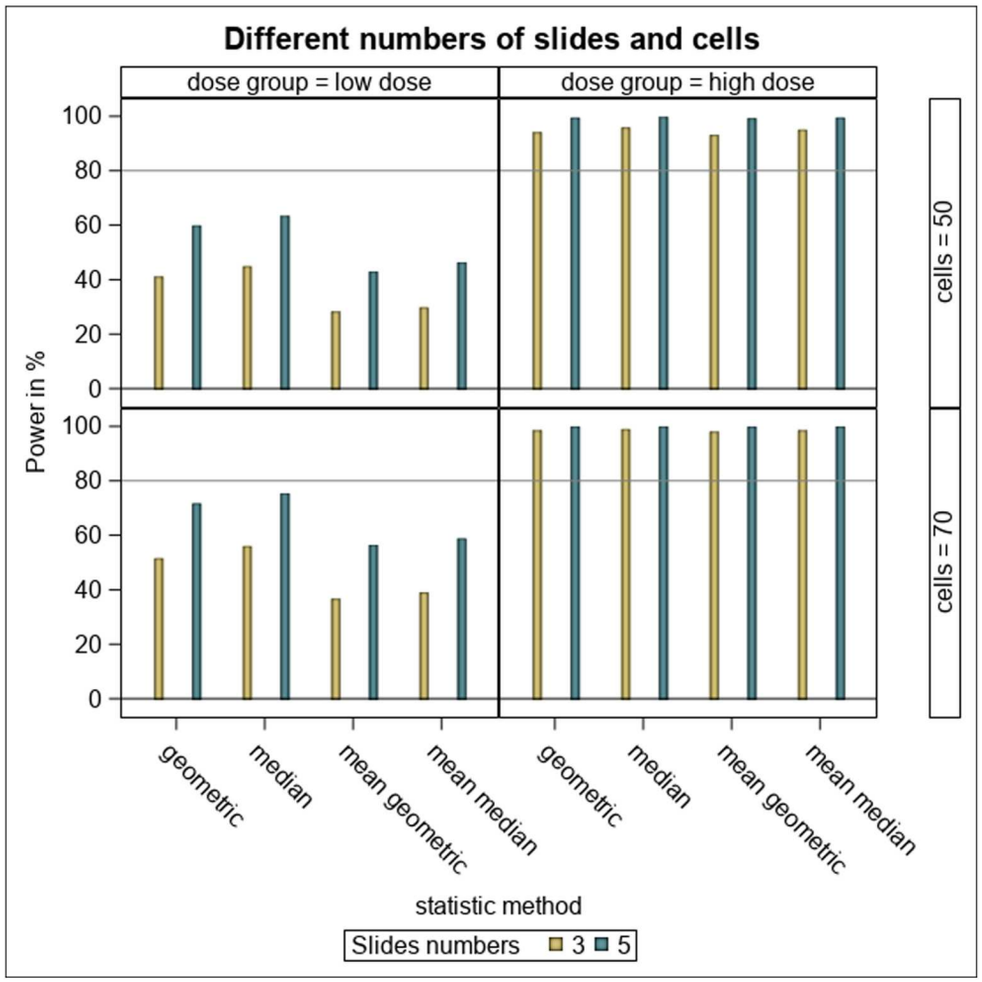



**Fig. 4.** Statistical power for different simulation scenarios comparing treatment vs. vehicle group using different numbers of slides and cells with 6 animals. The columns represent the respective dose groups (low and high) and the row represents the corresponding number of cells per slide (50 and 70). The bars display the power of the four statistical measures (geometric: geometric mean per slide, median: median per slide, mean geometric: arithmetic mean per animal of slide geometric means, mean median: arithmetic mean per animal of slide medians). The colour represents the number of slides. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

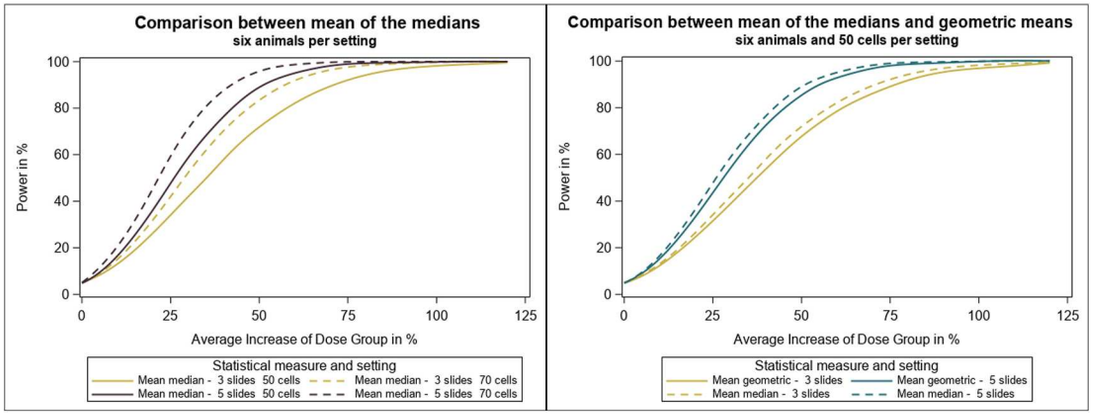



**Fig. 5.** Power curves for geometric means and medians (statistical animal model, gamma distributed data). LEFT: Power curves of the arithmetic mean per animal of the slide medians for different simulation scenarios based on 6 animals, on 50 or 70 gamma distributed measurements (cells) each observed on 3 or 5 slides. The y-axis shows the estimated power (proportion of rejected cases of null hypotheses); the x-axis illustrates differences between vehicle and dose groups [%]. RIGHT: Power curves of the arithmetic mean per animal of slide medians and geometric means, respectively, for different simulation scenarios for a varying number of slides (6 animals, 50 gamma distributed measurements (cells) per slide). The y-axis shows the estimated power (proportion of rejected cases of null hypotheses); the x-axis illustrates differences between vehicle and dose groups [%].

slide. This could be seen for slide summaries (used in the "statistical slide model") as well as for summaries per animal ("statistical animal model"). Although the amount of our experimental data was small, their basic behaviour, for example their skewness etc., coincides with what has been observed and discussed elsewhere (see e.g. Bright et al., 2011; Verd and Rottmann, 2015; Lovell and Omori, 2008).

For simulation purposes we chose a gamma-type distribution and employed a mixture of two distribution functions, one based on empirical moments (mean, variance) of existing data sets mixed with a second one based on the same mean, but a larger variance (5 times larger

plus an offset of 10). The second distribution described up to 10% noisy and extreme data per slide. This mixture of two different distribution functions should serve to create a realistic simulation framework.

In our work, we considered two basic models, one used several single slide summaries per animal and was denoted as "statistical slide model". The other model condensed the slide summaries per animal to one value ("statistical animal model"). Both models were deliberately kept simple, just including treatment as a fixed effect.

We refrained from involving possible animal, slide or sex effects into the statistical model, also knowing, that theoretically, there should be no slide effect, e.g., based on the location in the electrophoresis chamber, etc. By randomization of the slides. However, the incorporation of additional parameters, random factors, nested effects etc. into generalized linear models, non-parametric or Bayesian approaches are definitely interesting topics for future works (Wiklund and Agurell, 2003).

In our simulation settings, we used different experimental scenarios and varied the number of animals (6 or 8), slides (3, 4 or 5) and cells per slide (30, 50 and 70). Note that these simulation scenarios for the endpoint tail intensity were quite similar to those chosen by Wiklund & Agurell for tail length and tail moment.

At first, no systematic differences between these approaches were detected under $H_0$. In addition, all strategies seemed to have an empirical type I error of approximately 5%, i.e. none of the approaches was substantially conservative or liberal.

It is obvious that under $H_1$, the power of all different models increased with an increasing number of animals, slides and cells (see Fig. 4 and Figure B2).

For ethical and practical reasons it might be more appropriate to investigate more cells or slides than animals and therefore, we mainly focused on different combinations of cells and slides.

The key task of this work was to analyse whether there are substantial differences in statistical power between different models in dependency on the chosen slide summary. To this end, various simulation scenarios of animals, slides, and cells were developed, evaluated and displayed graphically. Assuming gamma distributed data, it turned out that median-based models showed a substantially increased statistical power compared to those involving the geometric mean per slide (see Fig. 5 (RIGHT)). Of course, results were based on the chosen simulation setting and varied between different configurations. Thus, differences in statistical power between different slide summaries reached up to about 7% for gamma distributed observations (see Fig. 5 (RIGHT)) or even 15% for lognormal data (see Appendix Figure B1 (RIGHT)), which were fundamental and somehow dramatic for the interpretation of experiments.

According to the present simulations and their ingredients, applications of geometric means instead of medians per slide consistently yielded a different statistical power to detect "moderate" differences between treatment groups. Firstly and mainly, this difference was due to the skewness of the data distribution; secondly, it depended on the amount of "outliers", and thirdly, it was based on certain statistical properties of the two centrality measures (e.g. the breakdown point, see section 2.3. "Statistical strategies"). Of course, distributions, "outliers" etc. go along with toxicological properties like the homogeneity of induced damage, mechanisms of action and/or the existence of different cell populations.

In principle, our results based on pairwise comparisons of vehicle and dose groups were confirmed using down-turn-protected trend tests (Bretz and Hothorn, 2001, 2003; Hothorn, 2004). These have the advantage of making statements on dose-response relationships, which are not limited to only two groups.

Of course, despite of these fundamental differences between different slide summaries on the final statistical outcome of an assay, future work involving additional experimental data, further types of distributions and summarizing measures liked trimmed, censored or weighted means, etc. might be valuable. In general, we have been investigating statistical properties in terms of power, though knowing that in this setting not only the power was of interest, but also the question which centrality measure provided the most suitable characterization of the treatment investigated.

## 5. Conclusion

The comet assay is a common test within a standard battery for genotoxicity testing able to detect chemically and physically induced DNA strand breaks in cells or nuclei isolated from various tissues. In this paper, we compared different summarizing measures per slide within various experimental settings and certain dose-response relationships.

It turned out that the use of the geometric mean as a summarizing measure per slide could lead to fundamentally different statistical test results compared to the use of medians, which is recommended in the current OECD guideline TG 489. Under certain circumstances, these differences were huge, such that, a negative statistical test outcome became positive or vice versa.

That means, one and the same data set and one and the same statistical strategy to compare treatment with vehicle observations can produce fundamentally different test results depending on the chosen summarizing mechanism per slide.

In general, our findings were mainly based on simulated data sets using a realistic framework stemming from existing observations.

According to our point of view, future work is needed on larger data sets from different tissues in order to investigate the influence of slide summaries (incl. further alternatives like trimmed, censored or weighted means), data distributions, fixed and random effects, or other models on the final statistical outcome in more detail.

## Funding

No funding received from any organisation other than Bayer AG, Genetic Toxicology and Research & Early Development Statistics, Berlin, Germany.

## 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.

## Acknowledgement

We would like to thank the involved colleagues from Bayer AG, Genetic Toxicology and Research & Early Development Statistics for providing data and for their valuable contributions to discussions.

## Appendix A

Median vs. Geometric mean estimates for gamma distributions.

In case of gamma distributed data, the relationship between geometric means and medians is not trivial, since there is no closed form of the median. However, an approximation of the median can be described as follows (Banneheka and Ekanayake, 2010):

$$\widehat{x}_{med} = \frac{(3\widehat{\alpha} - 0.8)}{(3\widehat{\alpha} + 0.2)}\bar{x},$$

where $\widehat{\alpha}$ is the moment estimator of the positive gamma parameter $\alpha$ and $\overline{x}$ denotes the arithmetic mean. The geometric mean is then given as follows (Johnson, 1994):

$\widehat{x}_{geo} = \overline{x}$*exp($\Psi(\widehat{\alpha}) - \log(\widehat{\alpha})$).

Here $\Psi(\widehat{\alpha}) = \frac{d}{d\widehat{\alpha}}(\log\Gamma(\widehat{\alpha}))$ denoted the digamma function.

It can be shown that $\widehat{x}_{med} > \widehat{x}_{geo}$ , *if* $\widehat{\alpha} \geq 0.31$ and $\widehat{x}_{med} \leq \widehat{x}_{geo}$ , *if* $\widehat{\alpha} < 0.31$.

In our simulations, $\widehat{\alpha}$is close to the threshold value of 0.31. Therefore, both inequalities can be observed, but in most cases, the median is greater than the geometric mean.

## *Appendix B*

In case of lognormal data, power curves are given in Figure B1 (LEFT and RIGHT):

## *Appendix C*

**Table C.1**
Contrasts for down-turn-protected trend test

| No | Hypothesis | Shape |
|---|---|---|
| 1 | $^{\mu}v=^{\mu}l=^{\mu}m=^{\mu}h$ (H0) |  |
| 2 | $^{\mu}v<^{\mu}l<^{\mu}m<^{\mu}h$ |  |
| 3 | $^{\mu}v=^{\mu}l<^{\mu}m<^{\mu}h$ |  |
| 4 | $^{\mu}v=^{\mu}l<^{\mu}m$ |  |
| 5 | $^{\mu}v=^{\mu}l<^{\mu}m=^{\mu}h$ |  |
| 6 | $^{\mu}v=^{\mu}l=^{\mu}m<^{\mu}h$ |  |
| 7 | $^{\mu}v<^{\mu}l=^{\mu}m<^{\mu}h$ |  |
| 8 | $^{\mu}v<^{\mu}l<^{\mu}m=^{\mu}h$ |  |
| 9 | $^{\mu}v<^{\mu}l=^{\mu}m=^{\mu}h$ |  |
| 10 | $^{\mu}v<^{\mu}l=^{\mu}m$ |  |
| 11 | $^{\mu}v<^{\mu}l<^{\mu}m$ |  |
| 12 | $^{\mu}v<^{\mu}l$ |  |

Index: $v$ = vehicle group, $l$ = low dose group, $m$ = medium dose group and $h$ = high dose group, grey: down-turn-protection.

**Table C.2**
Empirical mean and standard deviation (SD) per experiment

| Experiment | Mean | SD |
|---|---|---|
| 1 | 2.747 | 3.853 |
| 2 | 0.790 | 2.554 |
| 3 | 2.468 | 4.143 |
| 4 | 1.436 | 3.542 |
| 5 | 4.803 | 8.659 |
| All(mean) | 2.449 | 4.550 |

**Table C.3**
Down-turn protected trend tests, proportion of significant tests in simulation example ($p < 0.05$); for the different contrast see Appendix Table C.1

| Contrast | Method | | | |
|---|---|---|---|---|
| | Median | Mean Median | Geometric | Mean Geometric |
| No.2 | 0.8229 | 0.7685 | 0.7506 | 0.6990 |
| No.3 | 0.7329 | 0.6486 | 0.6474 | 0.5715 |
| No.4 | 0.3103 | 0.2706 | 0.2566 | 0.2231 |
| No.5 | 0.7441 | 0.7100 | 0.6676 | 0.6290 |
| No.6 | 0.5863 | 0.5014 | 0.4994 | 0.4355 |
| No.7 | 0.7869 | 0.7187 | 0.7168 | 0.6455 |
| No.8 | 0.7827 | 0.7545 | 0.7166 | 0.6856 |
| No.9 | 0.6956 | 0.6713 | 0.6327 | 0.6049 |
| No.10 | 0.4213 | 0.4003 | 0.3779 | 0.3491 |
| No.11 | 0.4350 | 0.3882 | 0.3788 | 0.3361 |
| No.12 | 0.1775 | 0.1634 | 0.1604 | 0.1442 |

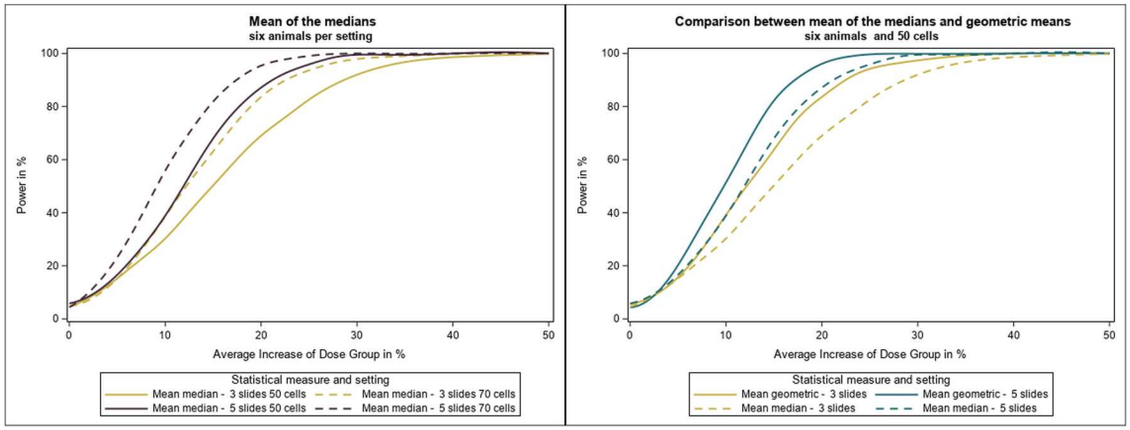



**Fig. B.1.** Power curves for geometric means and medians (statistical animal model, lognormal distributed data). LEFT: Power curves of the arithmetic mean per animal of the slide medians for different simulation scenarios based on 6 animals, on 50 or 70 lognormal distributed measurements (cells) each observed on 3 or 5 slides. The y-axis shows the estimated power (proportion of rejected cases of null hypotheses); the x-axis illustrates differences between vehicle and dose groups [%]. RIGHT: Power curves of the arithmetic mean per animal of slide medians and geometric means, respectively, for different simulation scenarios for a varying number of slides (6 animals, 50 lognormal distributed measurements (cells) per slide).

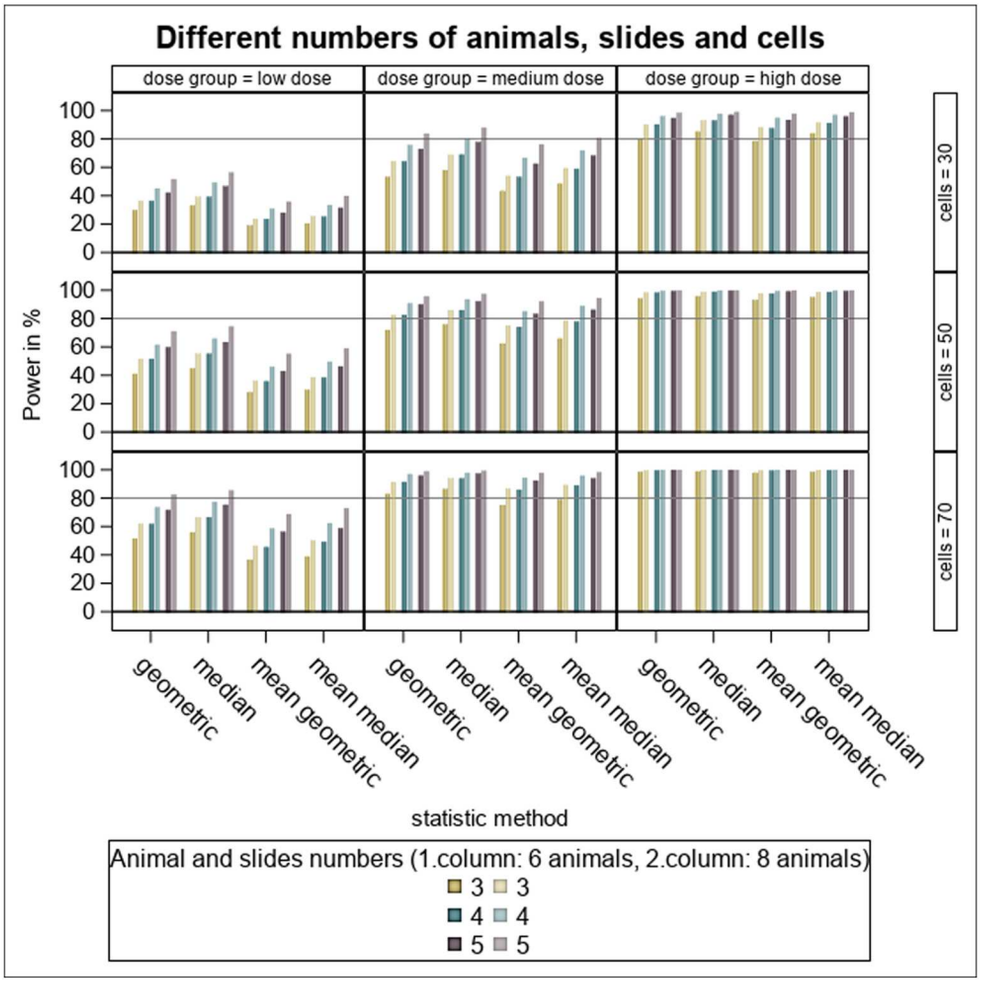



**Fig. B.2.** Statistical power for different simulation scenarios comparing treatment vs. vehicle group using different numbers of animals, slides and cells (gamma distributed data). The columns represent the respective dose groups (low, medium, high) and the row represents the corresponding number of cells per slide (30, 50 and 70). The bars display the power of the four statistical measures (geometric: geometric mean per slide, median: median per slide, mean geometric: arithmetic mean per animal of slide geometric means, mean median: arithmetic mean per animal of slide medians). The colour intensity (full/transparent) represents the number of animals and the colour itself the number of slides.

## References


Banneheka, B.M.S.G., Ekanayake, G.E.M.U.P.D., 2010. A new point estimator for the median of gamma distribution. Viyodaya Journal of Science 14, 95–103.

Bretz, F., Hothorn, L., 2001. Testing dose-response relationships with a priori unknown possibly nonmonotone shapes. J. Biopharm. Stat. 11, 193–207.

Bretz, F., Hothorn, L.A., 2003. Statistical analysis of monotone or non-monotone dose–response data from in vitro toxicological assays. Alternatives to Laboratory Animals 31 (1_Suppl. l), 81–96.

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.

Burlinson, B., Tice, R.R., Speit, G., Agurell, E., Brendler-Schwaab, S., Collins, A., Escobar, P., Honma, M., Kumaravel, T., Nakajima, M., Sasaki, Y., Thybaud, V., Uno, Y., Vasquez, M., Hartmann, A., 2007. Fourth international workgroup on genotoxicity testing: result of the in vivo comet assay Workgroup. Mutat. Res. 627, 31–35.

Collins, A.R., 2004. The comet assay for DNA damage and repair: principles, applications, and limitations. Mol. Biotechnol. 26 (3), 249–261.

Dierckx, P., 1993. Curve and Surface Fitting with Splines, first ed. Oxford University Press, New York.

Duez, P., Dehon, G., Kumps, A., Dubois, J., 2003. Statistics of the Comet assay: a key to discriminate between genotoxic effects. Mutagenesis 18 (2), 159–166.

Fairbairn, D.W., Olive, P.L., O'Neill, K.L., 1995. The comet assay: a comprehensive review. Mutat. Res. Genet. Toxicol. 339 (1), 37–59.

Guérard, M., Marchand, C., Plappert-Helbig, U., 2014. Influence of experimental conditions on data variability in the liver comet assay. Environ. Mol. Mutagen. 55, 114–121.

Hansen, M.K., Sharmab, A.K., Dybdahl, M., Bobergb, J., Kulahcia, M., 2014. In vivo Comet assay – statistical analysis and power calculations of mice testicular cells. Mutat. Res. 774, 29–40.

Hothorn, L.A., 2004. A robust statistical procedure for evaluating genotoxicity data. Environmetrics 15, 635–641.

Hothorn, L.A., 2016. Statistics in Toxicology Using R, first ed. Chapman and Hall/CRC, New York.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions, second ed., vol. 1. John Wiley, New York.

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 III. Appropriate follow-up testing in vivo. Mutat. Res. 654 (2), 114–132.

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.

Lovell, D.P., Omori, T., 2008. Statistical issues in the use of the comet assay. Mutagenesis 23 (3), 171–182.

OECD, 2016. Test No. 489. Vivo Mammalian Alkaline Comet Assay, OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris.

Östling, O., Johanson, K.J., 1984. Microelectrophoretic study of radiation - induced DNA damages in individual mammalian cells. Biochem. Biophys. Res. Commun. 123 (1), 291–298.

Shanmugam, R., Chattamvelli, R., 2015. Statistics for Scientists and Engineers, first ed. John Wiley & Sons, Hoboken, New Jersey.

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.

Stroup, W.W., Milliken, G.A., Claassen, E.A., Wolfinger, R.D., 2018. SAS for Mixed Models: Introduction and Basic Applications. SAS Institute Inc., Cary.

Tice, R.R., Agurell, E., Anderson, D., Burlinson, B., Hartmann, A., Kobayashi, H., 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.
Verbeke, G., Molenberghs, G., 2009. Linear Mixed Models for Longitudinal Data, reprint edition. Springer Verlag, New York.
Verd, P.E., Rottmann, A., 2015. Statistical models to analyze genotoxicological experiments with the comet assay data. Ann. Biom. Biostat. 2 (3), 1025.
Wiklund, S.J., Agurell, E., 2003. Aspects of design and statistical analysis in the Comet assay. Mutagenesis 18 (2), 167–175.

# *Article 2*

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 $x_1, \ldots, x_n$, the (arithmetic) mean is calculated as $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$. For the calculation of a median, data are arranged in ascending order such that $x_{(1)} \leq \ldots \leq x_{(n)}$ and the median $x_{\text{med}}$ 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 $x_{\text{geom}}$ is calculated as $\sqrt[n]{x_1 \bullet \ldots \bullet x_n}$. 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,

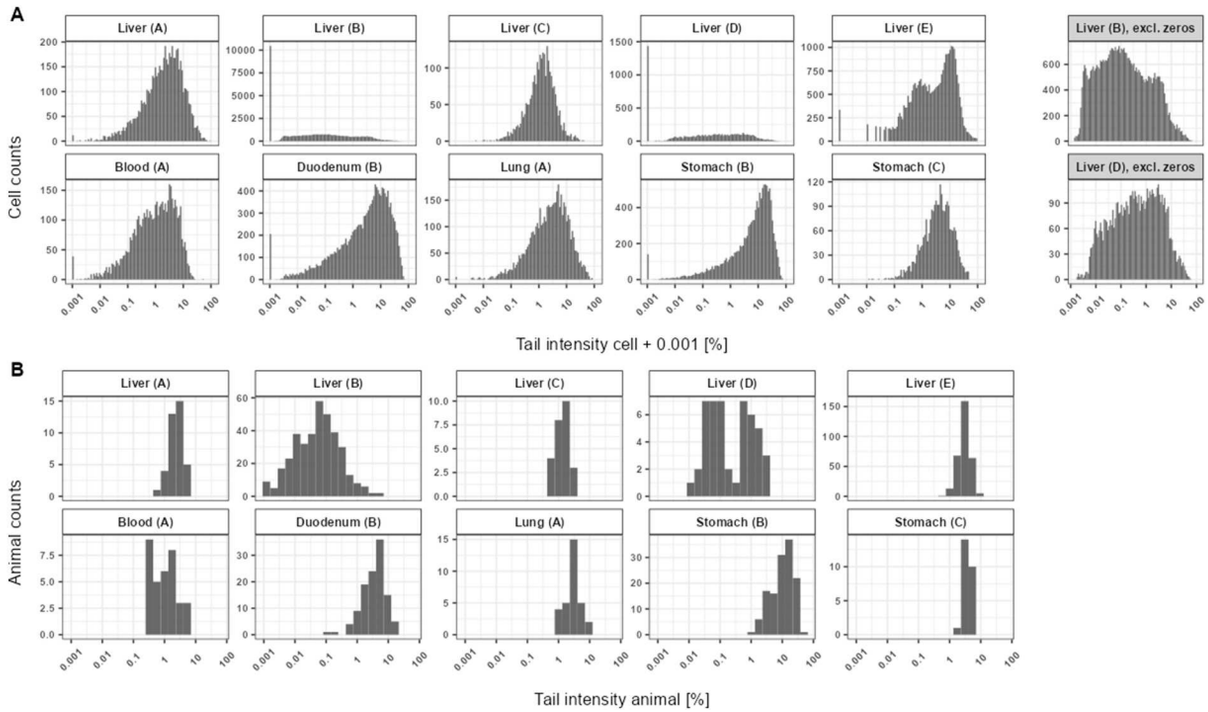



**Fig. 1.** (**A**) Left: Distribution of TI values for negative controls after adding + 0.001 on the cell level and log-transformation, stratified by organs and laboratories (A–E). For better interpretability, the values on the x-axis indicate the values prior to log-transformation, i.e., the original 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

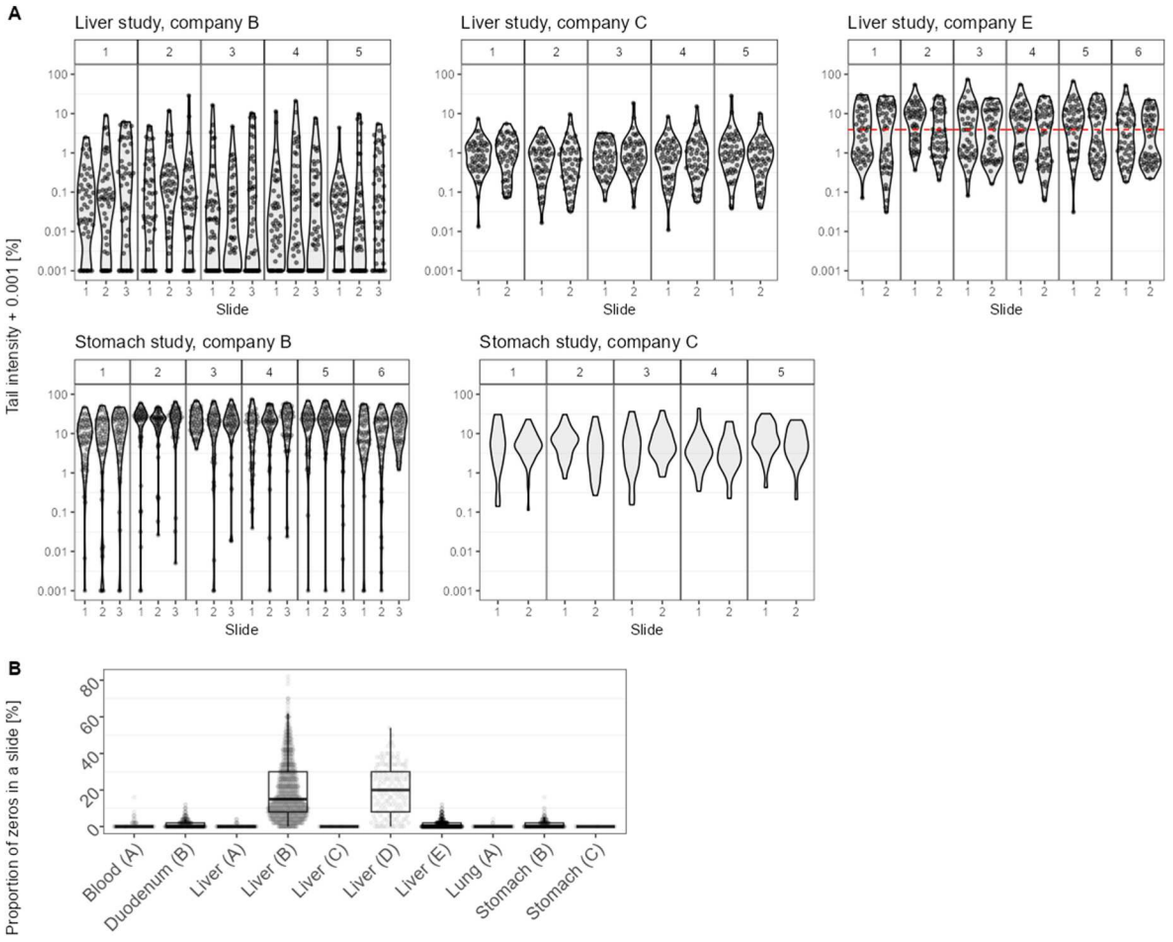



**Fig. 2.** (**A**) Violin plots of single-cell negative control data of five representative studies with five or six animals each for liver (top row) or stomach (bottom row) of laboratory B (left), C (middle) and E (right). The data per animal (the 5–6 panels within each sub-graphic) is subdivided into 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.

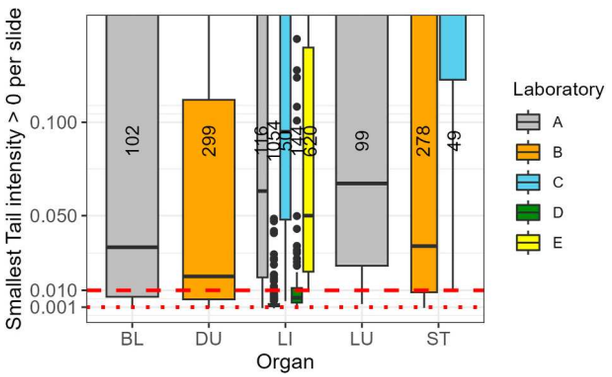



**Fig. 3.** *Smallest non-zero TI values* per *slide across laboratories and organs. The red dotted line indicates the OECD TG 489-proposed constant of 0.001 to be added to all data to enable processing steps such as a logarithmic transformation. The dashed line represents a heuristic alternative constant 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 b y 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.

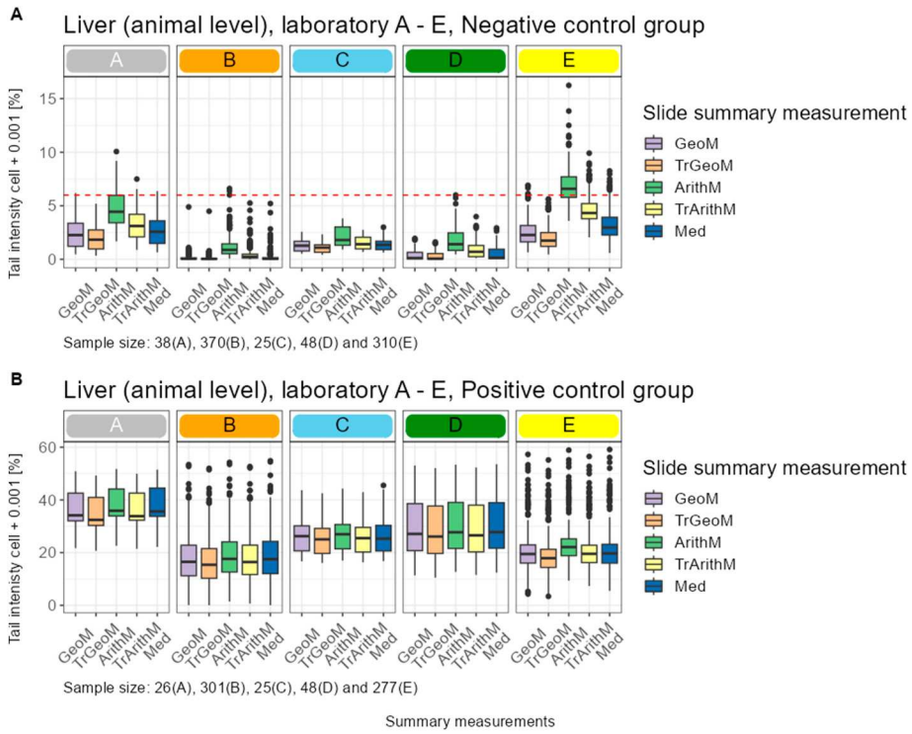



**Fig. 4.** (**A**): Negative control group data and (**B**) positive control group data from the different laboratories for liver tissue using five different statistical slide summary measures (GeoM: geometric mean, purple; TrGeoM: trimmed (10%) geometric mean, orange; ArithM: arithmetic mean, green, 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).

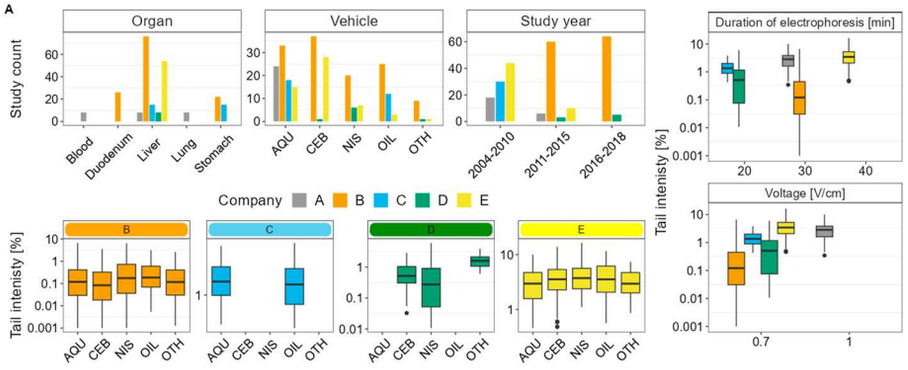
Regarding technical settings and the analysis system, each laboratory maintained a single protocol (Table S15). For example, all data generated by laboratory A used an electrophoresis duration of 30 min. In contrast, the used vehicle varied within the laboratories (Fig. 4A, upper left), depending on 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




| 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.

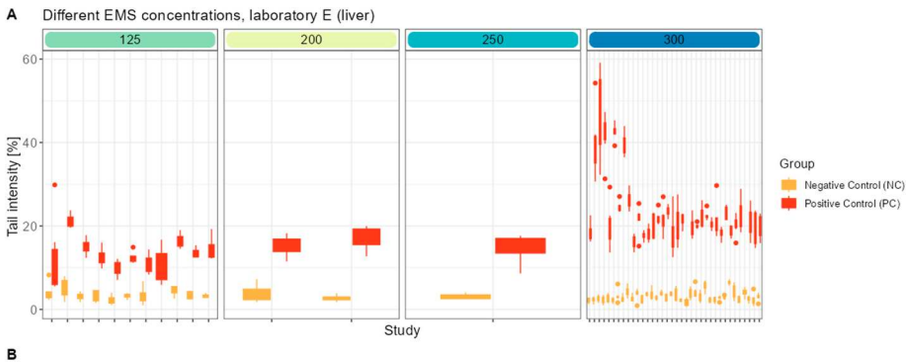
Finally, different empirical quantiles of the per study ratio between the two groups were determined for each laboratory separately. The values vary extremely for the same laboratory (Fig. 6B). Increasing values are alarming, as, e.g., a value of 0.3 means that the value for the NC is already 30% of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




B

| 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.*

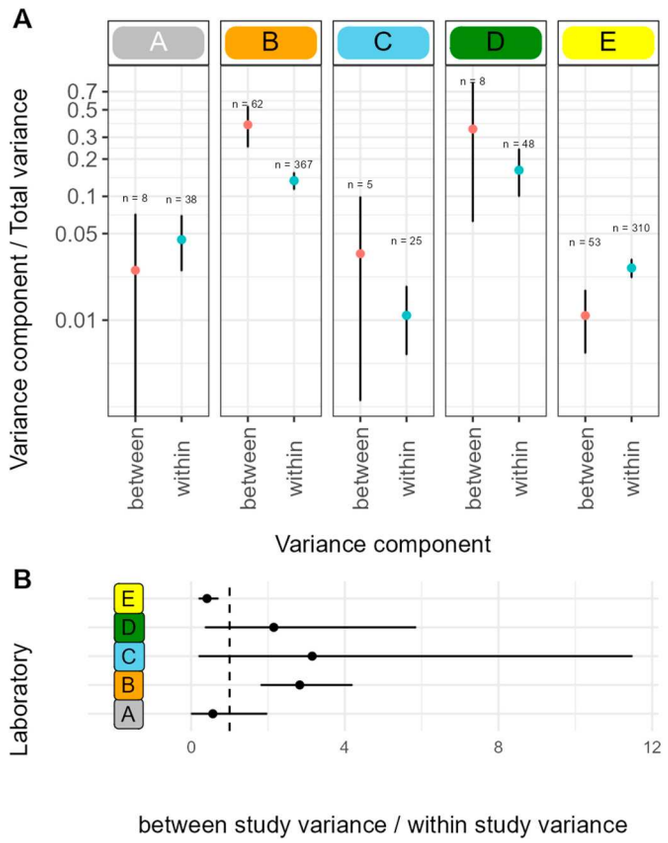



**Fig. 7.** (**A**) Estimated variance components. Red dots: Between study variance; Blue dots: Within study (residual) variance black bars: 95% pointwise confidence intervals, n: number of factor levels available for the estimation of the corresponding variance component. The different laboratories 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.

# *Article 3*

# *DEVELOPMENT AND APPLICATION OF BRAIN TISSUE BASED MULTI-OMICS PROFILE SCORES FOR ALZHEIMER'S DISEASE*

Timur Tug[1,2*], Donghai Liang[2,3], Sven Teschke[1], Youran Tan[3], Marla Gearing[4,5], Allan I. Levey[5], James J. Lah[5], Aliza P. Wingo[6], Thomas S. Wingo[7,8], Michael Lau[9,10], Katja Ickstadt[1], Anke Hüls[2,3*]

[1]Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany,
[2]Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA,[3]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA,
[4]Department of Pathology and Laboratory Medicine, Emory University, 100 Woodruff Circle, Atlanta, GA 30322 USA,
[5]Department of Neurology, Emory University School of Medicine, 100 Woodruff Circle, Atlanta, GA 30322 USA,
[6]Department of Psychiatry, University of California, Davis, 2230 Stockton Blvd, Sacramento, CA 95817, USA,[7]Department of Neurology, University of California, Davis, 4860 Y Street, Sacramento, CA 95817, USA,
[8]Alzheimer's Disease Research Center, University of California, Davis, 1651 Alhambra Blvd, Suite 200A, Sacramento, CA 95816,USA
[9]Mathematical Institute, Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
[10]eBay Inc., 2025 Hamilton Avenue, San Jose, CA 95125, USA

*: corresponding authors: timur.tug@tu-dortmund.de, anke.huels@emory.edu

Corresponding authors:

Timur Tug, Vogelpothsweg 87, 44227 Dortmund, Germany

Email: timur.tug@tu-dortmund.de

Anke Huels, PhD, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA.

Email: anke.huels@emory.edu

**Abstract (148/150 words)**

INTRODUCTION

Advances in omics technologies, such as epigenomics and metabolomics, provide novel insights into the biological mechanisms underlying Alzheimer's disease (AD). However, little is known how different omics layers interact and jointly relate to AD neuropathology.

METHODS

We performed a comprehensive single- and multi-omics analysis integrating genome-wide DNA methylation and high-resolution metabolomics data from 157 frontal cortex samples. We developed novel single and multi-omics profile scores (PS) for AD pathology, using a combination of machine learning, regression, and pathway analysis.

RESULTS

The best multi-omics PS showed an $R^2$ of 0.15 for the ABC score (Amyloid, Braak, CERAD), independent of age, sex, race and socioeconomic factors. The pathway analyses identified lipid metabolism and signal transduction as key biological pathways among the included metabolites and CpG sites.

DISCUSSION

The integration of DNA methylation and metabolomics provides deeper insights into AD pathophysiology and identifies promising molecular targets for potential disease-modifying approaches.

**Research in Context**

1. **Systematic Review:** Alzheimer's disease (AD) is a multifactorial disorder with complex molecular underpinnings. Advances in omics technologies, particularly DNA methylation (DNAm) and metabolomics, have provided insights into AD pathophysiology. Prior studies identified associations between DNAm and AD-related neuropathology, while metabolomics studies highlighted alterations in lipid metabolism and oxidative stress. However, most research examines these omics layers separately, limiting insights into their interplay.

2. **Interpretation:** Our study integrates DNAm and metabolomics data using novel PS approaches to improve AD prediction. The best DNAm-based profile score (PS) achieved $R^2 = 0.11$, outperforming metabolomics-based PS ($R^2 = 0.04$). Combining both omics layers improved predictive accuracy ($R^2 = 0.15$). Key pathways enriched in both layers include lipid metabolism and signal transduction, reinforcing their role in AD pathology.

3. **Future Direction:** Future studies should expand multi-omics approaches, conduct longitudinal analyses, validate findings in diverse cohorts, and explore translational applications for early diagnosis and therapy.

## BACKGROUND

Alzheimer's disease (AD) is a progressive neurological disorder that affects millions of Americans, with approximately 6.7 million people aged 65 and older currently living with the condition [1]. AD is the fifth-leading cause of death among older adults in the United States and poses a major public health challenge [2]. The financial burden is equally staggering, with the annual cost of care for AD patients projected to reach $580 billion in 2025, a figure expected to rise substantially in the coming decades [1,3]. As the prevalence of AD continues to rise globally due to aging populations, developing effective strategies to reduce its burden has become a critical priority. Substantial efforts are underway to create disease-modifying therapies targeting the molecular mechanisms of AD [4,5], but the complex and multifactorial nature of the disease has posed significant challenges.

Advances in high-throughput omics technologies, such as genomics, epigenomics, proteomics, and metabolomics, have provided new insights into the biological pathways and signatures underlying AD, offering promising avenues for innovative therapeutic strategies [6,7]. Among these, DNA methylation (DNAm) and metabolomics analyses have emerged as powerful tools for exploring AD etiology. These approaches offer complementary insights into AD's complex pathophysiology by capturing information on epigenetic changes and metabolic dysfunction—both recognized as core features of the disease [8]. DNAm, a key epigenetic mechanism, is closely linked to AD, with global hypomethylation, gene-specific methylation changes, and interactions with neuroinflammation, aging, oxidative stress, and environmental factors contributing to disease risk and progression. As the downstream product of the gene transcription and gene-environment interaction, metabolomics involves studying small molecules (metabolites) in biological systems, which can reveal changes in metabolic pathways associated with AD. Metabolic pathways that have been reported in association with AD include dysregulation in energy metabolism, lipid profiles, amino acid pathways, oxidative stress, and gut-brain axis interactions [9–12]. However, little is known how differences in DNAm and metabolomics interact and jointly influence the development of AD. Integrating DNAm and metabolomics through a multi-omics approach could illuminate shared biological pathways that may be particularly important for understanding AD etiology.

To address this knowledge gap, we conducted a comprehensive multi-omics analysis integrating genome-wide DNAm and high-resolution metabolomics data from 157 prefrontal cortex tissue samples of brain donors with varying stages of AD pathology, assessed with Braak staging, CERAD (**C**onsortium to **E**stablish a **R**egistry for **A**lzheimer's **D**isease) scoring, and the comprehensive ABC score (**A**myloid, Braak, **C**ERAD) [13,14]. Specifically, we calculated multi-omics profile scores (PS) integrating DNAm and metabolomic data, providing a holistic understanding of AD neuropathology by capturing both epigenetic and metabolic contributions. Our analysis framework not only identifies CpG sites and/or metabolomic features predictive of AD neuropathology levels but also elucidates mechanistic underpinnings by evaluating and comparing biological pathways enriched among the selected CpG sites and/or metabolomic features. The multi-omics framework developed in this study highlights the potential of combining epigenetic and metabolomic data to deepen our understanding of AD pathophysiology. This study extends the current state of research through several innovative approaches. We proposed novel brain

tissue-based multi-omics profile scores for AD, which integrate DNA methylation and metabolomics to better predict the neuropathological changes of the disease. Methodologically, advanced machine learning techniques such as Random Forest, Elastic Net and Boosting are used to efficiently analyze high-dimensional data and generate robust profile scores. These novel analysis strategies and methods and a comprehensive pathway analysis allow a more precise identification of relevant biological mechanisms.

## METHODS

### Study design

The Emory Goizueta Alzheimer's Disease Research Center (ADRC) established a brain bank to support Alzheimer's research, primarily enrolling research participants and patients clinically diagnosed with AD by Emory physicians. By the third quarter of 2020, the brain bank included 1,011 donors. Genome-wide DNAm and metabolomics profiling were conducted on 161 samples from donors deceased after 2007, with 159 samples passing quality control. All 161 donors had complete data for key covariates (e.g., age of death, race, sex and educational attainment) and outcome variables (e.g., ABC score, Braak Stage, CERAD). Informed consent and Institutional Review Board-approved protocols governed the research.

### Assessment of AD neuropathology

The ADRC conducted comprehensive neuropathologic evaluations on all donor brains using established research protocols and diagnostic criteria [13]. These assessments, performed by experienced neuropathologists, involved a variety of stains and immunohistochemical techniques, along with semi-quantitative scoring to evaluate AD and related neuropathologies in various brain regions [15]. AD neuropathology in this project was measured using the Braak staging, CERAD score, and ABC score - each assessing different aspects of disease progression. Braak Stage classifies the spread of neurofibrillary tangles (NFTs) tau-containing protein deposits, a hallmark of AD - in the brain across six stages [16]. Early stages (I and II) indicate NFTs in transentorhinal regions, while intermediate stages (III and IV) involve limbic regions, and later stages (V and VI) show NFTs spread throughout cortical areas. Higher stages reflect more extensive disease progression and broader NFT distribution in the brain [16]. The CERAD score evaluates the density of neuritic plaques, mainly composed of beta-amyloid, and categorizes them into four levels: none, sparse, moderate, and frequent [17]. These plaques are a primary indicator of AD, with higher CERAD scores signifying a greater accumulation of plaques, reflecting more advanced amyloid pathology [17]. The ABC score combines data from the Braak and CERAD scores with the Thal phase, which describes the spread of amyloid plaques in the brain. Thal staging ranges from phase 1 (amyloid in subcortical areas) to phase 5 (widespread distribution across the brain) [18]. The

ABC score synthesizes information on NFT spread and amyloid plaque density into a single assessment of AD pathology, categorizing it into four levels: none, low, intermediate, or high [14]. This score provides a more comprehensive evaluation of AD severity and helps to indicate the overall stage of the disease.

### Genome-wide DNA methylation

DNA was extracted from fresh-frozen prefrontal cortex tissues in 161 samples using the QIAGEN GenePure kit. DNAm was assessed using the Illumina Infinium MethylationEPIC BeadChips, processed in batches of 167 prefrontal cortex samples, which included six replicates. The raw intensity files were converted into a dataset containing beta values for each CpG site. These beta values were calculated as the ratio of the methylated signal to the total signal (methylated plus unmethylated), ranging from 0 to 1 on a continuous scale.

Preprocessing and quality control were conducted in R (v4.2.0) [19] using a validated quality control and normalization pipeline as previously described [20]. Out of the initial samples, 159 passed the quality control checks. After excluding single nucleotide polymorphism (SNP) probes, XY chromosome probes, and other low-quality probes, 789,286 CpG sites were retained for further analysis.

The final DNAm beta values were normalized to minimize probe-type differences and adjusted using ComBat to account for batch effects prior to downstream analyses [21]. Cell-type proportions (neuronal vs. non-neuronal cells) for each sample were estimated using the latest prefrontal cortex reference database and the R package *minfi* [22–29].

### High-resolution metabolomics

High-resolution metabolomic profiling of prefrontal cortex tissue was conducted using liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS) following established protocols [30–33]. Each sample was analyzed in triplicate with two complimentary chromatographic methods: hydrophilic interaction liquid chromatography (HILIC) for polar metabolites and reverse-phase chromatography (C18) for less polar compounds to enhance the coverage of feature extraction. Detected signals were characterized by accurate mass-to-charge ratio (m/z), retention time (RT), and ion intensity [34].

Raw data were converted to .mzML format and processed with apLCMS and xMSanalyzer for peak detection, alignment, feature quantification, batch correction, and quality filtering [35,36]. Metabolic features were further screened before being included in the final analysis based on strict criteria: detected in >15% of samples, median CV <30%, and Pearson correlation $\rho > 0.7$ among technical replicates. Data

were averaged across replicates with non-zero intensities and log2-transformed for statistical analyses. Finally, we included 20,051 features at the HILIC and 15,297 features at the C18, a total of 35,348 features.

### Covariates assessment

All models were adjusted for *a priori* selected covariates based on the literature, which include demographic and socioeconomic factors. Individual-level characteristics included sex, race (Black vs. White), educational attainment (high school degree or less, college degree, graduate degree) and age at death. The Area Deprivation Index (ADI) served as a proxy for neighborhood socioeconomic status, based on the 2015 Census Block Group data [37], indicating socioeconomic disadvantages in income, education, employment, and housing. The postmortem interval (PMI) refers to the time elapsed between a person's death and the collection of biological samples, such as tissue or blood. PMI is crucial in research involving post-mortem samples because it can impact on the quality and stability of molecular markers, including DNAm and metabolomics.

### Statistical analysis

Following a similar analysis pipeline as established for polygenic risk scores [38]), we developed single- and multi-omics PS based on metabolomics ($PS_{Metabolome}$) and DNAm ($PS_{DNAm}$) data, to predict AD neuropathology. A PS represents a weighted sum of selected variables:

$$PS = \sum_{k=1}^{K} \beta_k m_k \,, \qquad (1)$$

where $K$ is the number of selected variables (e.g., metabolites and/or CpG sites), $\beta_k$ the weight assigned to the $k$-th feature and $m_k$ is the actual value of the $k$-th feature.

The analytic pipeline is outlined in **Figure 1** and includes the following steps that are described in the following section: Stage 0 – DNAm and/or Metabolomic dataset linking and cleaning; Stage 1 – Split data randomly into training and test datasets; Stage 2 – Estimate the weights in the training dataset using different regression and machine learning methods (PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach); Stage 3 - Calculate the PS in the data dataset; Stage 4 - Validate the single- and multi-omics PS in the test data; Stage 5 - Conduct pathway analyses on identified CpG sites and/or metabolites.

Stages 0 and 1 – Data preparation and splitting into training and test data

Once the data sets are cleaned and merged (Stage 0), they are randomly split into training and test sets to prevent overfitting by ensuring that the models are tested on independent data. The training set is used for model building and feature selection, while the test set is reserved for evaluating model performance. In our analysis we split the data evenly into training and test data. However, with larger sample sizes, allocating a greater proportion to the training set can enhance the model's predictive accuracy (see [38] for related recommendations for polygenic scores). We repeated this process 10 times with 10 different seeds (10 iterations) to provide a robust evaluation of our model performance. Since some of the methods described in Step 2 do not allow for missing values, we used Random Forest as implemented in the R package missRanger to impute missing values in the metabolome data set [39]. For the multi-omics PS, the DNAm and metabolites beta values were transformed into z-scores to ensure that both omics datasets are on the same scale before calculating one PS for the combined dataset.

Stage 2 - Feature selection and weight calculation in the training data

In the training data, variable selection and weight calculation are performed using five different regression and machine learning approaches: Pruning & Thresholding (PT), Elastic Net (EN), Boosting (BO), Random Forest (RF) and Windows Approach (WA) using cross leverage scores. This step identifies the most informative CpG sites and/or metabolites related to neuropathological outcomes. A detailed description of the methods can be found in the Supplementary Materials.

Pruning and thresholding (PT) are complementary techniques to simplify high-dimensional datasets by reducing redundancy and focusing on the most relevant variables. Pruning removes highly correlated or redundant variables, decreasing multicollinearity and computational demands while retaining predictive accuracy. For the pruning step, we applied the agglomerative (bottom-up) clustering approach [40] using the R packages ClassDiscovery [41], flashClust [42] and cluster [43]. Complete-linkage clustering was used to define the distance between two clusters as the maximum distance between any pair of points, and the Pearson correlation was used as the distance metric. The analyses generate different numbers of clusters with different numbers of CpG sites/metabolomic variables. A single representative is then drawn from each cluster and used for pruning. Thresholding applies to a predefined cutoff (e.g., p-value threshold) to filter out weak associations and retain only significant variables. We used ordinal logistics regression models (adjusted for covariates) to estimate associations between each and the neuropathology markers. Only CpG site and/or variables with a p-value $< 0.05$ were included in the PS.

Elastic Net (EN) is a regularized regression technique that combines the penalties of Lasso (L1) and Ridge (L2) to select variables and assign weights. To perform the analysis, we used the R package ordinalNet [44].

Boosting (BO) combines several weakly predictive models to improve the prediction accuracy [45]. We used an extended version of the Boosting method proposed by [46] to allow for ordinal outcomes. The approach considers the response variable as an ordered factor and computes thresholds to distinguish between categories. The boosting process iteratively updates coefficients, computes gradients, and adjusts predictions, resulting in a final model with optimized parameters and scaling corrections.

The Random Forest (RF) algorithm operates by creating multiple decision trees using bootstrap sampling and feature subset selection. The final prediction is then aggregated across all trees, using majority voting for classification or averaging for regression, ensuring robust and accurate predictions. We obtained [47]variable importance measures (VIMs) within the RF fitting process using the R package ranger [47] [48]. This method performs an automatic selection of variables and the VIMs were used as weights to calculate the PS.

The sliding windows approach (WA) involves analyzing variables within specific "windows" or genomic regions. For each window, cross leverage scores are calculated for each variable based on matrix decomposition techniques. These scores quantify the association of the variables, their interactions and the outcome. We select those *q* variables with the absolute highest cross leverage score. This method is described in more detail in [49] with an R code in the corresponding in Supplementary Materials.

Stage 3 - Profile score calculation and prediction of AD neuropathology in the test data

With the selected variables from Stage 2, PS are calculated in the test dataset.

Three different PS were calculated (after equation (1)): Single-omics PS ($PS_{DNAm}$; $PS_{Metabolome}$) that only contain information from one omics layer (either DNAm or metabolomics), multi-omics PS that contain both the DNAm and metabolomics data ($PS_{DNAm+Metabolome}$), and joint PS models that contain the individual single-omics PS ($PS_{DNAm}$; $PS_{Metabolome}$) in the same prediction model with and without an interaction term between the single-omics PS ($PS_{DNAm}$ * $PS_{Metabolome}$). In our analysis, the outcome is ordinal, and all three models are ordinal logistic regression models [50,51] with the following equations:

Single-omics PS: $$Neurop.\ outcome \sim PS_{DNAm}\ (or\ PS_{Metabolome}) + Covariates \quad (2)$$

Multi-omics PS: $$Neurop.\ outcome \sim PS_{DNAm+Metabolome} + Covariates \quad (3)$$

Joint PS model: $$Neurop.\ outcome \sim PS_{DNAm} + PS_{Metabolome}\ (+PS_{DNAm} * PS_{Metabolome}) + Covariates \quad (4)$$

Stage 4 - Validation of PS models in the test data

The performance of the single-, multi-omics and joint PS models was evaluated in the test data based on a partial McFadden's $R^2$, also known as the partial pseudo-$R^2$, which is a measure of goodness of fit for logistic regression models, including ordinal logistic regression (can be found in the Supplementary Materials). Calculating partial $R^2$ allowed us to demonstrate the prediction $R^2$ for the PS, independent of the influence of the other covariates (sex, race, educational attainment, age at death, ADI, PMI). We further evaluated whether the PS was significantly associated with the neuropathology outcomes in the independent test data using a likelihood ratio test (p-value<0.05) in the ordinal logistic regression models from Step 3.

Stage 5 - Pathway analysis of selected metabolites and/or CpG sites

The final step involves a pathway analysis of the identified omics variables (CpG site and/or metabolomics features). Pathway analysis maps selected CpG sites and/or metabolomics features to known biological pathways, elucidating the biological mechanisms potentially underlying observed associations with neuropathology. Only the best performing PS were used for the pathway analyses.

For the DNAm PS, we conducted gene set enrichment analyses using the R package missMethyl [24,52–54] and the KEGG database [55–57].

We included the features/CpG sites that were selected in at least one, two, or three of the 10 iterations in the pathway enrichment analysis.

For the metabolomics PS, datasets for positive and negative ion modes were merged with experimental results to match mass-to-charge ratios (m/z) and retention times. Next, we used the R package metapone [58] to identify pathways from the KEGG database associated with the detected metabolites by leveraging adduct information and permutation-based statistical thresholds.

The top pathways were ranked based on p-values, and their significance was visualized with scatter plots showing the strength and size of CpG sites or metabolomic features contributing to each pathway.

Sensitivity Analyses

While the main analyses were conducted for the largest sample possible (N=154 for $PS_{DNAm}$, N=141 for $PS_{Metabolome}$, N=138 for the multi-omics PS and the joint PS

model), we conducted a sensitivity analysis, in which we restricted the $PS_{DNAm}$ and $PS_{Metabolome}$ to the donors with data on DNAm and metabolomics (N=138), to validate that differences between the different PS models are not due to differences in sample size.

## Results

### Study population

After excluding 4 brain donors with missing covariate information, a total of 157 samples were included in the current analysis (**Table 1**). 154 of them had DNAm data available, 141 of them had metabolomics data available and 138 of them had both DNAm and metabolomics data available.

The mean age at death was 76.4 years (standard deviation [SD]: 10.0). Most participants were White (89.2%) and 10.8% self-identified as Black or African American. The study population consisted of 54.8% males and 45.2% females. The study population was predominantly well-educated with 49.7% holding a college degree, and 28.0% a graduate degree. The mean ADI score was 36.1, with a standard deviation of 24.0, indicating a wide range in socioeconomic deprivation among the participants. The prevalence of the APOE ε4 allele (56.1% with at least one APOE ε4 allele) was much higher than that in the general population in the United States, which is estimated to be around 25-30% (Huang et al., 2017). Most donors (59.2%) had high levels of AD neuropathologic changes (ABC score of "high"), 47.1% of donors were classified as Braak Stage 6 and 70.1% of donors had frequent neuritic plaques on the CERAD score, indicating a high prevalence of AD neuropathology in this study population.

### Single-omics PS

First, we calculated single-omics PS (equation (1)) based on DNAm and metabolomics data (Figure 1). $PS_{DNAm}$ and $PS_{Metabolome}$ were calculated for each of the three neuropathological outcomes (ABC score, Braak stage and CERAD score) using six different approaches (PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach). The results for the ABC score are shown in Figure 2 and the results for the other two outcomes are included in the appendix (Figures S1 and S2). The results from all three outcomes were similar.

For the $PS_{DNAm}$ (Figure 2A), PT reaches the highest median $R^2$ (0.11) over all other methods and found a significant association between the DNAm PS and the ABC score in all 10 iterations. Compared to the PT approach, the RF approach performed worse in terms of both the median $R^2$ value of 0.02 and the number of significant associations (4 out of 10). Other noteworthy approaches for the DNAm PS are PT+EN and BO. The median $R^2$ was similar to the RF approach (PT+EN: 0.04; BO: 0.01) and PT+EN identified a significant association between DNAm PS and ABC

score in 6 out of 10 iterations and the BO approach in 3 out of 10 iterations. WA and WA+EN resulted in a smaller median $R^2$ and no significant associations with the ABC score.

For the $PS_{Metabolome}$ (Figure 2B), the RF approach performed best, leading to a median $R^2$ value of 0.04 across the 10 iterations, identifying a significant association between the metabolomics PS and the ABC score in 8 out of 10 iterations. While the PT approach detected a significant association in 7 out of 10 iterations, its median $R^2$ value was similar to the RF approach ($R^2$=0.04). The other approaches resulted in fewer significant associations and lower $R^2$ values.

The results were similar in a sensitivity analysis, in which we restricted the $PS_{DNAm}$ and $PS_{Metabolome}$ to the donors with data on DNAm and metabolomics (N=138; Supplementary Figure S6).

**Multi-omics PS**

When combining the DNAm and metabolomics data in one dataset to derive a multi-omics PS ($PS_{DNAm+Metabolome}$; equation (3)), the PT approach ($R^2$=0.15, 10 out of 10 significant associations) and the RF approach ($R^2$=0.01, 2 out of 10 significant associations) performed the best (Figure 2C). The $R^2$ values were greater than in the single-omics PS (PT for the DNAm PS (Figure 2A)).

**Joint PS models**

Next, we included the individual DNAm PS and metabolomics PS in the joint PS model (with and without an interaction term between the two PS; equation (4)) to evaluate whether this results in a higher model prediction performance (Figure 3).

In general, there was only a small improvement in $R^2$ when combining both PS in comparison to the single-omics scores. The models with an interaction term between the two PS always performed a little better than the models without the interaction term. The best model $R^2$ (0.15) was reached when combining the DNAm PS based on the PT approach with the metabolomics PS based on the RF approach, which were also identified as the best PS in the individual omics data (Figure 1). $R^2$ for the model with the interaction term ($R^2$=0.15) was slightly higher than $R^2$ for the model without the interaction term between the two PS ($R^2$=0.13). For most methods (including RF and PT), the multi-omics PS $R^2$ were higher than the joint PS $R^2$.

**Secondary analyses of the best-performing PS**

The two best-performing single-omics PS (PT for DNAm and RF for metabolomics) showed a Pearson correlation of $\rho = 0.25$) (Figure 4). Of note, absolute correlations between the other single-omics PS were lower ranging from 0.00 to 0.19.

For the two best-performing single-omics PS, 100 iterations were carried out to receive a more precise estimate for the prediction $R^2$ ($R^2$=0.13 PT for DNAm and $R^2$ = 0.01 RF for metabolomics) (Supplementary Figure S7). The results reflect the previous results ($R^2$=0.11 PT for DNAm and $R^2$ = 0.04 RF for metabolomics) and

most important statements with a relatively large variability. A similar behavior is also observed for the captured variables: the values found from the 10 iterations are in the range of the 100 iterations found (e.g. 1061 (10 iterations) to 1032 (100 iterations) captured variables with RF for metabolomics).

To map CpG sites and/or metabolites that were selected by the best-performing PS (PT for DNAm and RF for metabolomics) to known biological pathways, we conducted a KEGG pathway enrichment analysis for all CpG sites/metabolites that were selected in one, two or three of the ten iterations (Figure 5). The results for all significantly enriched pathways (p-value < 0.05) in at least one of the ten iterations are summarized in Figure S5 and Figure 5 presents the enriched pathway classes which had at least one pathway with significant p-values in more than one iteration. For $PS_{DNAm}$, 20 KEGG pathways that can be summarized in 13 classes, were significantly enriched (p-value < 0.05) in at least one of the ten iterations (Figure 5A, Figure S5A). Among these, the most prevalent pathway classes were lipid metabolism and digestive system. For $PS_{Metabolome}$, 21 KEGG pathways that can be summarized in 9 classes, were significantly enriched (p-value < 0.05) in at least one of the ten iterations (Figure 5B, Figure S5B). Among these, the most prevalent pathway classes were lipid metabolism and signal transduction. Of note, both of these pathway classes were also identified in the $PS_{DNAm}$ analyses.

## Discussion

We performed a comprehensive single- and multi-omics analysis integrating genome-wide DNAm and high-resolution metabolomics data derived from 157 frontal cortex samples, aiming to gain a better understanding of AD neuropathology. We developed single- and multi-omics PS based on DNAm and metabolomics data to predict the neuropathological features of AD independent of age, sex, race and socioeconomic factors, using various machine learning and regression-based approaches. The best-performing $PS_{DNAm}$, which was calculated using the PT approach, predicted AD neuropathology levels with a median partial $R^2$ of 0.11 and the best-performing $PS_{Metabolome}$, which was calculated using RF, reached a median $R^2$ of 0.04. Combining the DNAm and metabolomics data in the same PS model only led to a small improvement in the prediction accuracy ($R^2$ = 0.15 for the best-performing joint PS model). Interestingly, $PS_{DNAm}$ and $PS_{Metabolome}$ were moderately correlated with a Pearson correlation of 0.25 and the biological pathways lipid metabolism and signal transduction were enriched among the identified CpG sites as well as the identified metabolites, emphasizing the importance of these two AD-related pathways across various omics layers.

Our analysis showed that DNAm-based PS had a better predictive performance for AD neuropathology than metabolomics-based PS. The DNAm PS achieved a median $R^2$ value of 0.11, while the best metabolomics PS only achieved a median $R^2$ value of 0.04. These results suggest that DNAm may be a stronger indicator of

neuropathologic changes in AD than metabolomics. Current evidence highlights the distinct yet complementary roles of brain DNAm and metabolomics in understanding neuropathology markers associated with AD. Alterations in DNAm patterns are closely linked to neuroinflammation, oxidative stress, and other pathological processes in AD. For instance, global hypomethylation and gene-specific methylation changes have been identified as significant contributors to disease progression, impacting genes involved in synaptic function and neurodegeneration [8]. On the other hand, brain metabolomics has recently emerged as a powerful tool for identifying metabolic dysfunctions associated with AD. Research indicates that specific metabolic pathways—such as those involving lipid metabolism, energy production, and amino acid metabolism—are significantly altered in the brains of individuals with AD [10,11].

We identified two overlapping KEGG pathway classes between the CpGs identified in the $PS_{DNAm}$ and the metabolites identified in the $PS_{Metabolome}$ that are related to AD. Notably, both analyses identified key pathway classes such as "lipid metabolism" and "signal transduction," which play significant roles in AD pathology. For instance, the alpha linolenic acid and metabolism and linoleic acid metabolism have been previously linked with AD [59–61]. Alpha-linolenic acid is an essential omega-3 fatty acid known for its anti-inflammatory properties and its role in maintaining neuronal health. Studies have shown that an imbalance in fatty acid metabolism, including alpha-linolenic acid, is linked to neurodegenerative diseases such as AD [59,60]. Linoleic acid, on the other hand, is an essential omega-6 fatty acid whose dysregulation has also been associated with AD. It influences inflammatory processes and can lead to the formation of bioactive lipids that are important for neuronal health [61].

In addition to the two shared pathways, CpG sites selected by the $PS_{DNAm}$ were mapped to eight AD-related pathways that were unique for the DNAm data, including the pentose phosphate pathway [62], glycerophospholipid metabolism [63], ether lipids [64], arachidonic acid [65], linoleic acid [66], Ras signaling [67], the complement system [68] and impaired thermogenesis [69]. For the metabolomics data, ten unique pathways were identified with a known link to AD, including arginine metabolism and proline [70], sphingolipids [61], steroid hormones [71], cholesterol [72], C21 steroids [73], insulin signaling [74], adenosine A2A receptors [75], drug metabolism [76] and cytochrome P450 enzymes [77]. These unique pathways emphasize that different biological signatures of AD are detected by different omics layers.

Among the statistical methods used, there were clear differences in the performance for different omics data. For the DNAm data, PT was the best-performing method based on the $R^2$. For the metabolomics data, RF produced the highest $R^2$. While extensive simulation studies are needed to explain the observed differences in performance for different omics data, it could be due to the different number of variables in DNAm (789,286 CpG sites) versus metabolomics (35,348 variables) data or differences in the distribution, correlation structures, or underlying interactions

between variables. In comparison, EN and WA led to a substantially lower $R^2$ for both omics layers and EN also clearly underperformed in terms of computational time. BO showed a similar performance as EN, but future studies need to determine whether an improved hyperparameter optimization could increase the performance of this method.

In our study, we developed novel brain tissue–based multi-omics profile scores for AD neuropathology that integrate genome-wide DNAm and high-resolution metabolomics data, revealing that combining these omics layers modestly improves predictive accuracy compared to using either layer alone. This finding aligns with a growing body of literature exploring multi-modal prediction models for AD-related outcomes. Wang et al. [78] used unsupervised machine learning to develop an AD risk score that integrates cerebrospinal fluid, MRI, while Cary et al. [79]further advanced the field by integrating genetic, transcriptomic, and proteomic data to map AD risk onto core biological domains such as synapse function, immune response, and lipid metabolism. In a study by Liu et al. [80], comprehensive genetic prediction models were developed to compute risk scores for blood metabolites, which were subsequently employed in association analyses with Alzheimer's disease risk. Together, these studies emphasize that several omics layers are associated with AD-related outcomes, and they provide a valuable context for our results which show that the integration of these omics layers can offer deeper insights into the molecular underpinnings of AD.

There are several strengths of our study to be noted. The unique dataset includes both well-characterized DNAm and metabolomics data from 138 brain donors, allowing a detailed examination of brain-based multi-omics profiles related to AD neuropathology. Our study is characterized by methodological advances that enable a comprehensive analysis of DNAm and metabolomics. While previous studies often analyzed isolated omics data (e.g., only DNA methylation or only metabolomics), this analysis combines both types of data to obtain a holistic picture of biological processes. This opens new perspectives for understanding the interactions between epigenetic changes and metabolic dysfunctions. The calculation of PS from both datasets enables a differentiated view of the influence of various factors on AD neuropathology. Our PS are based on sophisticated machine learning techniques, such as RF, EN, and BO methods, which allow for robust feature selection even in high-dimensional datasets. By employing these innovative methodologies, our analysis not only identified CpG sites and/or metabolomic features predictive of AD neuropathology levels but also elucidated mechanistic underpinnings by evaluating enriched biological pathways among selected variables. Overall, these methodological improvements facilitate a deeper understanding of the intricate relationships between epigenetic modifications and metabolic changes in AD pathology.

In addition to its strengths, our study has some limitations that should be considered. One notable aspect is that the ADRC brain bank is enriched with AD patients and

other dementias, which makes the brain bank a convenience sample rather than a population-based one. This concentration of AD cases may reduce variability in neuropathology markers within the sample. Another consideration is the relatively small sample size when splitting our data into training and test sets to prevent overfitting. However, it is important to highlight that few studies have access to such a large autopsy sample, which is crucial for accurately measuring neuropathology markers as well as brain tissue-based DNAm and metabolomics data—this represents a significant strength of our work. While integrating DNAm and metabolomics has improved predictive power, further exploration is needed to understand how these epigenetic and metabolic changes interact and jointly influence AD pathology. The use of postmortem tissue samples also introduces some limitations; they may not fully capture dynamic metabolic and epigenetic changes throughout disease progression, potentially overlooking important temporal variations in biomarkers. Additionally, applying machine learning methods such as Pruning & Thresholding, Random Forests, Elastic Net, Boosting, and Sliding Windows can present challenges related to hyperparameter selection and optimization. While a general optimization of the hyperparameter was performed and used in each model, future work should evaluate whether these factors can be further optimized to improve the model performance and the robustness of analyses.

Future research directions should focus on expanding the sample size and diversity to validate the results and ensure their generalizability. Integrating additional levels of omics such as genomics, proteomics, transcriptomics, lipidomics, and microbiomics could provide a more comprehensive view of the pathophysiology of AD. Exploring interactions between these different layers could help to identify novel biomarkers and signaling pathways. In addition, functional studies are needed to confirm the biological relevance of the identified signaling pathways.

## Conclusion

The present research highlights the potential of integrating DNAm and metabolomics data to deepen our understanding of the pathophysiology of AD. Future research should focus on expanding and diversifying study populations and longitudinal designs to translate the multi-omics insights gained into clinical applications.

**References**


[1] Alzheimer`s Association (2022). 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*. https://doi.org/https://doi.org/10.1002/alz.12638.

[2] Centers for Disease Control and Prevention (2023). .

[3] Nandi, A. *et al.* (2024). Cost of care for Alzheimer's disease and related dementias in the United States: 2016 to 2060. *npj Aging*. https://doi.org/10.1038/s41514-024-00136-6.

[4] Long, J.M. and Holtzman, D.M. (2019). Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell*. https://doi.org/10.1016/j.cell.2019.09.001.

[5] Selkoe, D.J. (2002). Alzheimer's Disease Is a Synaptic Failure. *Science*. https://doi.org/10.1126/science.1074069.

[6] Mathys, H. *et al.* (2019). Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*. https://doi.org/10.1038/s41586-019-1195-2.

[7] Dai, X. and Shen, L. (2022). Advances and Trends in Omics Technology Development. *Frontiers in Medicine*. https://doi.org/10.3389/fmed.2022.911861.

[8] Wilkins, J.M. and Trushina, E. (2018). Application of metabolomics in Alzheimer's disease. *Frontiers in Neurology*. https://doi.org/10.3389/fneur.2017.00719.

[9] Yuan, Y. *et al.* (2025). Dysregulation of energy metabolism in Alzheimer's disease. *Journal of Neurology*. https://doi.org/10.1007/s00415-024-12800-8.

[10] Cleland, N.R.W. *et al.* (2021). Altered substrate metabolism in neurodegenerative disease: new insights from metabolic imaging. *Journal of Neuroinflammation*. https://doi.org/10.1186/s12974-021-02305-w.

[11] Korczowska-Łącka, I. *et al.* (2023). Selected Biomarkers of Oxidative Stress and Energy Metabolism Disorders in Neurological Diseases. *Molecular Neurobiology*. https://doi.org/10.1007/s12035-023-03329-4.

[12] Zeng, Y. *et al.* (2023). Identification of key lipid metabolism-related genes in Alzheimer's disease. *Lipids in Health and Disease*. https://doi.org/10.1186/s12944-023-01918-9.

[13] Besser, L.M. *et al.* (2018). The revised national Alzheimer's coordinating center's neuropathology form-available data and new analyses. *Journal of Neuropathology and Experimental Neurology*. https://doi.org/10.1093/jnen/nly049.

[14] Montine, T.J. *et al.* (2012). National Institute on aging-Alzheimer's association guidelines for the neuropathologic assessment of Alzheimer's disease: A practical approach. *Acta Neuropathologica*. https://doi.org/10.1007/s00401-011-0910-3.

[15] Deture, M.A. and Dickson, D.W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration*. https://doi.org/10.1186/s13024-019-0333-5.

[16] Braak, H. and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol*.

[17] Mirra, S.S. *et al.* (1991). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). *Neurology*. https://doi.org/10.1212/WNL.41.4.479.

[18] Thal, D.R. *et al.* (2002). Phases of Aβ-deposition in the human brain and its relevance for the development of AD. *Neurology*. https://doi.org/10.1212/WNL.58.12.1791.

[19] R Core Team (2021). R: A Language and Environment for Statistical Computing.

[20] Pett, L. *et al.* (2024). The association between neighborhood deprivation and DNA methylation in an autopsy cohort. *Aging*. https://doi.org/10.18632/aging.205764.

[21] Johnson, W.E. *et al.* (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxj037.

[22] Aryee, M.J. *et al.* (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu049.

[23] Guintivano, J. *et al.* (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*. https://doi.org/10.4161/epi.23924.

[24] Maksimovic, J. *et al.* (2012). SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biology*.

[25] Fortin, J.-P. and Hansen, K.D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*. https://doi.org/10.1186/s13059-015-0741-y.

[26] Fortin, J.-P. *et al.* (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw691.

[27] Fortin, J.-P. *et al.* (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*. https://doi.org/10.1186/s13059-014-0503-2.

[28] Triche, T.J. *et al.* (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt090.

[29] Andrews, S. V. *et al.* (2016). “Gap hunting” to characterize clustered probe signals in Illumina methylation array data. *Epigenetics & Chromatin*. https://doi.org/10.1186/s13072-016-0107-z.

[30] Go, Y.M. *et al.* (2015). Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicological Sciences*. https://doi.org/10.1093/toxsci/kfv198.

[31] Ladva, C.N. *et al.* (2018). Particulate metal exposures induce plasma metabolome changes in a commuter panel study. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0203468.

[32] Liang, D. *et al.* (2019). Perturbations of the arginine metabolome following exposures to traffic-related air pollution in a panel of commuters with and without asthma. *Environment International*. https://doi.org/10.1016/j.envint.2019.04.003.

[33] Liang, D. *et al.* (2018). Use of high-resolution metabolomics for the identification of metabolic signals associated with traffic-related air pollution. *Environment International*. https://doi.org/10.1016/j.envint.2018.07.044.

[34] Ribbenstedt, A. *et al.* (2018). Development, characterization and comparisons of targeted and non-targeted metabolomics methods. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0207082.

[35] Uppal, K. *et al.* (2013). xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-15.

[36] Yu, T. *et al.* (2009). apLCMS—adaptive processing of high-resolution LC/MS data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp291.

[37] Kind, A.J.H. and Buckingham, W.R. (2018). Making Neighborhood-Disadvantage Metrics Accessible — The Neighborhood Atlas. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMp1802313.

[38] Choi, S.W. *et al.* (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*. https://doi.org/10.1038/s41596-020-0353-1.

[39] Mayer, M. (2024). missRanger: Fast Imputation of Missing Values.

[40] Everitt, B.S. *et al.* (2011). Cluster Analysis 5th Edition Cluster Analysis 5th Edition WILEY SERIES IN PROBABILITY AND STATISTICS Cluster Analysis 5th Edition.

[41] Coombes, K.R. (2024). ClassDiscovery: Classes and Methods for “Class Discovery” with Microarrays or Proteomics.

[42] Langfelder, P. and Horvath, S. (2012). Fast *R* Functions for Robust Correlations and Hierarchical Clustering. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v046.i11.

[43] Maechler, M. *et al.* (2023). cluster: Cluster Analysis Basics and Extensions.

[44] Wurm, M.J. *et al.* (2021). Regularized Ordinal Regression and the ordinalNet R Package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v099.i06.

[45] Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*. https://doi.org/10.1214/07-STS242.

[46] Lau, M. *et al.* (2024). logicDT: a procedure for identifying response-associated interactions between binary predictors. *Machine Learning*. https://doi.org/10.1007/s10994-023-06488-6.

[47] Wright, M.N. and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v077.i01.

[48] Janitza, S. *et al.* (2015). A computationally fast variable importance test for random forests for high-dimensional data.

[49] Teschke, S. *et al.* (2024). Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores. *Biometrical Journal*. https://doi.org/10.1002/bimj.70014.

[50] McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*.

[51] Simon, N. *et al.* (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2012.681250.

[52] Phipson, B. *et al.* (2015). missMethyl: an R package for analysing methylation data from Illuminas HumanMethylation450 platform. *Bioinformatics*.

[53] Maksimovic, J. *et al.* (2015). Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic acids research*.

[54] Phipson, B. and Oshlack, A. (2014). DiffVar: A new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biology*. https://doi.org/https://doi.org/10.1186/s13059-014-0465-4.

[55] Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/28.1.27.

[56] Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Science*. https://doi.org/10.1002/pro.3715.

[57] Kanehisa, M. *et al.* (2025). KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkae909.

[58] Tian, L. and Yu, T. (2024). metapone: Conducts pathway test of metabolomics data using a weighted permutation test. https://doi.org/10.18129/B9.bioc.metapone.

[59] Grimm, M.O.W. *et al.* (2017). Omega-3 fatty acids, lipids, and apoE lipidation in Alzheimer's disease: a rationale for multi-nutrient dementia prevention. *Journal of Lipid Research*. https://doi.org/10.1194/jlr.R076331.

[60] Penke, B. *et al.* (2018). The Role of Lipids and Membranes in the Pathogenesis of Alzheimer's Disease: A Comprehensive View. *Current Alzheimer Research*. https://doi.org/10.2174/1567205015666180911151716.

[61] He, X. *et al.* (2010). Deregulation of sphingolipid metabolism in Alzheimer's disease. *Neurobiology of Aging*. https://doi.org/10.1016/j.neurobiolaging.2008.05.010.

[62] Butterfield, D.A. *et al.* (2012). Redox Proteomics in Selected Neurodegenerative Disorders: From its Infancy to Future Applications. *Antioxidants & Redox Signaling*. https://doi.org/10.1089/ars.2011.4109.

[63] Yin, F. (2023). Lipid metabolism and Alzheimer's disease: clinical evidence, mechanistic link and therapeutic promise. *The FEBS Journal*. https://doi.org/10.1111/febs.16344.

[64] Fabelo, N. *et al.* (2014). Altered lipid composition in cortical lipid rafts occurs at early stages of sporadic Alzheimer's disease and facilitates APP/BACE1 interactions. *Neurobiology of Aging*. https://doi.org/10.1016/j.neurobiolaging.2014.02.005.

[65] Olivier, J.-L. (2016). Arachidonic acid in Alzheimer's disease. *Journal of Neurology and Neuromedicine*. https://doi.org/10.29245/2572.942X/2016/9.1086.

[66] Grimm, M. *et al.* (2013). Effect of Different Phospholipids on α-Secretase Activity in the Non-Amyloidogenic Pathway of Alzheimer's Disease. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms14035879.

[67] Kumari, S. *et al.* (2023). Apoptosis in Alzheimer's disease: insight into the signaling pathways and therapeutic avenues. *Apoptosis*. https://doi.org/10.1007/s10495-023-01848-y.

[68] Tenner, A.J. (2020). Complement-Mediated Events in Alzheimer's Disease: Mechanisms and Potential Therapeutic Targets. *The Journal of Immunology*. https://doi.org/10.4049/jimmunol.1901068.

[69] Clarke, J.R. *et al.* (2018). Metabolic Dysfunction in Alzheimer's Disease: From Basic Neurobiology to Clinical Approaches. *Journal of Alzheimer's Disease*. https://doi.org/10.3233/JAD-179911.

[70] Kan, M.J. *et al.* (2015). Arginine Deprivation and Immune Suppression in a Mouse Model of Alzheimer's Disease. *The Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.4668-14.2015.

[71] Pike, C.J. *et al.* (2009). Protective actions of sex steroid hormones in Alzheimer's disease. *Frontiers in Neuroendocrinology*. https://doi.org/10.1016/j.yfrne.2009.04.015.

[72] Chang, T.-Y. *et al.* (2017). Cellular cholesterol homeostasis and Alzheimer's disease. *Journal of Lipid Research*. https://doi.org/10.1194/jlr.R075630.

[73] Vaňková, M. *et al.* (2023). The Role of Steroidomics in the Diagnosis of Alzheimer's Disease and Type 2 Diabetes Mellitus. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms24108575.

[74] De la Monte, S.M. (2012). Brain Insulin Resistance and Deficiency as Therapeutic Targets in Alzheimers Disease. *Current Alzheimer Research*. https://doi.org/10.2174/156720512799015037.

[75] Cunha, R.A. (2008). Different cellular sources and different roles of adenosine: A1 receptor-mediated inhibition through astrocytic-driven volume transmission and synapse-restricted A2A receptor-mediated facilitation of plasticity. *Neurochemistry International*. https://doi.org/10.1016/j.neuint.2007.06.026.

[76] Mangialasche, F. *et al.* (2010). Alzheimer's disease: clinical trials and drug development. *The Lancet Neurology*. https://doi.org/10.1016/S1474-4422(10)70119-8.

[77] Bahado-Singh, R.O. *et al.* (2023). Alzheimer's Precision Neurology: Epigenetics of Cytochrome P450 Genes in Circulating Cell-Free DNA for Disease Prediction and Mechanism. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms24032876.

[78] Wang, Z. *et al.* (2020). AD risk score for the early phases of disease based on unsupervised machine learning. *Alzheimer's and Dementia*. https://doi.org/10.1002/alz.12140.

[79] Cary, G.A. *et al.* (2024). Genetic and multi-omic risk assessment of Alzheimer's disease implicates core associated biological domains. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*. https://doi.org/10.1002/trc2.12461.

[80] Liu, S. *et al*. (2024). Identification of blood metabolites associated with risk of Alzheimer's disease by integrating genomics and metabolomics data. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-023-02400-9.

## ACKNOWLEDGMENTS

T.T. was supported by the Research Training Group “Biostatistical Methods for High-Dimensional Data in Toxicology” (RTG 2624, Project I1) funded by the German Research Foundation (DFG, Project Number 427806116). This work was supported by the HERCULES Pilot Project via NIEHS P30ES019776 (Huels), the Goizueta Alzheimer’s Disease Research Center: Pilot Grant via NIA P30 AG055611 (Huels/Liang), the Rollins School of Public Health Dean’s Pilot and Innovation Grant (Huels), NIEHS R21ES032117 (Liang), R01ES035738 (Liang), NIA R01AG079170 (Huels/Wingo), U01AG088425 (Huels/Liang/Wingo), and R01AG087250 (Huels/Liang).

## CONFLICT OF INTEREST STATEMENT

The authors report no competing interests.

## CONSENT STATEMENT

All relevant ethical guidelines have been followed, and any necessary IRB and/or ethics committee approvals have been obtained. Written informed consent was obtained from all participants before inclusion in the study.

## KEYWORDS

Alzheimer’s disease, multi-omics, profile scores, DNA methylation, metabolomics, neuropathology, machine learning

## Figures and tables

**Table 1. Characteristics of the study population.**

| | | Total (N=157) | DNAm (N=154) | Metabolomics (N=141) | Multi-omics analyses (N=138) |
|---|---|---|---|---|---|
| **Age at Death** | | | | | |
| Mean (SD) | | 76.4 (10.0) | 76.4 (10.0) | 76.7 (10.2) | 76.7 (10.2) |
| Range | | 57.0 - 105.0 | 57.0 - 105.0 | 57.0 - 105.0 | 57.0 - 105.0 |
| **Race** | | | | | |
| White | | 140 (89.2%) | 137 (89.0%) | 124 (87.9%) | 121 (87.7%) |
| Black | | 17 (10.8%) | 17 (11.0%) | 17 (12.1%) | 17 (12.3%) |
| **Sex** | | | | | |
| Female | | 71 (45.2%) | 70 (45.5%) | 61 (43.3%) | 60 (43.5%) |
| Male | | 86 (54.8%) | 84 (54.5%) | 80 (56.7%) | 78 (56.5%) |
| **Post Mortal Index (PMI) (in hours)** | | | | | |
| Mean (SD) | | 11.6 (9.6) | 11.6 (9.6) | 11.5 (9.7) | 11.5 (9.7) |
| Range | | 1.5 - 64.0 | 1.5 - 64.0 | 1.5 - 64.0 | 1.5 - 64.0 |
| **Education (cat.)** | | | | | |
| High school or less | | 35 (22.3%) | 35 (22.7%) | 29 (20.6%) | 29 (21.0%) |
| College degree | | 78 (49.7%) | 76 (49.4%) | 70 (49.6%) | 68 (49.3%) |
| Graduate degree | | 44 (28.0%) | 43 (27.9%) | 42 (29.8%) | 41 (29.7%) |
| **Area Deprivation Index (ADI)** | | | | | |
| Mean (SD) | | 36.1 (24.0) | 36.3 (24.1) | 34.4 (23.7) | 34.6 (23.9) |
| Range | | 1.0 - 94.0 | 1.0 - 94.0 | 1.0 - 94.0 | 1.0 - 94.0 |
| **APOE (yes/no)** | | | | | |
| E4 absent | | 69 (43.9%) | 69 (44.8%) | 60 (42.6%) | 60 (43.5%) |
| E4 present | | 88 (56.1%) | 85 (55.2%) | 81 (57.4%) | 78 (56.5%) |
| **ABC score** | | | | | |
| Not | | 15 (9.6%) | 15 (9.7%) | 13 (9.2%) | 13 (9.4%) |
| Low | | 28 (17.8%) | 28 (18.2%) | 24 (17.0%) | 24 (17.4%) |

| | | | | | |
|---|---|---|---|---|---|
| Intermediate | | 21 (13.4%) | 20 (13.0%) | 20 (14.2%) | 19 (13.8%) |
| High | | 93 (59.2%) | 91 (59.1%) | 84 (59.6%) | 82 (59.4%) |
| **Braak Stage** | | | | | |
| Stage 0 | | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| Stage 1 | | 16 (10.2%) | 16 (10.4%) | 15 (10.6%) | 15 (10.9%) |
| Stage 2 | | 11 (7.0%) | 11 (7.1%) | 10 (7.1%) | 10 (7.2%) |
| Stage 3 | | 17 (10.8%) | 17 (11.0%) | 13 (9.2%) | 13 (9.4%) |
| Stage 4 | | 18 (11.5%) | 17 (11.0%) | 17 (12.1%) | 16 (11.6%) |
| Stage 5 | | 21 (13.4%) | 21 (13.6%) | 19 (13.5%) | 19 (13.8%) |
| Stage 6 | | 74 (47.1%) | 72 (46.8%) | 67 (47.5%) | 65 (47.1%) |
| **CERAD score** | | | | | |
| No | | 34 (21.7%) | 34 (22.1%) | 30 (21.3%) | 30 (21.7%) |
| Sparse | | 3 (1.9%) | 3 (1.9%) | 3 (2.1%) | 3 (2.2%) |
| Moderate | | 10 (6.4%) | 10 (6.5%) | 9 (6.4%) | 9 (6.5%) |
| Frequent | | 110 (70.1%) | 107 (69.5%) | 99 (70.2%) | 96 (69.6%) |

Donors included in the column “total” had information on all covariates, neuropathology outcomes and either DNAm or metabolomics data available. Of these, 154 donors (listed in the column “DNAm” had DNAm available and 141 donors (listed in the column “metabolomics”) had metabolomics data available. Donors included in the column “multi-omics analyses” had data on DNAm and metabolomics and were included in the multi-omics analyses.

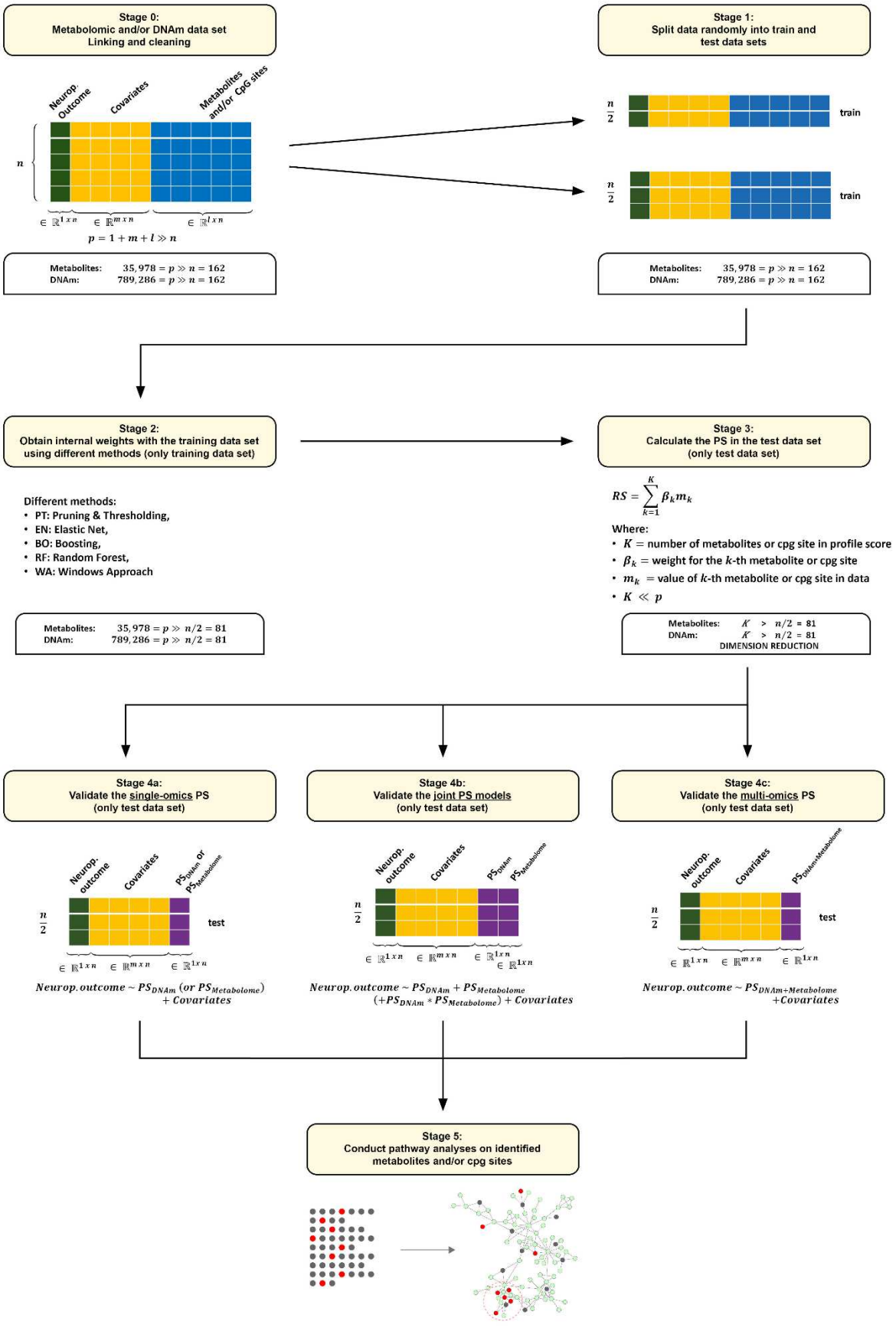

Stage 0:
Metabolomic and/or DNAm data set
Linking and cleaning
Neurop. Outcome
Covariates
Metabolites and/or CpG sites
$n$
$\in \mathbb{R}^{1\,x\,n}$
$\in \mathbb{R}^{m\,x\,n}$
$\in \mathbb{R}^{l\,x\,n}$
$p = 1 + m + l \gg n$
Metabolites: $35,978 = p \gg n = 162$
DNAm: $789,286 = p \gg n = 162$
Stage 1:
Split data randomly into train and test data sets
$\frac{n}{2}$
train
$\frac{n}{2}$
train
Metabolites: $35,978 = p \gg n = 162$
DNAm: $789,286 = p \gg n = 162$
Stage 2:
Obtain internal weights with the training data set using different methods (only training data set)
Different methods:
• PT: Pruning & Thresholding,
• EN: Elastic Net,
• BO: Boosting,
• RF: Random Forest,
• WA: Windows Approach
Metabolites: $35,978 = p \gg n/2 = 81$
DNAm: $789,286 = p \gg n/2 = 81$
Stage 3:
Calculate the PS in the test data set
(only test data set)
$RS = \sum_{k=1}^{K} \beta_k m_k$
Where:
• $K$ = number of metabolites or cpg site in profile score
• $\beta_k$ = weight for the $k$-th metabolite or cpg site
• $m_k$ = value of $k$-th metabolite or cpg site in data
• $K \ll p$
Metabolites: $K > n/2 = 81$
DNAm: $K > n/2 = 81$
DIMENSION REDUCTION
Stage 4a:
Validate the single-omics PS
(only test data set)
Neurop. outcome
Covariates
$PS_{DNAm}$ or $PS_{Metabolome}$
$\frac{n}{2}$
test
$\in \mathbb{R}^{1\,x\,n}$
$\in \mathbb{R}^{m\,x\,n}$
$\in \mathbb{R}^{1\,x\,n}$
$Neurop.outcome \sim PS_{DNAm}\ (or\ PS_{Metabolome}) + Covariates$
Stage 4b:
Validate the joint PS models
(only test data set)
Neurop. outcome
Covariates
$PS_{DNAm}$
$PS_{Metabolome}$
$\frac{n}{2}$
$\in \mathbb{R}^{1\,x\,n}$
$\in \mathbb{R}^{m\,x\,n}$
$\in \mathbb{R}^{1\,x\,n}$
$\in \mathbb{R}^{1x\,n}$
$Neurop.outcome \sim PS_{DNAm} + PS_{Metabolome} (+PS_{DNAm} * PS_{Metabolome}) + Covariates$
Stage 4c:
Validate the multi-omics PS
(only test data set)
Neurop. outcome
Covariates
$PS_{DNAm+Metabolome}$
$\frac{n}{2}$
test
$\in \mathbb{R}^{1\,x\,n}$
$\in \mathbb{R}^{m\,x\,n}$
$\in \mathbb{R}^{1\,x\,n}$
$Neurop.outcome \sim PS_{DNAm+Metabolome} + Covariates$
Stage 5:
Conduct pathway analyses on identified metabolites and/or cpg sites

**Figure 1.** Overview of the statistical procedure for determining the weights and for calculating and validating the individual PS. First, the respective data are prepared and linked to relevant demographic covariates (age at death, race, gender, Post Mortal Index, Education and Area Deprivation Index) and outcomes (neuropathologic scores such as ABC score, Braak Stage and CERAD score). Then the data is split into training and test data. Various methods are used to determine the required internal weights based on the training data set. The following methods are used here: PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and combinations of these methods. The PS is now determined on the test data by means of the specific weighting of the respective PS. We validated the PS by regression against the outcomes for the single-omics PS (for each data set) and the multi-omics PS (for both data sets in one model optional with an interaction term). Finally, an optional metabolic pathway analysis using the cpg-sites (DNAm) or features (metabolomics) identified by the respective PS can be performed to evaluate relevant biological pathways.

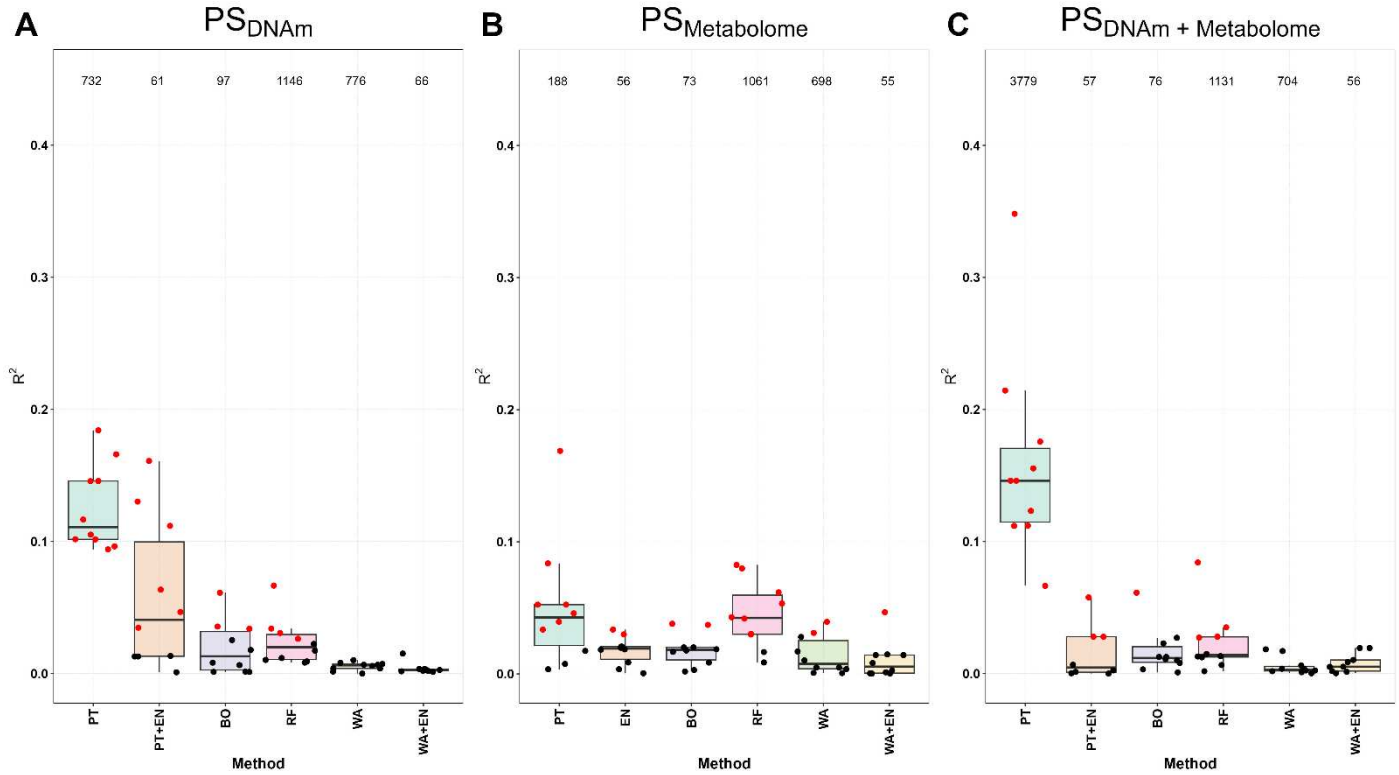



**Figure 2.** Overview of the results of PS calculation using the various methods and their accuracy of fitting from the individual and combined data sets (McFadden $R^2$ and p-value) on ABC score outcome. All three sub-graphs are structured in the same way and therefore the explanation can be made on one graph and apply to all three graphs: A) DNAm data set (single-omics PS), B) metabolome data set (single-omics PS) and C) combination of both individual data sets (multi-omics PS). The following 6 methods are shown on the x-axis: PT: Pruning & Thresholding, EN: Elastic Net or EN+PT, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN (due to the high dimension, EN can only be applied to the smaller data set and then PT is used before for dimension reduction). McFadden $R^2$ is shown on the y-axis, so that for each method a boxplot is shown for the 10 iterations, with the individual results shown as black (not significant) and red (significant) points. At the top are the recorded number of features or cpg-sites included in the PS (mean value across the 10 iterations).


**Figure 3.** Results from single- and multi-omics PS each individual graph shows the $R^2$ for the PS of the individual and combined datasets, with the method from the DNAm dataset shown on the x-axis and the metabolome dataset on the y-axis (ABC score). The methods are PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN. McFadden R2 is shown

on the y-axis, so that for each method the median with standard error (SE) is drawn as bar plot with SE bars. The values in a subgraph are from the following data sets or combinations: Joint PS models with interaction term (green), Joint PS models without interaction term (orange) and multi-omics PS from both combined data sets (golden). Since the single PS are at most as good (median) as the joint PS, we have omitted these for better consideration and drawn only the PS with the highest value as a line in the subgraph. If the single PS from DNAm data set is higher than the PS from the metabolome data set, we took the value of the DNAm PS and drew a blue line (vice versa for the single PS from metabolome data set purple). Due to the calculation, the golden boxplots are only present in the subgraphs that use the same methods in both data sets.

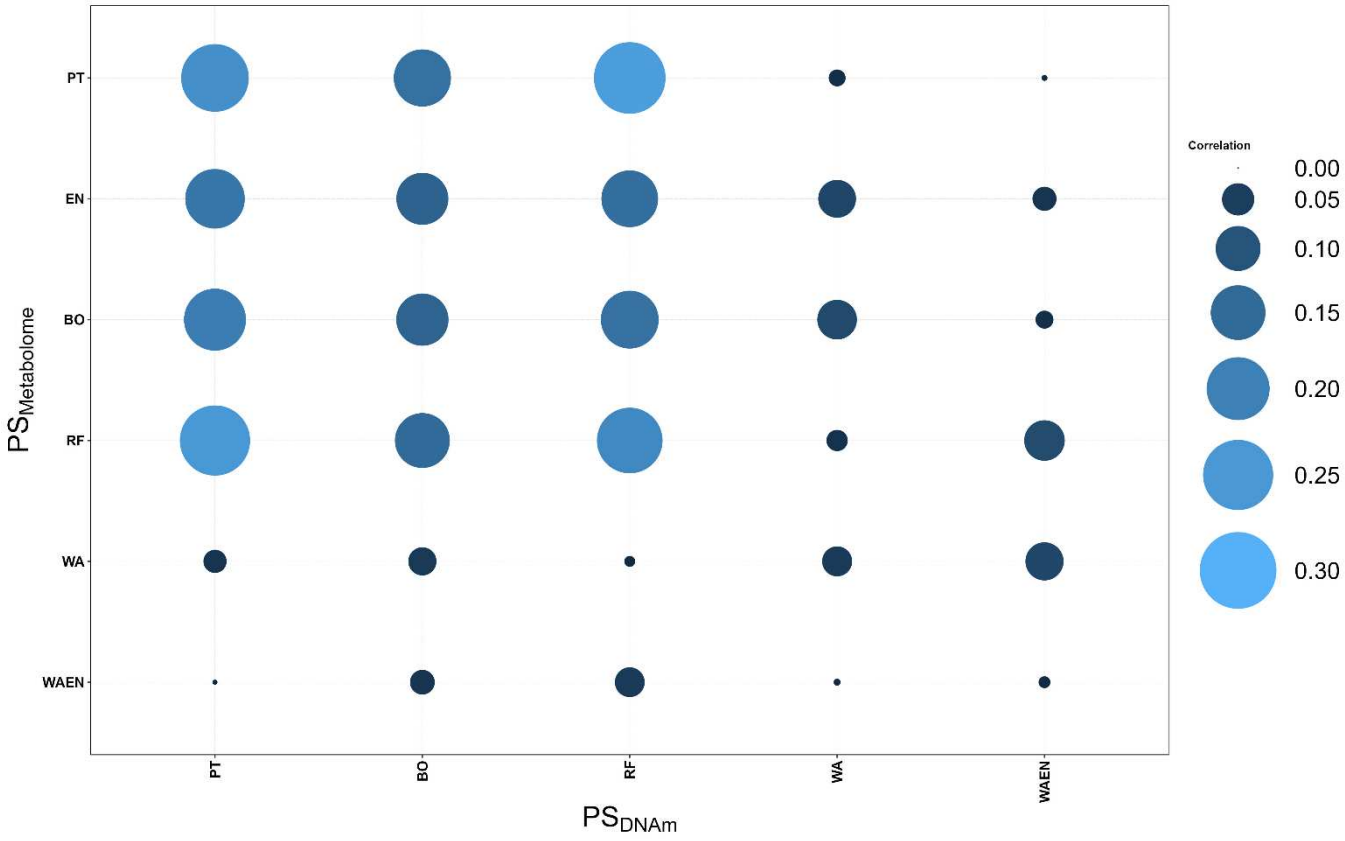



**Figure 4**. Pearson Correlation structure between the different methods and PS of each data set. On the x-axis are the PS of the different methods based on DNAm data set and on the y-axis the PS of the different methods based on the metabolome data set. The methods are PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN. The color intensity and the size of the dots indicate the respective correlation coefficient according to Bravais-Pearson.

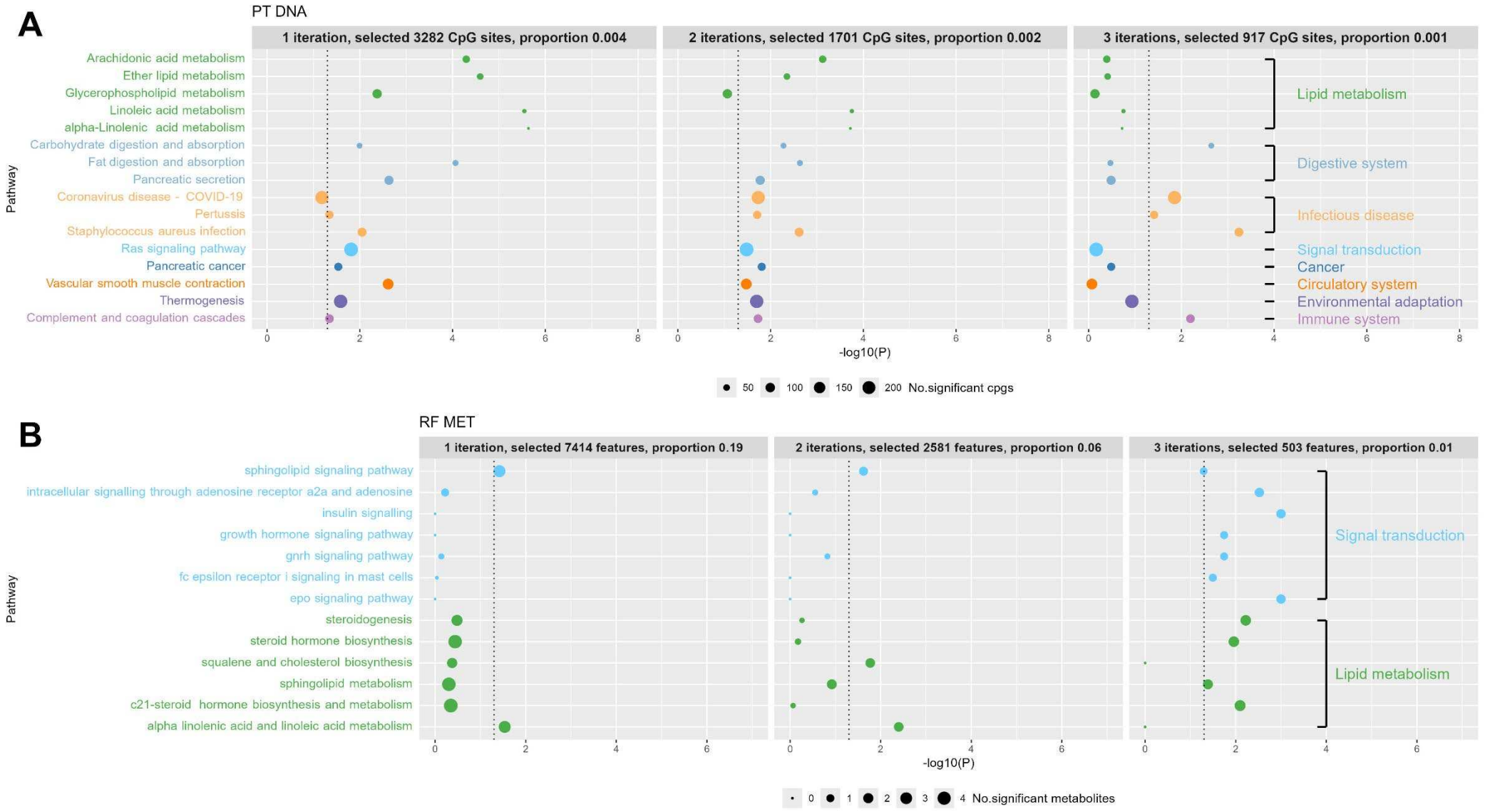

A
PT DNA
1 iteration, selected 3282 CpG sites, proportion 0.004
2 iterations, selected 1701 CpG sites, proportion 0.002
3 iterations, selected 917 CpG sites, proportion 0.001
Pathway
Arachidonic acid metabolism
Ether lipid metabolism
Glycerophospholipid metabolism
Linoleic acid metabolism
alpha-Linolenic acid metabolism
Carbohydrate digestion and absorption
Fat digestion and absorption
Pancreatic secretion
Coronavirus disease - COVID-19
Pertussis
Staphylococcus aureus infection
Ras signaling pathway
Pancreatic cancer
Vascular smooth muscle contraction
Thermogenesis
Complement and coagulation cascades
Lipid metabolism
Digestive system
Infectious disease
Signal transduction
Cancer
Circulatory system
Environmental adaptation
Immune system
-log10(P)
50 100 150 200 No.significant cpgs
B
RF MET
1 iteration, selected 7414 features, proportion 0.19
2 iterations, selected 2581 features, proportion 0.06
3 iterations, selected 503 features, proportion 0.01
Pathway
sphingolipid signaling pathway
intracellular signalling through adenosine receptor a2a and adenosine
insulin signalling
growth hormone signaling pathway
gnrh signaling pathway
fc epsilon receptor i signaling in mast cells
epo signaling pathway
steroidogenesis
steroid hormone biosynthesis
squalene and cholesterol biosynthesis
sphingolipid metabolism
c21-steroid hormone biosynthesis and metabolism
alpha linolenic acid and linoleic acid metabolism
Signal transduction
Lipid metabolism
-log10(P)
0 1 2 3 4 No.significant metabolites

**Figure 5**. For the DNAm data set (A) with the PT method scatter plots of KEGG pathway enrichment analysis, where at least in (1), (2) or (3) the weighted CpG-sites were found. For the Metabolome data set (B) with the RF method scatter plots of KEGG pathway enrichment analysis, where at least in (1), (2) or (3) iterations the weighted features were found. Only pathways classes, which had at least one pathway with significant p-values in more than one iteration. The number of significant CpG-Sites/ features in the pathway is indicated by the circle area, and the circle color represents the predefined pathway classes. On the x-axis the p-value (as $-\log_{10}(p)$) is represented with the significance level of 0.05 (dotted line). The proportion refers to the number of selected variables divided by the total number of variables (789,286 for DNAm; 35,978 for metabolomics). We display on the y-axis the different pathway terms enriched by KEGG database; the single pathway stands on the y-axis and the classes of them in the last plot. The following pathways classes are not presented in the figure because the corresponding pathway class was only significant in one iteration: A) For DNAm: Carbohydrate metabolism (pentose phosphate pathway), Metabolism of terpenoids and polyketides (terpenoid backbone biosynthesis), Information processing in viruses (virion – flavivirus and alphavirus) and Infection disease: parasitic (Chagas disease); B) For metabolome: Cancer (pathways in cancer), Endocrine system (ovarian steroidogenesis), Infection disease: parasitic (African trypanosomiasis), Biosynthesis of other secondary metabolites (pterine biosynthesis), Digestive system (mineral absorption), Amino acid metabolism (arginine and proline metabolism) and Xenobiotics biodegradation and metabolism (metabolism of xenobiotics by cytochrome p450; drug metabolism – other enzymes)

## Supplement Files:

### A) Additional figures

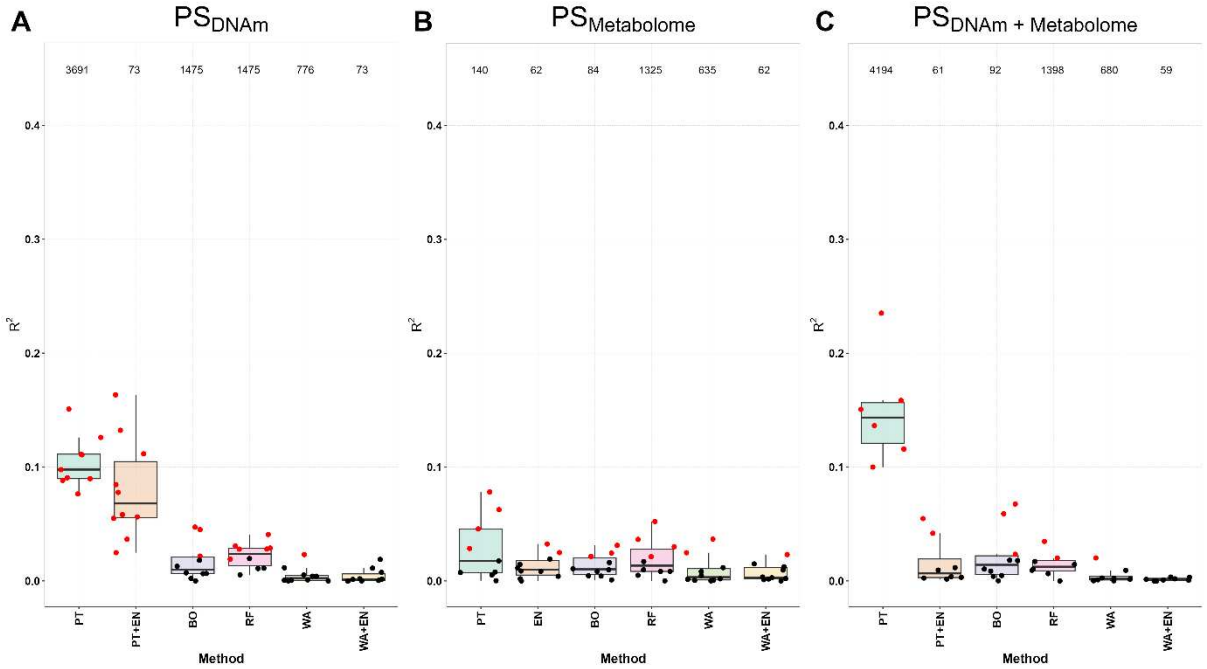



**Figure S1.** Overview of the results of PS calculation using the various methods and their accuracy of fitting from the individual and combined data sets (McFadden $R^2$ and p-value) on Braak stage outcome. All three sub-graphs are structured in the same way and therefore the explanation can be made on one graph and apply to all three graphs: A) DNAm data set (single-omics PS), B) ) metabolome data set (single-omics PS) and C) combination of both individual data sets (multi-omics PS). The following 6 methods are shown on the x-axis: PT: Pruning & Thresholding, EN: Elastic Net or EN+PT, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN (due to the high dimension, EN can only be applied to the smaller data set and then PT is used before for dimension reduction). Partial McFadden $R^2$ is shown on the y-axis, so that for each method a boxplot is shown for the 10 iterations, with the individual results shown as black (not

significant) and red (significant) points. At the top are the recorded number of features or cpg-sites included in the PS (mean value across the 10 iterations).


**Figure S3.** Results from single- and multi-omics PS each individual graph shows the $R^2$ for the PS of the individual and combined datasets, with the method from the DNAm dataset shown on the x-axis and the metabolome dataset on the y-axis (Braak stage). The methods are PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN. McFadden R2 is shown on the y-axis, so that for each method the median with standard error (SE) is drawn as bar plot with SE bars. The values in a subgraph are from the following data sets or combinations: Joint PS models with interaction term (green), Joint PS models without interaction term (orange) and multi-omics PS from both combined data sets (golden). Since the single PS are at most as good (median) as the joint PS, we have omitted these for better consideration and drawn only the PS with the highest value as a line in the subgraph. If the single PS from

DNAm data set is higher than the PS from the metabolome data set, we took the value of the DNAm PS and drew a blue line (vice versa for the single PS from metabolome data set purple). Due to the calculation, the golden boxplots are only present in the subgraphs that use the same methods in both data sets.

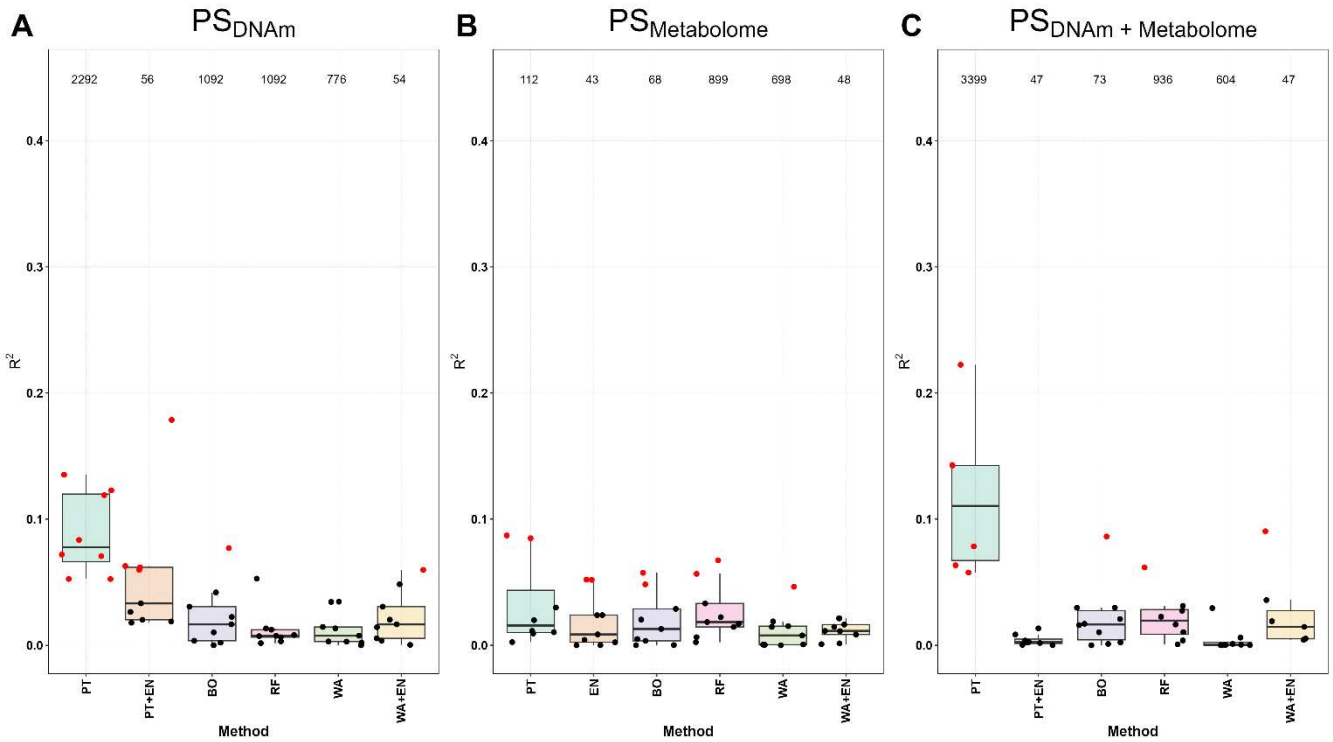



**Figure S2.** Overview of the results of PS calculation using the various methods and their accuracy of fitting from the individual and combined data sets (McFadden $R^2$ and p-value) on CERAD outcome. All three sub-graphs are structured in the same way and therefore the explanation can be made on one graph and apply to all three graphs: A) DNAm data set (single-omics PS), B) metabolome data set (single-omics PS) and C) combination of both individual data sets (multi-omics PS). The following 6 methods are shown on the x-axis: PT: Pruning &

Thresholding, EN: Elastic Net or EN+PT, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN (due to the high dimension, EN can only be applied to the smaller data set and then PT is used before for dimension reduction). McFadden $R^2$ is shown on the y-axis, so that for each method a boxplot is shown for the 10 iterations, with the individual results shown as black (not significant) and red (significant) points. At the top are the recorded number of features or cpg-sites included in the PS (mean value across the 10 iterations).

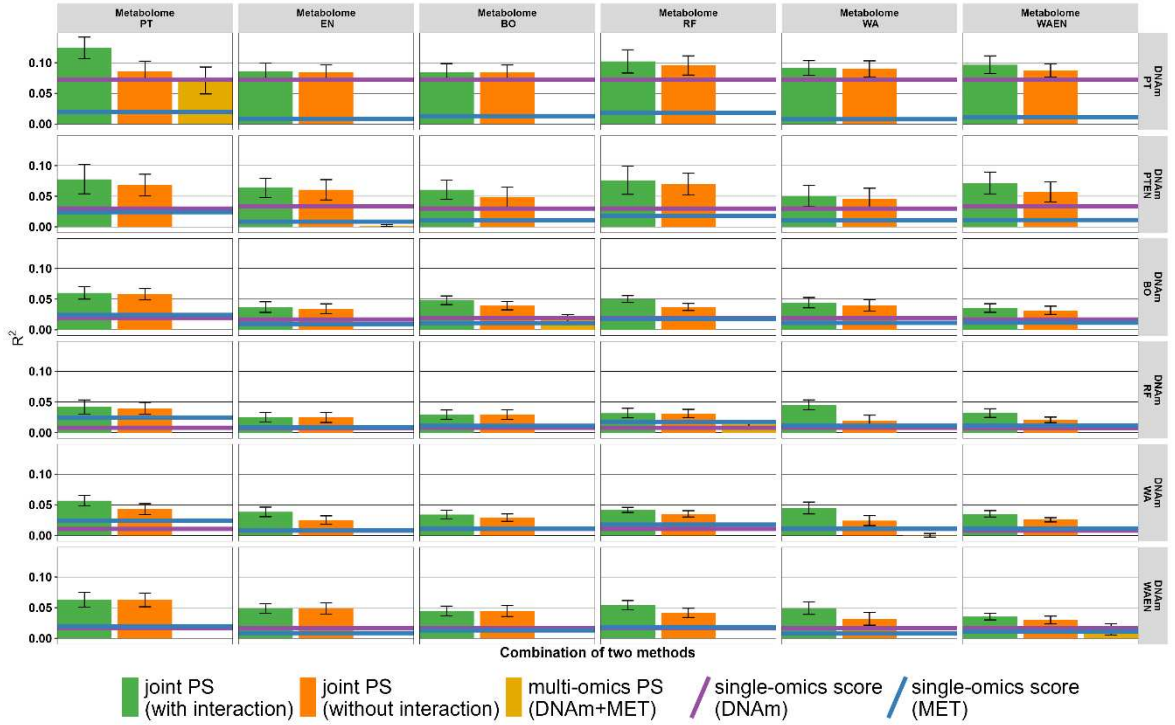



**Figure S4.** Results from single- and multi-omics PS each individual graph shows the $R^2$ for the PS of the individual and combined datasets, with the method from the DNAm dataset shown on the x-axis and the metabolome dataset on the y-axis (CERAD score). The methods are PT: Pruning & Thresholding, EN: Elastic Net, BO: Boosting, RF: Random Forest, WA: Windows Approach and WA+EN. McFadden R2 is shown on the y-axis, so that for each method the median with standard error (SE) is drawn as bar plot with SE bars. The values in a subgraph are from the following data sets or combinations: Joint PS models with interaction term (green), Joint PS models without interaction term

(orange) and multi-omics PS from both combined data sets (golden). Since the single PS are at most as good (median) as the joint PS, we have omitted these for better consideration and drawn only the PS with the highest value as a line in the subgraph. If the single PS from DNAm data set is higher than the PS from the metabolome data set, we took the value of the DNAm PS and drew a blue line (vice versa for the single PS from metabolome data set purple). Due to the calculation, the golden boxplots are only present in the subgraphs that use the same methods in both data sets.

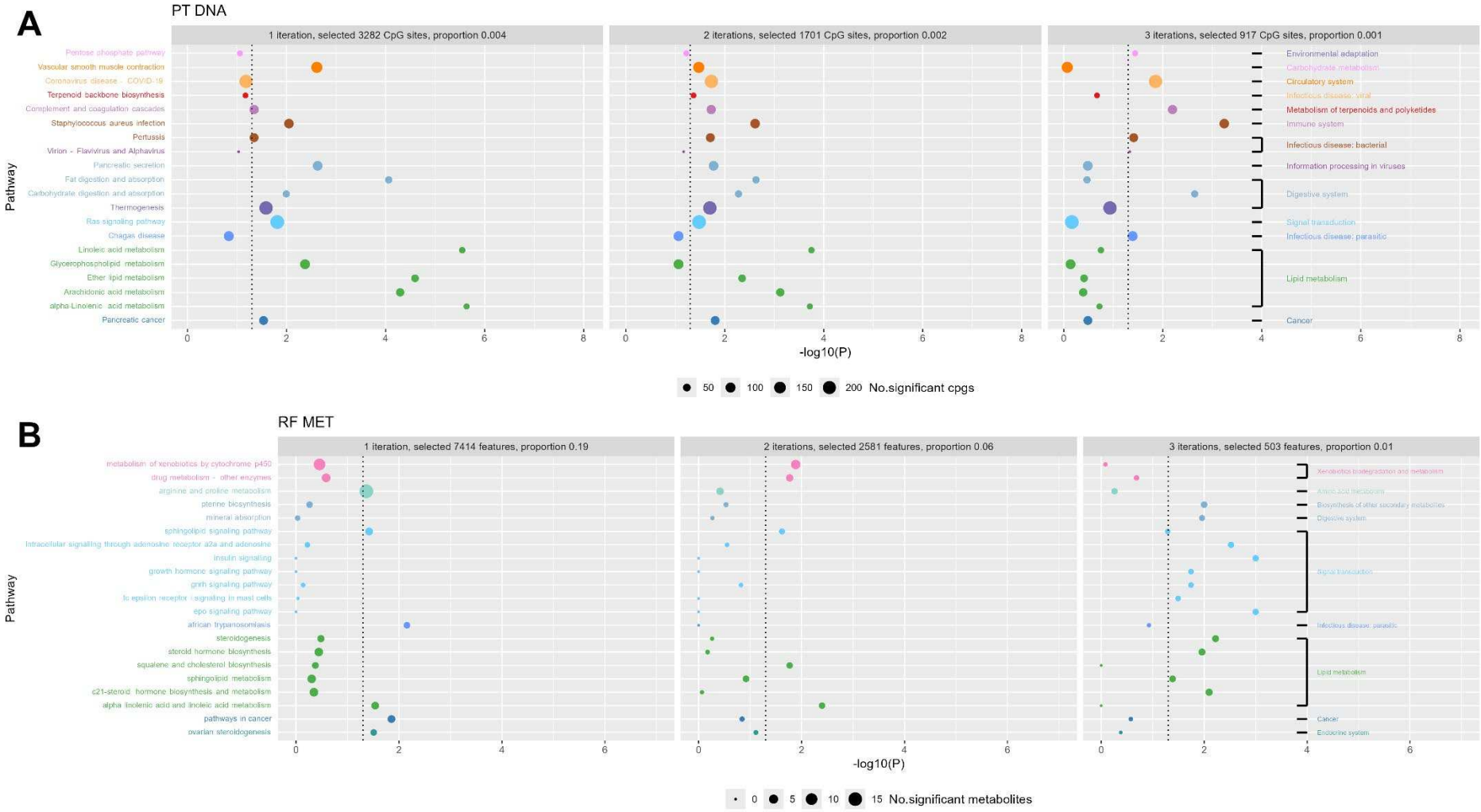



Figure S5. For the DNAm data set (A) with the PT method scatter plots of KEGG pathway enrichment analysis, where at least in (1), (2) or (3) the weighted CpG-sites were found. For the Metabolome data set (B) with the RF method scatter plots of KEGG pathway enrichment analysis, where at least in (1), (2) or (3) iterations the weighted features were found. Only pathways classes, which had at least one pathway

with significant p-values in more than one iteration. The number of significant CpG-Sites/ features in the pathway is indicated by the circle area, and the circle color represents the predefined pathway classes. On the x-axis the p-value (as -log10(p)) is represented with the significance level of 0.05 (dotted line). The proportion refers to the number of selected variables divided by the total number of variables (789,286 for DNAm; 35,348 for metabolomics). We display on the y-axis the different pathway terms enriched by KEGG database; the single pathway stands on the y-axis and the classes of them in the last plot.

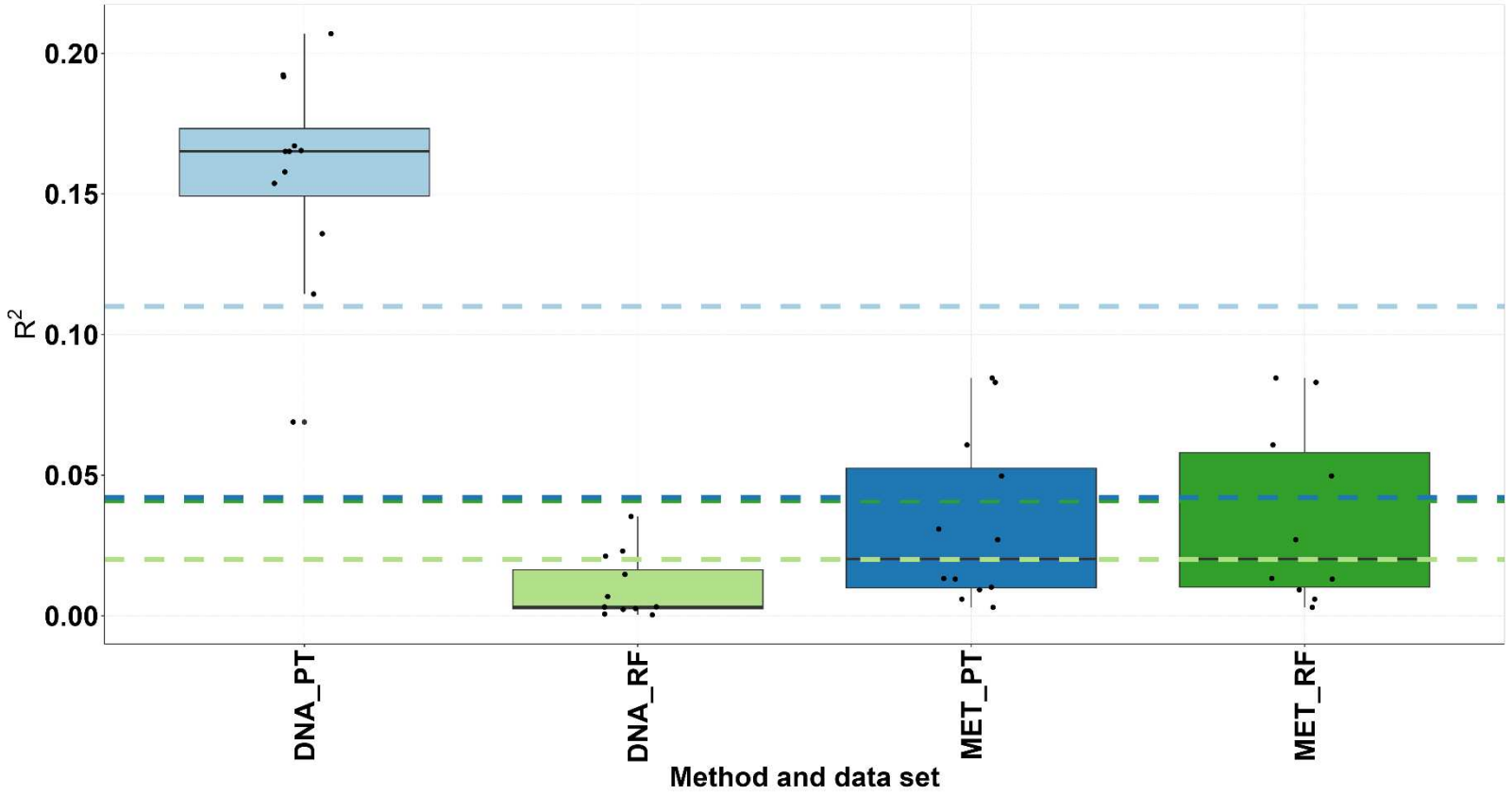


Figure S6. Comparison of the two best methods in the DNAm and metabolome datasets of the $R^2$ of different sample sizes. In the previous analysis of the single-omics PS, the maximum number of observations of the respective data sets were taken, i.e. DNAm (n=154) and metabolome (n=141) and $R^2$ determined. Now the joint observations were used as in the joint or multi-omics PS determination of n=138. We ran RF and PT as best methods for both data sets with 10 iterations. The methods and corresponding data sets are plotted on the x-axis and the $R^2$ on the y-axis. For comparison with the previous analyses, the medians were entered as dashed lines.

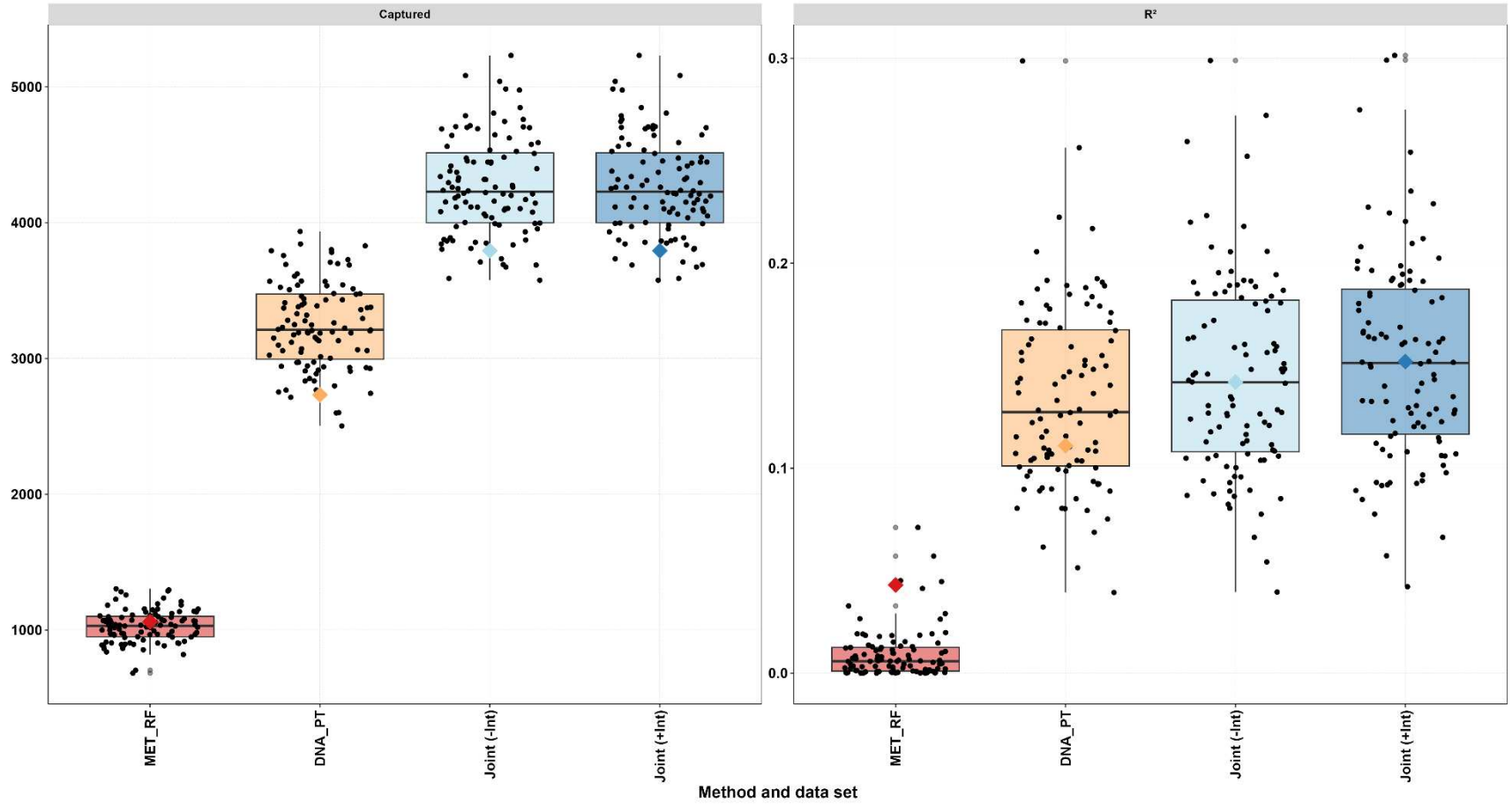


Figure S7. Comparison of the results of the PS for the best methods in the DNAm (PT) and metabolome (RF) datasets and the joint PS with and without the interaction term in 100 iterations. In the left figure we can see the numbers of captured features / CpG sites (y-axis) for the four different PS (x-axis) as a boxplot with 100 jittered points. In the right figure the $R^2$ for the four methods and datasets is showing for the 100 iterations. The colored star is the mean value of the 10 iterations from the main part of this work for both values (captured features/CpG sites and $R^2$).

## B) Additional tables

**Table ST1.** Comparison of $R^2$ values using Median (Q1: 1st, Q3: 3rd quartile and IQR: interquartile range) for the combination of both data sets (DNAm and metabolome data set) in the joint model with and without interaction term (equation (4)).

| | additive with interaction (N:10) | additive without interaction (N:10) |
|---|---|---|
| **DNA_BO vs. MET_BO** | | |
| Median (Q1, Q3) | 0.008 (0.004, 0.033) | 0.004 (0.001, 0.012) |
| IQR | 0.029 | 0.011 |
| **DNA_PT vs. MET_BO** | | |
| Median (Q1, Q3) | 0.017 (0.007, 0.024) | 0.006 (0.001, 0.021) |
| IQR | 0.017 | 0.019 |
| **DNA_PTEN vs. MET_BO** | | |
| Median (Q1, Q3) | 0.022 (0.015, 0.023) | 0.015 (0.014, 0.016) |
| IQR | 0.008 | 0.002 |
| **DNA_RF vs. MET_BO** | | |
| Median (Q1, Q3) | 0.020 (0.016, 0.028) | 0.011 (0.003, 0.017) |
| IQR | 0.012 | 0.015 |
| **DNA_WA vs. MET_BO** | | |
| Median (Q1, Q3) | 0.019 (0.008, 0.025) | 0.006 (0.002, 0.008) |
| IQR | 0.017 | 0.006 |
| **DNA_WAEN vs. MET_BO** | | |
| Median (Q1, Q3) | 0.009 (0.003, 0.018) | 0.003 (0.002, 0.004) |
| IQR | 0.014 | 0.003 |
| **DNA_BO vs. MET_EN** | | |
| Median (Q1, Q3) | 0.009 (0.004, 0.035) | 0.004 (0.001, 0.012) |
| IQR | 0.031 | 0.011 |
| **DNA_PT vs. MET_EN** | | |
| Median (Q1, Q3) | 0.016 (0.008, 0.025) | 0.007 (0.001, 0.018) |
| IQR | 0.017 | 0.016 |
| **DNA_PTEN vs. MET_EN** | | |
| Median (Q1, Q3) | 0.020 (0.014, 0.025) | 0.016 (0.011, 0.017) |
| IQR | 0.011 | 0.006 |
| **DNA_RF vs. MET_EN** | | |
| Median (Q1, Q3) | 0.021 (0.017, 0.028) | 0.011 (0.004, 0.016) |
| IQR | 0.011 | 0.012 |
| **DNA_WA vs. MET_EN** | | |
| Median (Q1, Q3) | 0.019 (0.008, 0.024) | 0.006 (0.002, 0.009) |
| IQR | 0.016 | 0.006 |
| **DNA_WAEN vs. MET_EN** | | |
| Median (Q1, Q3) | 0.008 (0.003, 0.016) | 0.003 (0.001, 0.004) |
| IQR | 0.013 | 0.003 |
| **DNA_BO vs. MET_PT** | | |
| Median (Q1, Q3) | 0.019 (0.006, 0.020) | 0.005 (0.001, 0.013) |

| | additive with interaction (N:10) | additive without interaction (N:10) |
|---|---|---|
| IQR | 0.014 | 0.012 |
| **DNA_PT vs. MET_PT** | | |
| Median (Q1, Q3) | 0.027 (0.005, 0.058) | 0.005 (0.001, 0.024) |
| IQR | 0.053 | 0.023 |
| **DNA_PTEN vs. MET_PT** | | |
| Median (Q1, Q3) | 0.043 (0.012, 0.055) | 0.032 (0.010, 0.043) |
| IQR | 0.044 | 0.033 |
| **DNA_RF vs. MET_PT** | | |
| Median (Q1, Q3) | 0.019 (0.015, 0.026) | 0.008 (0.006, 0.012) |
| IQR | 0.012 | 0.006 |
| **DNA_WA vs. MET_PT** | | |
| Median (Q1, Q3) | 0.014 (0.004, 0.026) | 0.003 (0.002, 0.004) |
| IQR | 0.023 | 0.002 |
| **DNA_WAEN vs. MET_PT** | | |
| Median (Q1, Q3) | 0.004 (0.003, 0.006) | 0.003 (0.002, 0.004) |
| IQR | 0.003 | 0.003 |
| **DNA_BO vs. MET_RF** | | |
| Median (Q1, Q3) | 0.017 (0.008, 0.027) | 0.006 (0.002, 0.013) |
| IQR | 0.020 | 0.010 |
| **DNA_PT vs. MET_RF** | | |
| Median (Q1, Q3) | 0.029 (0.024, 0.032) | 0.016 (0.009, 0.030) |
| IQR | 0.009 | 0.021 |
| **DNA_PTEN vs. MET_RF** | | |
| Median (Q1, Q3) | 0.025 (0.019, 0.030) | 0.019 (0.012, 0.025) |
| IQR | 0.011 | 0.013 |
| **DNA_RF vs. MET_RF** | | |
| Median (Q1, Q3) | 0.022 (0.015, 0.023) | 0.010 (0.006, 0.012) |
| IQR | 0.008 | 0.006 |
| **DNA_WA vs. MET_RF** | | |
| Median (Q1, Q3) | 0.013 (0.008, 0.015) | 0.005 (0.001, 0.010) |
| IQR | 0.008 | 0.010 |
| **DNA_WAEN vs. MET_RF** | | |
| Median (Q1, Q3) | 0.005 (0.002, 0.005) | 0.003 (0.001, 0.004) |
| IQR | 0.003 | 0.004 |
| **DNA_BO vs. MET_WA** | | |
| Median (Q1, Q3) | 0.008 (0.005, 0.014) | 0.004 (0.002, 0.011) |
| IQR | 0.009 | 0.010 |
| **DNA_PT vs. MET_WA** | | |
| Median (Q1, Q3) | 0.015 (0.009, 0.025) | 0.009 (0.007, 0.022) |
| IQR | 0.016 | 0.015 |
| **DNA_PTEN vs. MET_WA** | | |
| Median (Q1, Q3) | 0.010 (0.009, 0.015) | 0.006 (0.002, 0.010) |
| IQR | 0.006 | 0.007 |
| **DNA_RF vs. MET_WA** | | |

| | additive with interaction (N:10) | additive without interaction (N:10) |
|---|---|---|
| Median (Q1, Q3) | 0.008 (0.007, 0.021) | 0.007 (0.005, 0.016) |
| IQR | 0.014 | 0.011 |
| **DNA_WA vs. MET_WA** | | |
| Median (Q1, Q3) | 0.010 (0.007, 0.018) | 0.003 (0.001, 0.005) |
| IQR | 0.010 | 0.004 |
| **DNA_WAEN vs. MET_WA** | | |
| Median (Q1, Q3) | 0.015 (0.010, 0.020) | 0.003 (0.001, 0.005) |
| IQR | 0.010 | 0.004 |

**C) Further statistical description**

**i) Stage 4**

Stage 4a (Single-Omics PS): Here, each omics type (e.g., DNAm or metabolomic PS) is associated independently with the outcome variable. Models are adjusted for relevant covariates, such as sex, race, age at death, educational attainment, PMI and ADI, to accurately assess the association. For an ordinal outcome $Y$ with $L$ ordered categories ($Y \in \{1,2,\dots,L\}$), the cumulative probability is modeled as:

$$P(Y_i \leq k) = \Phi(\theta_k - \eta_i), k = 1, \dots, L-1,$$

where $P(Y_i \leq k)$ is the cumulative probability that $Y$ falls in category $k$ or below, $\Phi$ cumulative distribution function (e.g., logistic or normal), $\theta_k$ are thresholds separating the categories and $\eta_i$ are linear combination of predictors $\eta_i = \beta_0 + \beta_1 \cdot PS_i + \sum_{j=1}^{p} \gamma_j * X_{ij}$ [1]. The most widely used model for ordinal outcomes is the ordinal logistic regression model, which assumes proportional odds:

$$\log\left(\frac{P(y_i \leq k \mid X_i)}{P(y_i > k \mid X_i)}\right) = \theta_k - \eta_i, \qquad (1)$$

where $\frac{P(Yi>k)}{P(Yi\leq k)}$ are the cumulative odds of being in category $k$ or below, $\theta_k$ category-specific threshold $\eta_i = \beta_0 + \beta_1 \cdot PS_i + \sum_{j=1}^{p} \gamma_j * X_{ij}$ [2].

McFadden's $R^2$, also known as the pseudo-$R^2$, is a measure of goodness of fit for logistic regression models, including ordinal and multinomial logistic regression. For a given model, McFadden's $R^2$ is calculated as:

$$R^2 = 1 - \frac{ln\ L(full\ model)}{ln\ L(null\ model)},$$

where $ln\ L(full\ model)$ is the log-likelihood of the fitted model, including all predictors and $ln\ L(null\ model)$ is the log-likelihood of the null model, which includes only the intercept. McFadden's $R^2$ is typically lower than traditional $R^2$ because it is based on the improvement in log-likelihood rather than the explained variance, reflecting the model's predictive power for ordinal outcomes rather than continuous variance. A value between 0.2 and 0.4 indicates a meaningful improvement over the null model and is considered a good fit in this context. Unlike traditional $R^2$, which explains the proportion of variance, McFadden's $R^2$ should be interpreted as a relative measure of model fit rather than a direct percentage [3]. While the standard McFadden $R^2$ provides an overall measure of model fit, it does not indicate the unique contribution of specific predictors. This is where the partial McFadden $R^2$ becomes valuable.

The partial McFadden $R^2$ quantifies the contribution of a set of predictors by comparing two models: Restricted model that includes all covariates except for the predictors whose effect is being isolated and the full model that includes all covariates, including

the predictors of interest. By evaluating the improvement in log-likelihood when the predictors of interest are added, the partial McFadden $R^2$ provides insight into their explanatory power. The formula for the partial McFadden $R^2$ is given by: $R^2_{partial} = \frac{lnL_{restricted} - lnL_{full}}{lnL_{restricted}} = 1 - \frac{\ln L_{full}}{\ln L_{restricted}}$, where $\ln L_{restricted}$ is the log-likelihood of the model that excludes the predictors of interest and $\ln L_{full}$ is the log-likelihood of the model that includes the predictors of interest.

The partial McFadden $R^2$ can be viewed as analogous to a partial $R^2$ in linear regression, quantifying the incremental improvement provided by the additional predictors. Since this value is derived from a ratio of log-likelihoods, it reflects the relative explanatory power of the added predictors compared to the restricted model. In practice, researchers and practitioners often use this measure to decide whether the inclusion of extra predictors justifies the added model complexity.

Stage 4b (Joint-Omics PS): This step involves integrating multiple omics Risk Scores into a joint model to test if they collectively improve predictive power Assuming that we have two PS from different omics data the regression equation (see (1)) is extended to $\eta_i = \beta_0 + \beta_1 \cdot PS_{1,i} + \beta_2 \cdot PS_{2,i} + \sum_{j=1}^{p} \gamma_j * X_{ij}$. This is the combined integrated model for both omics data sets. In ordinal outcome models, interaction terms are used to assess whether the effect of one predictor depends on the level of another predictor, resulting in the following equation: $\eta_i = \beta_0 + \beta_1 \cdot PS_{1,i} + \beta_2 \cdot PS_{2,i} + \beta_3 \cdot (PS_{1,i} * PS_{2,i}) + \sum_{j=1}^{p} \gamma_j * X_{ij}$.

Incorporating an interaction term between the two single-omics PS allows for a more nuanced understanding of relationships between predictors and ordinal outcomes. For the goodness of fit for logistic regression models we use like in the single-omics PS the partial McFadden's $R^2$ [3].

Stage 4c (Multi-Omics PS):

As mentioned above here we combine both data sets (DNAm and metabolome) with some preprocessing steps like z-score transform. With this combined new data set the calculation stages are the same as in Stage 4a for the single-omics PS was described.

### ii) Methods

## Statistical and machine learning methods for variable selection & estimation of weights

### 1. Pruning & Thresholding (PT)

Pruning involves systematically reducing the number of features, or CpG-sites in a model by removing those that are highly correlated or redundant. By eliminating these

correlated markers, pruning simplifies the model, reduces multicollinearity, and decreases computational demands without significantly compromising prediction accuracy [4]. Pruning generally involves iteratively removing variants/variables based on their correlations without relying on association significance, focusing primarily on reducing multicollinearity across the entire set of variants. In the present analysis, pruning was integrated into hierarchical clustering. Hierarchical clustering with pruning is a technique that applies pruning principles to hierarchical clustering to reduce redundancy in high-dimensional data. By combining clustering with pruning, it is possible to create a streamlined clustering result that retains the most informative clusters and minimizes noise or redundancies, particularly useful in DNAm, metabolomic, and other "omics" datasets. Hierarchical clustering organizes data points into a tree-like structure (dendrogram) based on their pairwise distances. It begins with each data point as its own cluster and iteratively merges the two closest clusters until a single cluster encompassing all data points is formed. We use one of two main types of hierarchical clustering: the Agglomerative (bottom-up) clustering: Start with each data point as a separate cluster and merge them iteratively. Complete-linkage clustering is a corresponding method for agglomerative clustering where the distance between two clusters is defined as the maximum distance between any pair of points in the two clusters. For the distance metric in hierarchical clustering, we use the Pearson correlation to calculate the distance between the clusters. When applying pruning to hierarchical clustering, the goal is to simplify the dendrogram by removing clusters or data points that add little new information, thus focusing on clusters with distinct, meaningful information. In high-dimensional data, each data point (e.g., CpG sites, metabolites features) has an associated importance score (e.g., association strength, p-value). During hierarchical clustering, prune variables within each cluster that fall below a set importance threshold, reducing redundancy by keeping only the top variables within each cluster.

Thresholding is a technique commonly used in high-dimensional data analysis, particularly in the context of genetic, epigenetic, and metabolomic studies, to filter out weak associations and focus on the most relevant variables. By applying a statistical or p-value-based cutoff, thresholding helps to reduce noise, improve interpretability, and enhance the predictive power of models by including only those variables that meet a predefined level of importance or association [4].

Thresholding can be formally described as follows:

1. Define a Threshold $t$: The first step is to define a threshold $t$, which could be a p-value, a correlation coefficient, or any other measure of association or importance.

2. Apply the Threshold: For each feature $x_i$ (e.g., CpG site, metabolite) with an associated score $s(x_i)$, retain $x_i$ if and only if $s(x_i) \geq t$. This score $s(x_i)$ could represent statistical significance (e.g., p-value), effect size, correlation

coefficient, or another measure relevant to the context. Mathematically: $x_i$ is retained if $s(x_i) \geq t$.

3. Binary Masking: Thresholding can be represented by a binary indicator function, where each feature is assigned a value of 1 (included) or 0 (excluded) based on whether it meets the threshold:

4. $I(x_i) = \begin{cases} 1 & if\ s(x_i) \geq t \\ 0 & if\ s(x_i) < t \end{cases}$

This binary masking function $I(x_i)$ can then be used to filter variables in downstream analyses, retaining only those that meet the threshold.

Together, pruning and thresholding provide complementary techniques to manage the vast number of variables, leading to more streamlined and interpretable prediction models [5].

**2. Elastic Net (EN)**

Elastic Net (regression) is a regularization technique that linearly combines the penalties of Lasso (L1) and Ridge (L2) regression to address some of the limitations of each approach individually. It is particularly useful when dealing with datasets with many predictors or when predictors are highly correlated. Elastic Net aims to retain the feature selection ability of Lasso and the stability provided by Ridge in a single model [6].

The Elastic Net objective function is defined as:

$$\min_{\beta} \left( \frac{1}{2N} \sum_{i=1}^{N} (y_i - X_i\beta)^2 + \lambda \left( \alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right) \right),$$

where $N$ is the number of observations, $p$ is the number of variables, $y_i$ is the target outcome, $X_i$ is the vector of variables for the $i$-th observation, $\beta$ is the vector of coefficients to be estimated, $\lambda$ is the regularization parameter controlling the overall strength of the penalty and $\alpha$ is the mixing parameter that balances the L1 (Lasso) and L2 (Ridge) penalties.

The Elastic Net penalty term consists of two components: 1. L1 Penalty (Lasso): $\alpha \sum_{j=1}^{p} |\beta_j|$, which encourages sparsity by driving some coefficients to zero. 2. L2 Penalty (Ridge): $\frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2$, which helps handle multicollinearity by shrinking coefficients without necessarily setting them to zero. Both key parameters can be explained in the following manner: The regularization parameter $\lambda$ controls the extent of regularization. Higher values of $\lambda$ increase the penalty on the coefficients, reducing model complexity. Mixing Parameter $\alpha$ determines the balance between Lasso and Ridge penalties: For $\alpha = 1$ the model behaves like Lasso, focusing on feature selection. For $\alpha = 0$ the model

behaves like Ridge, focusing on coefficient shrinkage without setting them to zero. And between $0 < \alpha < 1$ the model applies a combination of both L1 and L2 penalties, achieving a trade-off between sparsity and stability.

Elastic Net is advantageous for datasets where predictors are highly correlated because it can select groups of correlated variables together [6]. In contrast, Lasso tends to select only one feature from a group of correlated variables, while Elastic Net encourages joint selection.

When applying Elastic Net regression to an ordinal outcome (i.e., a dependent variable with ordered categories, such as rating scales from "low" to "high"), additional considerations are necessary to account for the ordered nature of the outcome. Standard Elastic Net regression is designed for continuous outcomes, but with ordinal outcomes, ordinal regression models (e.g., ordinal logistic regression) can be adapted with Elastic Net regularization. This approach ensures that the model respects the ordinal structure while applying regularization to manage multicollinearity and improve prediction accuracy.

The Elastic Net regularized ordinal regression objective function becomes:

$$\min_{\beta}\left(-\sum_{i=1}^{N}\log P(y_i\,|X_i\beta) + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2\right)\right),$$

where $P(y_i\,|X_i\beta)$ is the probability of observing the ordinal outcome $y_i$ for the $i$ -th observation, given the predictors $X_i$ and coefficients $\beta$. The probability $P(y_i\,|X_i\beta)$ is often modeled using an ordinal logistic (proportional odds) or probit function, respecting the ordinal structure by assuming that outcomes fall within specific ranges of a latent continuous variable.

Example: Proportional Odds Model with Elastic Net Regularization

For ordinal logistic regression, the proportional odds assumption assumes a common slope for each category of the ordinal outcome but different intercepts, defined as:

$$\log\left(\frac{P(y_i \le k\,|X_i)}{P(y_i > k\,|X_i)}\right) = \theta_k - \eta_i,$$

where $k$ indexes the thresholds between ordinal categories, $\theta_k$ are the threshold parameters that separate the ordinal categories and $X_i\beta$ is the linear predictor with the Elastic Net penalty applied to $\beta$ as in the Elastic Net objective function.

This formulation, incorporating Elastic Net regularization, encourages a sparse or grouped solution depending on the values of α and $\lambda$ and is advantageous when predictors are numerous or correlated [2].

### 3. Boosting (BO)

Gradient boosting [7]is a state-of-the-art ensemble method for constructing predictive models in an iterative fashion. The idea behind gradient boosting is performing functional gradient descent in the linear span of simple models (such as simple linear regression models or shallow decision trees). More precisely, in each iteration,

1. the gradient of the current loss (between the outcome and the current model predictions) is computed,
2. a simple model is fit to the gradient, and
3. the ensemble model is updated by adding the new model.

Thus, the resulting model $f(\boldsymbol{X}) = \rho_0 + \sum_{b=1}^{B} f_b \rho_b(\boldsymbol{X})$ is a sum of *B* simple models $f_b$ weighted by boosting coefficients $\rho_b$.

In high-dimensional applications in which only a fraction of predictors influence the outcome, gradient boosting can be particularly efficient for constructing multiple linear regression models, as the computational complexity is given by $O(Bpn)$ (see, e.g., [8]), where *B* is the (pre-specified) number of boosting iterations/maximum number of terms, compared to ordinary linear regression or regularized procedures such as elastic net that exhibit a complexity of $O(p^2 n)$ [9].

Gradient boosting can be also employed for other outcome types that belong to the exponential family [10]. In this paper, an ordinal outcome $Y \epsilon \{1, \dots, K\}$ is studied. Hence, an ordinal boosting algorithm is used that models the outcome with the commonly used proportional odds assumption

$$P(Y \leq k) = \frac{1}{1 + \exp(f(X) - \theta_k)}, \qquad k = 1, \dots, K,$$

for category thresholds $\theta_1, \dots, \theta_K$ . Schmid et al. [11] derived the necessary negative log-likelihood loss function with corresponding gradient for applying gradient boosting to ordinal outcomes and proposed optimizing the thresholds in every boosting iteration with respect to the current model. More details can be found in [11]. We implemented the algorithm in R and did not use standard boosting libraries, as their application crashed when attempting to fit models to our high-dimensional data set (*p* = 789*,286*).

**4. Random Forest (RF)**

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and robustness, making it effective for both classification and regression tasks [12]. The model builds a "forest" of trees, where each tree is grown on a random subset of the data, helping to reduce overfitting and improve generalizability.

The Random Forest algorithm follows these main steps:

1. Bootstrap Sampling: From the original dataset of size $N$, multiple samples are drawn with replacement to create bootstrap samples. Each sample will serve as the training set for a single decision tree.
2. Feature Subset Selection: At each node within each tree, a random subset of variables (of size $m$, where $m < p$, with $p$ as the total number of variables) is selected. The best split among these variables is chosen, a process that helps reduce the correlation between trees and thus improves ensemble diversity.
3. Tree Growth: Each decision tree is grown without pruning, allowing it to reach its maximum depth until a stopping criterion (e.g., minimum node size) is met. This ensures high variance among individual trees.
4. Aggregation (Voting or Averaging): For a classification problem, the Random Forest prediction is given by: $\hat{y} = mode(\{T_b(x)\}_{b=1}^{B})$, where: $\hat{y}$ is the predicted class label, $T_b(x)$ is the prediction from the $b$ -th tree in the ensemble, given input $x$ and $B$ is the total number of trees.

For regression, the Random Forest prediction is the mean of the predictions from each tree:$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$,where: $\hat{y}$ is the predicted class label, $T_b(x)$ is the prediction from the $b$ -th tree in the ensemble, given input $x$ and $B$ is the total number of trees.

Since each tree in RF is trained on a bootstrap sample, roughly one-third of the original samples are left out (not included in the training sample). These are known as Out-of-Bag (OOB) samples and can be used to estimate the model's error without requiring a separate test set [12]. Random Forest provides a measure of feature importance based on how much each feature improves the split criterion (e.g., Gini impurity or mean squared error) when used in nodes across all trees. By combining multiple trees and using random feature selection at each split, Random Forest reduces the risk of overfitting compared to individual decision trees [13].

To identify statistically significant variables in a Random Forest model, a variable importance test was developed. The work from Janitza et al. [14],introduces a computationally efficient approach designed for high-dimensional data, where traditional permutation-based tests are often too resource-intensive. The proposed method leverages a modified version of the permutation Variable Importance Measures (VIMs), inspired by cross-validation, to estimate an empirical null distribution based on non-positive importance scores. Compared to standard permutation-based tests, it

maintains Type I error control while achieving similar or even higher power at a significantly lower computational cost. The function var.sel.vita() implements the Vita approach for variable selection, ensuring that feature importance is statistically validated rather than merely relying on raw importance scores. By constructing an empirical null distribution from non-positive VIMs, as described by Janitza et al. [14], and utilizing the importance_pvalues function from the ranger package [15], it assigns p-values to each variable. This process enables the identification of statistically significant predictors, retaining only truly relevant features and enhancing model interpretability while mitigating overfitting risks.

**5. (Sliding) Windows approach (WA)**

Another method, particularly suitable for large amounts of data, is the so-called cross leverage score (CLS) [16]. These were originally developed for variable selection in genome-wide studies but can also be applied to other types of data, such as omics data. We will use these scores for a pre-selection of variables with which we will later calculate the risk scores. The CLS indicate for each variable its leverage on the outcome variable. Theoretically, each CLS is equal to its corresponding parameter in a least squares solution up to a small bounded additive error. At the same time, the score also contains information on whether the corresponding variable is part of a significant interaction effect, even if there is no significant main effect of the variable. The great advantage of the CLS is that these scores can also be calculated streamwise. The data stream algorithm allows data to be read in sequentially, resulting in a very fast and effective calculation that uses only a fraction of the computer's memory.

The approach is motivated by dimension reduction methods, whereby a $n \ll p$ problem has to be addressed here. Therefore, we consider the transposed matrix

$$\tilde{X} = [X, y]^T \in \mathbb{R}^{\tilde{p}\, x\, n}$$

with $\tilde{p} = p + 1$ and $p$ the number of variables and n the number of observations. $X \in \mathbb{R}^{n\, x\, p}$ is the data matrix and $y \in \mathbb{R}^n$ denotes the response. The CLS are given by the off-diagonal entries of the hat matrix $H = QQ^T$ of $\tilde{X}$. So, we need to calculate the QR-decomposition $\tilde{X} = QR$. Since we are only interested in the CLS between the variables $j \in \{1, \dots, p\}$ and the response $y$, we have to calculate the dot products of rows $Q_{j.}$ with row $Q_{\tilde{p}.}$ :

$$c_{j\tilde{p}} = \langle Q_{j.}, Q_{\tilde{p}.} \rangle, \qquad j \in \{1, \dots, p\}\,.$$

This score provides information on the mutual influence of $x_j$ and $y$ [16]. To avoid computing the QR decomposition with running time $O(n^2 p)$ [17], which is prohibitively slow for p very large (e.g. millions of omics variables), we consider a streamwise computation, here the sliding window approach [16]. We calculate the QR-decomposition for many small matrices instead of one QR-decomposition for a very

large matrix and merge the results a suitable way. We the select the $\mathrm{q} = \lceil n \log n \rceil$ variables that have the most extreme CLS for subsequent analyses. Theoretically, this is motivated by the coupon collector's problem [18] , which requires oversampling by a $\log n$ factor so that the selection contains at least as many variables to ensure that the submatrix preserves the full rank n of the original matrix [19].

**References (Supplement only)**


[1] McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*.

[2] Simon, N. *et al.* (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*.

[3] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior.

[4] Wu, J. *et al.* (2013). Strategies for Developing Prediction Models From Genome-Wide Association Studies. *Genetic Epidemiology*. https://doi.org/10.1002/gepi.21762.

[5] Privé, F. *et al.* (2019). Making the Most of Clumping and Thresholding for Polygenic Scores. *The American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2019.11.001.

[6] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[7] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. https://doi.org/10.1214/aos/1013203451.

[8] Lau, M. *et al.* (2025). Boosting interaction tree stumps for modeling statistical interactions. .

[9] Efron, B. *et al.* (2004). Least angle regression. *The Annals of Statistics*. https://doi.org/10.1214/009053604000000067.

[10] Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*. https://doi.org/10.1214/07-STS242.

[11] Schmid, M. *et al.* (2011). Geoadditive regression modeling of stream biological condition. *Environmental and Ecological Statistics*. https://doi.org/10.1007/s10651-010-0158-4.

[12] Breiman, L. (2001). Random Forests. *Machine Learning*. https://doi.org/10.1023/A:1010933404324.

[13] Tin Kam HoRandom decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*. https://doi.org/10.1109/ICDAR.1995.598994.

[14] Janitza, S. *et al.* (2015). A computationally fast variable importance test for random forests for high-dimensional data.

[15] Wright, M.N. and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v077.i01.

[16] Teschke, S. *et al.* (2024). Detecting Interactions in High-Dimensional Data Using Cross Leverage Scores. *Biometrical Journal*. https://doi.org/10.1002/bimj.70014.

[17] Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computations* , Johns Hopkins University Press.

[18] TROPP, J.A. (2011). IMPROVED ANALYSIS OF THE SUBSAMPLED RANDOMIZED HADAMARD TRANSFORM. *Advances in Adaptive Data Analysis*. https://doi.org/10.1142/S1793536911000787.

[19] Erdős, P. and Rényi, A.. (1961). *On a Classical Problem of Probability Theory.A Magyar Tudományos Akadémia Matematikai. Kutató Intézetének Közleményei 6*.

# *Article 4*

# Hierarchical modelling of risk factors with and without prior information – the process of regression model evaluation for an example of respiratory diseases in piglet production from daily practice data

Timur Tug[1*], Fiona Mers[1], Franziska Schäkel[2], Doris Höltig[3], Lothar Kreienbrock[2†], Katja Ickstadt[1†]

[1] Department of Statistics, TU Dortmund university, Dortmund, Germany

[2] Institute of Biometry, Epidemiology and Information Processing, WHO Collaborating Centre for Research and Training for Health at the Human-Animal-Environment Interface, University of Veterinary Medicine, Hannover, Germany

[3] Clinical Centre for Farm Animals, University of Veterinary Medicine, Hannover, Germany

*** Correspondence:**

Corresponding Authors

timur.tug@tu-dortmund.de; lothar.kreienbrock@tiho-hannover.de

†: These authors contributed equally to this work and share last authorship.

Keywords: pork production, biosecurity, herd health management, respiratory health, hierarchical regression, frequentistic modelling, Bayesian modelling, model evaluation.

## Abstract

In veterinary epidemiology, regression models are commonly used to describe animal health and related risk factors. However, model selection and evaluation present ongoing challenges - especially when many potential predictors, complex interactions, and limited sample sizes are involved. The VASIB project serves as a representative example, focusing on piglet-producing farms with persistent respiratory disease problems. Across 30 farms, a wide array of variables was collected at the farm, barn, compartment, pen, and individual animal levels, aiming to support optimized treatment and management strategies to improve respiratory health.

This study investigates the occurrence of coughing in pigs using various epidemiological models, including hierarchical frequentist logistic regression, non-hierarchical Bayesian logistic regression (with full and partial pooling), and hierarchical Bayesian models with informative and non-informative priors. These approaches are evaluated and compared using statistical measures such as the corrected Akaike Information Criterion ($AIC_c$), marginal and conditional $R^2$, and intra-class correlation coefficients ($ICC_c$ and $ICC_{adj}$).

In the frequentist models, convergence issues arose due to limited observations within clusters, which did not occur in the Bayesian framework. While the choice of priors had limited influence on Bayesian model results, differences between models suggest that prior specification can still be relevant. Thus, it is important to assess and compare various model structures—including both hierarchical and non-hierarchical, and Bayesian versus frequentist approaches - to capture the data's complexity and ensure robust inference.

Across all models, floor condition emerged as a consistently significant factor influencing the likelihood of coughing. Other variables - such as water flow rate, stocking density, $CO_2$, $H_2S$, and temperature - varied in their significance depending on the model type. This highlights the importance of tailoring the model structure to the specific epidemiological question and data context.

Overall, this work emphasizes that there is no universal rule for model selection in veterinary data analysis. Instead, a balanced, context-sensitive modeling strategy that considers both statistical and epidemiological perspectives is essential to derive meaningful and actionable conclusions for improving animal health.

## 1 Introduction

The hygiene status of a farm is a major factor influencing animal health and is therefore a fundamental part of veterinary advice (Robertson 2020). However, the subjective rating of the veterinarian is prone to a lot of internal and external factors and may differ from visit to visit and can be difficult to understand. For this reason, careful evaluation is needed, which factors measure the health outcome on farms in a harmonized form and highlight critical points in an understandable way. In general, from the epidemiological point of view, this evaluation is in line with a careful development of regression models, connecting the health outcome with a more or less complex set of interacting factors under study, which describe the several biosecurity measures and other influencing factors (Wilson and Lorenz 2015; Dohoo, Martin, and Stryhn 2003; Amenu et al. 2023; Gelman and Hill 2007).

In this work, the development of regression models for data from daily veterinary practice is described. Within the research project VASIB, in selected farms with sustainable respiratory disease problems, the aim is to examine whether targeted diagnostic measures, optimization of the treatment strategy and comprehensive, intensive management advice can minimize respiratory symptoms and with this the use of antibiotics, and thus make an active contribution to reducing the general development of resistance in livestock farming. To this end, the project is working on the development and validation of a model that can be used with onsite-farming data from veterinary practices with the aim of synergizing epidemiological data from veterinary preventive medicine and farm data (Müller 2020).

However, to describe respiratory health in piglet production different herd measures, which are based on direct veterinary inspection and on information from the farmers may be used. Overall, this may be interpreted as a multivariate health outcome or the need of selection a representative surrogate to describe respiratory health. Within this investigation we choose as a surrogate "coughing in piglets", which is generally used in practice (Pessoa et al. 2022).

Nevertheless, many other factors must be considered, if coughing has to be described during the veterinary inspection visit. First, animal health data on a farm appears at farm, barn, compartment, pen and individual animal level. These

hierarchical structures must be considered especially if factors respond in different ways, like the air- or feedborne transmission (Maes et al. 2018). Second, manyfold direct (causal) and indirect factors effect animal health, which are more or less associated within an interacting and partial correlated structure (Grosse-Kleimann et al. 2021; Sanchez-Vazquez et al. 2011). And, if these multiple factors lead to a large number of different classes, they break down into multiple substructures, which usually do not contain a sufficient number of animals for a powerful epidemiological analysis. This at the end, causes missing data, which finally restricts the prognostic value of an epidemiological model (Dohoo, Martin, and Stryhn 2003).

Against this background the development of an epidemiological statistical model to be used for prevention in livestock farming is not a matter of highest quality and precision only, but on a sophisticated model building process, which takes into account the needs of daily work data.

Classical statistical models often struggle to account for complex data structures, leading to potential bias when questionable or poorly defined covariates are included (Dohoo, Martin, and Stryhn 2003). To address this, advanced statistical methodologies, such as hierarchical or generalized models, are employed to better capture variability and dependencies in the data (Gelman and Hill 2007). Bayesian methods are particularly valuable in cases where prior knowledge exists, enabling the incorporation of expert insights into the modelling process for more robust and informed inference (McElreath 2020).

This paper shows a process of regression model evaluation using the example of respiratory diseases in piglet production from daily practice data used from the VASIB-project.

## 2 Material & Methods

### *2.1 Study Design and Data Acquisition*

The data used for this investigation was enrolled during the VASIB-project on farms with sustainable respiratory health problems. For this, 30 piglet-producing farms with recurring respiratory tract disease problems in weaners were selected for an in-deep investigation. Overall, data was enrolled within 72 pigsties, 130 compartments, and 300 pens respectively, and finally from 450 single animals. All farms were connected to one veterinary practice network that is located in the Federal States of North Rhine Westphalia and Schleswig Holstein, Germany.

In preparation for the veterinary visit, a questionnaire was developed and evaluated beforehand. The questionnaire included management, biosecurity, feeding, medication and medical history aspects (the original German version is on view at XXXX, @Editor this link will be active only, if the paper is accepted for publication).

The questionnaire was sent to the participating farmers to obtain general information about the farm before the veterinarian's inspection. At the beginning of the veterinarian's inspection visit, the questionnaire was reviewed in a face-to-face interview, and missing values or implausible answers were clarified.

The information of the questionnaires, the checklist and the results of the clinical examination were entered in a SQL database developed for the study. All datasets were checked for plausibility and completeness.

For demonstration of the regression modelling process, the variables addressed in Table 1 were selected for this investigation.

**Table 1: Variables and Descriptions in the "Initial" Dataset**

| Variable | Description (Categories) | Level |
|---|---|---|
| Cough | Do the pigs cough? (0 = no; 1 = yes) | Animal |

**Table 1: Variables and Descriptions in the "Initial" Dataset**

| Variable | Description (Categories) | Level |
|---|---|---|
| Clinical variables | Sum of all clinical variables (0 = no symptoms; 1 = mild symptoms) | Animal |
| Laboratory variables | Sum of all laboratory variables, including blood | Animal |
| Blood lab variables | Sum of blood lab variables | Animal |
| Pen ID | ID variable for the pens | Pen |
| Age | Age of animals (in days) | Pen |
| Animal pollution | Degree of dirtiness of the animals (0 = no findings; 1 = slightly dirty; 2 = moderately dirty; 3 = heavily dirty) | Pen |
| Skin injuries | Animals with skin injuries (0 = none; 1 = few (up to 10%); 2 = some (up to 50%); 3 = many (over 50%)) | Pen |
| Pen size | Size of the pen (in $m^2$) | Pen |
| Stocking density | Stocking density in the pens | Pen |
| Compartment ID | ID variable for the compartments | Compartment |
| Water flow rate | Water flow rate (in ml/min) | Compartment |
| Temperature | Recorded temperature (in °C) | Compartment |
| Air pressure | Air pressure (in Pa) | Compartment |
| $CO_2$ level | $CO_2$ level (in ppm) | Compartment |
| Relative humidity | Relative humidity (in %) | Compartment |
| $NH_3$ level | $NH_3$-adjusted value (in ppm) | Compartment |
| $H_2S$ level | $H_2S$ level (in ppm) | Compartment |

**Table 1: Variables and Descriptions in the "Initial" Dataset**

| Variable | Description (Categories) | Level |
|---|---|---|
| Floor condition | Condition of the floor (1 = new; 2 = moderately worn; 3 = heavily worn; 4 = damaged) | Compartment |
| Farm ID | ID variable for the farms | Farm |
| Disinfectant | Frequency of disinfectant replacement in disinfection baths (1 = daily; 2 = weekly; 3 = when dirty; 4 = after emptying; 5 = irregularly) | Farm |
| Proximity to next farm | Proximity to the nearest pig farm (1 = <0.5 km; 2 = 0.5 km - 10 km; 3 = >10 km) | Farm |
| Target temperature - In | Average target temperature when animals are housed (in °C) | Farm |
| Target temperature - Out | Average target temperature when animals are removed (in °C) | Farm |
| Respiratory diseases | Batches affected by respiratory diseases last year (1 = few (up to 10%); 2 = some (up to 50%); 3 = many (over 50%)) | Farm |
| Protective clothing | Use of protective clothing outside barns (0 = no; 1 = yes, to cross the yard; 2 = yes, for other tasks) | Farm |
| Minimum quarantine | Minimum quarantine duration (in days) | Farm |
| Maximum quarantine | Maximum quarantine duration (in days) | Farm |
| Winter | Was the farm visited in winter? (0 = no; 1 = yes) | Farm |

In order to investigate the influence on cough in the study population hierarchical logistic (frequentist and Bayesian) regression models were used. Here, we include all levels (farms → pigsties → compartments → pens) as mentioned above. The pen sizes vary between 12 and 85 animals, from which overall 450 animals were included into individual veterinary inspection (see Figure 1).

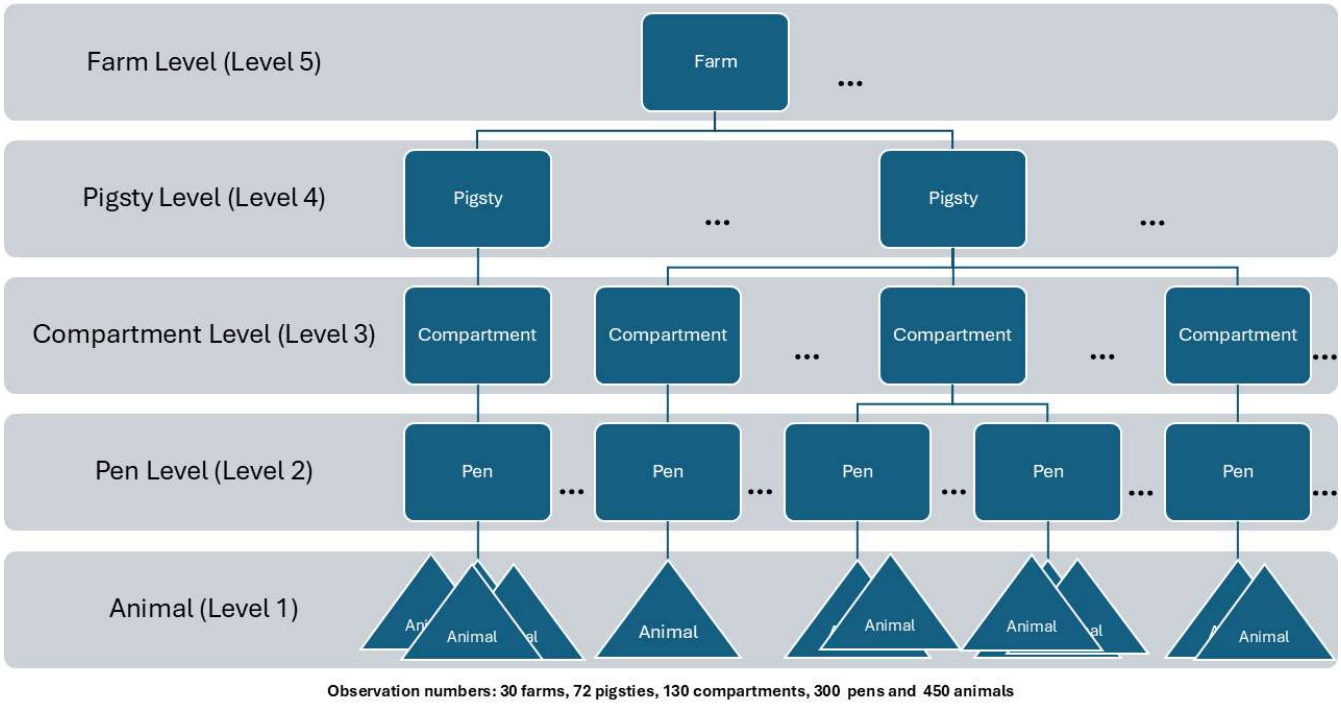



**Figure 1: Hierarchical Levels within the Sample Population**

In this work, multiple imputation was used to handle missing values in order to ensure data integrity and avoid bias in the results. For quantitative variables, the predictive mean matching (PMM) method was used, in which real values are drawn from the available data to generate plausible replacement values for missing entries (Rubin 1986). Categorical variables were imputed using a proportional odds logistic regression (polr) model to adequately account for ordinal relationships between categories (Grund, Lüdtke, and Robitzsch 2023; White, Daniel, and Royston 2010).

These methods made it possible to minimize the impact of missing data and perform a more robust analysis of factors influencing cough in pigs.The R software packages "mice" and "miceadds" are used for both imputation methods (Robitzsch and Grund 2025; Buuren and Groothuis-Oudshoorn 2011). The hierarchical structure of the data (animal - pen - compartment - farm) was also considered to ensure that the imputation procedure met the specific requirements of the dataset (Grund, Lüdtke, and Robitzsch 2023; White, Daniel, and Royston 2010).

### *2.2 Classical analysis of hierarchical model*

Hierarchical regression extends classical regression by handling clustered data structured across multiple levels (see Figure 1). Here, the goal is to account for the variability at each level to analyze cluster effects(Wilson and Lorenz 2015). The key elements are varying coefficients and a model for these coefficients, possibly incorporating cluster-level predictors - features that distinguish hierarchical models from classical ones. Here, we examine frequentist hierarchical models and compare their applicability in the context of cough prevention in pig production.

For multilevel data, each level contributes a variance component that measures intraclass-correlation. As an example here, consider a three-level model for the cough response, $y_{ijk} \sim Ber(\pi_{ijk})$ , with cough probability $\pi_{ijk}$ for the $k$-th pig, located in the $j$-th pen in the $i$-th compartment. Pigs (level 1) are nested in pens (level 2), which are nested in compartments (level 3); compartments are the primary units, pens the secondary units, and pigs the units of observation. These clusters are treated as random effects with an average effect of zero, and the analysis is performed using logistic regression. Then, these results are modelled by a logistic regression model (1)

$$\boldsymbol{log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \beta_{0ij} + \beta_{1ij} pig^{1}_{ijk}} \quad (1)$$

where $\beta_{0ij}$ is the intercept, $\beta_{1ij}$ the coefficient associated with the predictor $pig^1_{ijk}$for $k = 1,\ldots,n_{ij}$ pigs, $j = 1,\ldots,n_i$ pens and $i = 1,\ldots,n, n\ (n \epsilon \mathbb{N})$ compartments.

Each pen is modelled with its own logistic regression. While the pig variable $pig^1$ is a fixed effect with the same impact across pens, the pen-specific intercept $\beta_{0ij}$ captures differences between pens in model (2) as

$$\boldsymbol{\beta_{0ij} = \beta_{0i} + \beta_{2i}\, pen^2_{ij} + u_{0ij}} \quad (2)$$

where $pen^2_{ij}$ is a pen-level covariate and $u_{0i}$ is a random effect. At the compartment level, the intercept is

$$\boldsymbol{\beta_{0i} = \beta_{0i} + u_{0i}}. \quad (3)$$

Substituting (3) into the (1) gives

$$\boldsymbol{log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \beta_0 + \beta_{1ij} pig^1_{ijk} + \beta_{2i} pen^2_{ij} + \beta_3 compartment^3_i + u_{0i} + u_{0ij}}. \quad (4)$$

This generalized linear mixed model includes two random effects - $u_{0i}$ for compartments ($u_{0i} \sim N(0, \sigma^2_{u_i})$ ) and $u_{0ij}$ for pens ($u_{0ij} \sim N(0, \sigma^2_{u_{0ij}})$) - which account for cluster-specific variability (Wilson and Lorenz 2015). Note that with only five individuals per cluster, variance estimates may be unreliable, and high variability in random effects can bias fixed effect estimates. Hierarchical models are thus better able to handle dependencies in nested data than non-hierarchical approaches (Austin 2010). Estimation of variance components is unreliable with as few as five individuals per cluster(Austin 2010), and high variability in random effects can bias estimates, making hierarchical models that incorporate random effects to account

for nested data superior in accuracy and fit to non-hierarchical models that assume independent observations.

## *2.3 Bayesian analysis of hierarchical model*

Bayesian regression estimates parameters as distributions by combining sample data with prior knowledge, making it useful for complex relationships, non-convergence in ML methods, and small samples (Finch, Bolin, and Kelley 2019). Non-hierarchical Bayesian models assume independence and ignore clustering, and with this risking bias, while hierarchical models assign priors to capture nested data structures.

For example here, the parameters $\beta_{0ij}$ and $\beta_{1ij}$ can be modelled with Gaussian priors as $\beta_{0ij}, \beta_{1i} \sim N(\mu, \sigma^2)$ treating the data as a single population ("complete pooling"). The response variable is modelled as $y_{ijk} | \pi_{ijk}, \theta \sim Ber(\pi_{ijk})$ with prior vector θ containing all level-specific β coefficients. For the compartment-specific intercepts, we assume (5), (6) and (7) repectively

$$\boldsymbol{\beta_{0i}} | \boldsymbol{\beta_0}, \boldsymbol{\sigma_0} \sim \boldsymbol{N}(\boldsymbol{\beta_0}, \boldsymbol{\sigma_0^2}), \tag{5}$$

$$\boldsymbol{\beta_0} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2}), \tag{6}$$

$$\boldsymbol{\sigma_0} \sim \boldsymbol{InvG}(\boldsymbol{a}, \boldsymbol{b}), \tag{7}$$

and the level-specific coefficients follow (8)

$$\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_p} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2}). \tag{8}$$

Continuous predictors are standardized and later back-transformed using

$$\widehat{\boldsymbol{\beta}}_{\mathbf{0}} - \sum_{\boldsymbol{m}=\mathbf{1}}^{\boldsymbol{p}} \widehat{\boldsymbol{\beta}}_{\boldsymbol{m}} \frac{\bar{\boldsymbol{x}}_{\boldsymbol{m}}}{\boldsymbol{s}_{\boldsymbol{m}}}, \tag{9}$$

$$\frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{m}}}{\boldsymbol{s}_{\boldsymbol{m}}}. \tag{10}$$

Overall, the full Bayesian model with highly informative priors is given by

$$\begin{aligned} \boldsymbol{log}\left(\frac{\boldsymbol{\pi}_{\boldsymbol{ijk}}}{\mathbf{1}-\boldsymbol{\pi}_{\boldsymbol{ijk}}}\right) &= \boldsymbol{\beta}_{\mathbf{0}} + \boldsymbol{\beta}_{\mathbf{1}\boldsymbol{ij}}\boldsymbol{pig}^{\mathbf{1}}_{\boldsymbol{ijk}} + \boldsymbol{\beta}_{\mathbf{2}\boldsymbol{i}}\boldsymbol{pen}^{\mathbf{2}}_{\boldsymbol{ij}} + \boldsymbol{\beta}_{\mathbf{3}}\boldsymbol{compartment}^{\mathbf{3}}_{\boldsymbol{i}} + \boldsymbol{u}_{\mathbf{0}\boldsymbol{i}} + \boldsymbol{u}_{\mathbf{0}\boldsymbol{ij}}, \\ \boldsymbol{\beta}_{\mathbf{0}} &\sim \boldsymbol{N}(\mathbf{0},\mathbf{1}) \\ \boldsymbol{\beta}_{\mathbf{1}},\ldots,\boldsymbol{\beta}_{\mathbf{15}} &\sim \boldsymbol{N}(\mathbf{0},\mathbf{1}) \\ \boldsymbol{\sigma}_{\mathbf{0}\boldsymbol{i}} &\sim \boldsymbol{InvG}(\mathbf{0.5},\mathbf{0.5}) \\ \boldsymbol{\sigma}_{\mathbf{0}\boldsymbol{ij}} &\sim \boldsymbol{InvG}(\mathbf{0.5},\mathbf{0.5}) \\ \boldsymbol{\sigma}_{\mathbf{0}} &\sim \boldsymbol{InvG}(\mathbf{0.5},\mathbf{0.5}). \end{aligned} \tag{11}$$

In this investigation, we compare one frequentist hierarchical model (Model 1) with three Bayesian models: non-hierarchical with non-informative priors (Model 2), hierarchical with non-informative priors (Model 3), and hierarchical with highly informative priors (Model 4).

Firstly, it should be noted that for all Bayesian models we use four Markov chains each with 5,000 iterations of which half are for warmup. This leaves us with a total of 10,000 post-warmup draws. Calculations are done with the brms R package, version 2.17.0, and the statistical software R, version 4.0.5 (R Core Team 2021). The brms package (Bürkner 2018; 2017; 2021), with the help of the rstan package (Stan Development Team 2024), uses the Stan platform to fit Bayesian hierarchical models. For further calculations and graphical representation ggmcmc (Fernández-

i-Marín 2016), ggplot2 (Wickham 2016), bayesplot (Gabry and Mahr 2024; Gabry et al. 2019), performance (Lüdecke et al. 2021), tidybayes (Kay 2024) and lme4 (Bates et al. 2015) were used.

### *2.4 Evaluation measures*

In this study, several goodness-of-fit measures were employed to assess and compare model performance in both Bayesian and frequentist frameworks. Two key metrics were the $R^2$ measures and the Akaike Information Criterion (AIC), along with its corrected version (AICC), as well as the Intraclass-Correlation Coefficient (ICC).

The $R^2$ measure is divided into two types for hierarchical models: the marginal $R^2$ ($R^2_m$) and the conditional $R^2$ ($R^2_c$) (Nakagawa and Schielzeth 2013). The marginal $R^2$ represents the variance explained solely by the fixed effects, while the conditional $R^2$ accounts for the total variance explained by both fixed and random effects. A large difference between these two indicates that a substantial portion of the variance is attributable to the grouping (random) effects, emphasizing the importance of properly modeling the hierarchical structure.

The ICC further breaks down the variance by measuring the proportion attributable to the random grouping factors (such as compartments or pens). This measure is critical for hierarchical models as it quantifies the degree of similarity within clusters. An adjusted ICC, which considers only the variance of the random effects relative to the total variance (random effects plus residual error), provides insight into the cluster-specific influence on the outcome.

In frequentist models, these measures are derived from maximum likelihood estimates and are used alongside the AIC and AICC to compare models. The AIC balances model fit against complexity, where lower values indicate a better trade-off between the goodness of fit and parsimony. The AICC further adjusts for small sample sizes, providing a more reliable basis for model.

# 3 Results

## *3.1 General Data Structure and Description of the Sample Population*

In this study, the variables collected were considered from a cross-sectional study. For this modelling exercise, a total of 29 variables were collected (see Table 1), whereby these variables are both qualitative and quantitative in nature. The basis descriptive measures of these variables are displayed in Table 22.

After necessary imputation steps, a complete data set with 29 variables is available for further analysis. The most important 16 variables are described descriptively in the following Table 2.

**Table 2: Generals descriptive measures of the sample population (n = 450 animals from 30 farms)**

| Quantitative variables | | | | |
|---|---|---|---|---|
| **Variable** | **mean** | **std dev** | **min** | **max** |
| **Age in days** | 51.842 | 14.099 | 28.000 | 89.000 |
| **Pen size in m²** | 12.061 | 4.530 | 5.760 | 25.750 |
| **Stocking density in animals/m²** | 0.377 | 0.116 | 0.170 | 1.080 |
| **Water flow rate in ml/min** | 928.600 | 489.033 | 100.000 | 2,200.000 |
| **Temperature in °C** | 27.829 | 1.984 | 22.000 | 32.800 |
| **Air pressure in Pa** | 1,010.724 | 9.539 | 989.000 | 1,032.200 |
| **$CO_2$ in ppm** | 2,178.222 | 918.326 | 800.000 | 5,000.000 |
| **$NH_3$ in ppm** | 7.517 | 5.582 | 2.000 | 30.150 |
| **$H_2S$ in ppm** | 0.615 | 0.465 | 0.000 | 2.000 |
| **Relative humidity in %** | 62.987 | 7.260 | 45.200 | 78.500 |
| **Qualitative variables** | | | | |
| **Variable** | **category** | | **n** | **%** |
| **Floor condition** | moderately worn | | 308 | 68.4 |
| | new | | 142 | 31.6 |
| **Skin injuries 1** | no | | 244 | 54.2 |
| | yes | | 206 | 45.8 |
| **Skin injuries 2** | no | | 365 | 81.1 |
| | yes | | 85 | 18.9 |
| **Pollution animals 1** | no | | 173 | 38.4 |
| | yes | | 277 | 61.6 |
| **Pollution animals 2** | no | | 345 | 76.7 |
| | yes | | 105 | 23.3 |
| **Cough** | no | | 264 | 58.7 |
| | yes | | 186 | 41.3 |

A total of 450 animals were included in this study. The average age of the animals was approx. 52 days, with an age range of 28 to 89 days. The average pen size was approximately 12 m², with a range of 5.760 to 25.750 m². The stocking density averaged 0.377 animals per $m^2$, i.e. 1 animal per 2.653 m². Floor conditions were predominantly classified as moderately worn (68.4%), while only a small proportion of pens were classified as new (31.6%). The water flow rate averaged 929 ml/min, with a range from 100 to maximum values of up to 2,200 ml/min. The average temperature was a pleasant 27.8 °C. The mean air pressure was approximately 1,010.7 Pa, with a range from a minimum of 989 Pa to a maximum of 1,032 Pa. The $CO_2$ value averaged about 2,178 ppm and ranged from 800 ppm to 5,000 ppm;

while the $NH_3$ value averaged about 7.5 ppm and varied between 2 ppm and 30 ppm. The $H_2S$ concentration averaged approximately 0.6 ppm with a range of 0 to 2 ppm and a relative humidity of approximately 63.0%, which varied between 45.2% and 78.5%.

Finally, 264 animals reported no coughing (58.7%), while 186 animals suffered from coughing (41.3%). For the analyses, we standardized the numerical variables and used them for the calculation. These results provide valuable insights into the health and husbandry conditions of the animal populations studied and their potential impact on animal welfare.

### *3.2 Results of Frequentist Models*

Starting the model selection process, we fit three hierarchical frequentist models (HFM) without any explanatory variables first to assess the impact of the cluster structure of the data. Table 3 shows the basic characteristics and measures of model fit for these basic models.

**Table 3: Measures for Model Specificity Comparing Hierarchical Frequentist Model (HFM) 1, 2 and 3**

| | **HFM 1** | **HFM 2** | **HFM 3** |
|---|---|---|---|
| **Random intercept (for each)** | Pens | Pens, Compartment | Pens, Compartment, Farms |
| **estimated Intercept** | -0.48 | -0.59 | -0.59 |
| **log-Likelihood** | -291.14 | -261.10 | -261.10 |
| **estimated variance pens** | 1.73 | <0.01 | <0.01 |
| **estimated variance compartments** | | 2.78 | 2.78 |
| **estimated variance farms** | | | <0.01 |
| **$ICC_c^2$** | 0.30 | 0.40 | 0.40 |
| **$ICC_{adj}$** | 0.34 | 0.46 | 0.46 |

From the model estimates (using the Laplace approximation; see Table 3 for details) the log-likelihood of a pig coughing in an "average" pen is estimated as $\widehat{\beta_0} = -0.48$, i.e., the probability of suffering from coughing without the influence of other variables is $\frac{exp(-0.48)}{1+exp(-0.48)}$=0.38 or 38 %. The adjusted intraclass-correlation ($ICC_{adj}$) shows that between 34 % and 46 %, respectively, of the variation (compared with the total variance) in the outcome variable cough can be explained by the respective clustering structure of the data in the models. For HFM2 and HFM3, the ICC-values are the same, as adding the farm level does not seem to provide any additional information. There is not enough additional farm-level variation to justify adding an additional random effect at this level to explain all of the observed variation. Although the HFM1 confirms that there is variation between pens, the magnitude of this variation can be nearly fully explained by the variation between compartments and the residual variance term. It should be noted that the HFM3 failed to converge, even though we have yet to add explanatory variables to the models. Therefore, the farm level was not further investigated in the model building process.

Before including explanatory variables in the model, the estimates of the compartmental effects or residuals, $\widehat{u_{0l}}$ have to be considered in more detail. These are calculated from the HFM2.

Figure 2 plots these conditional modes of the random compartment effect with all 63 compartments in total in rank order along with the associated 95 %-confidence intervals. The graph shows the estimated residuals for all compartments in the sample. For nine of the 63 total compartments, the 95 %-confidence intervals do not overlap with the horizontal line at zero, indicating that coughing in these compartments is above average. The confidence intervals are quite wide for some compartments, which is in line with the restricted sample sizes within these compartments. A corresponding graph (for HFM2) of the bay effects would simply consist of a horizontal line at zero, so this is not shown here.

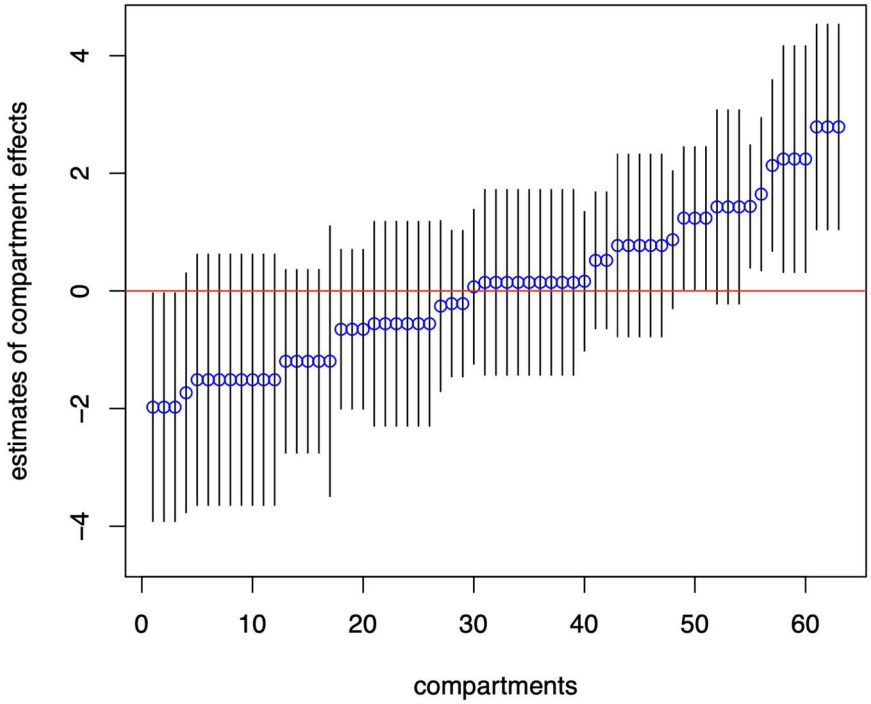



**Figure 2: Estimated Residuals per Compartment from Null-Model 2**

Subsequently, explanatory variables are included in the model in addition to the random intercepts. The starting point is initially the HFM2 from above. For model 1 the variables collected at pen level are included first. These are the age of pigs in a specific pen, the pen size or the stocking density in a pen and others. Model 1 assumes that the relationship of the explanatory variables with cough is the same across pens or compartments. Furthermore, we now add explanatory variables collected at the compartment level to our model. For both models, a random intercept is allowed for each of the pens or compartments. Table 4 shows the resulting odds ratios with associated 95 %-confidence intervals for the fixed effects of the fitted hierarchical logistic regression models, where model 1 has explanatory variables for pen level and model 2 has explanatory variables for the pen- and compartment level.

**Table 4: Fixed effects (Odds Ratios and associated 95%-confidence intervals; for categorical factors reference is underlined; statistically significant effects are marked bold) for fitted hierarchical logistic regression models**

| Factor | model 1 | model 2 |
|---|---|---|
| **age** (in days) | 1.38 [0.91, 2.09] | 1.27 [0.84, 1.93] |
| **pen size** (in $m^2$) | 0.79 [0.51, 1.24] | 0.84 [0.51, 1.37] |
| **stocking density** (in animals/$m^2$) | 1.25 [0.89, 1.75] | 1.38 [0.97, 1.98] |
| **animal pollution 1** (none vs. low) | 2.25 [0.94, 5.37] | 1.99 [0.82, 4.86] |
| **animal pollution 2** (none vs. high) | **3.28 [1.03, 10.47]** | 2.89 [0.89, 9.41] |
| **skin lesions 1** (none vs. low) | 1.22 [0.60, 2.51] | 1.52 [0.73, 3.17] |
| **skin lesions 2** (none vs. high) | 0.81 [0.32, 2.06] | 0.95 [0.37, 2.44] |
| **water flow rate** (in ml/min) | | **0.57 [0.33, 0.97]** |
| **temperature** (in °C) | | 3.85 [0.32, 46.68] |
| **air pressure** (in Pa) | | 0.70 [0.40, 1.23] |
| **$CO_2$-value** (in ppm) | | **2.12 [1.12, 4.01]** |
| **$NH_3$-value** (in ppm) | | 0.81 [0.49, 1.33] |
| **$H_2S$-value** (in ppm) | | 0.80 [0.48, 1.32] |
| **relative humidity** (in %) | | 0.61 [0.35, 1.09] |
| **floor condition** (as new vs. worn) | | **5.35 [1.72, 16.60]** |

Explanatory variables for the pen level have been considered in model 1. These are the pen size (in $m^2$), the age of the animals (in days), the stocking density within a pen, animals with skin lesions ("low"- or "medium"-/high-grade lesions in each case

in comparison with the reference category no lesions) and the degree of animal pollution within a pen.

Comparing the general outcome of both models show similar estimates at pen-level. Always all factors under study show no statistical significant effect. However, high animal pollution has a statistically significant effect on cough within model 1. The estimated odds ratio for high animal soiling is 3.28.

In model 2, compartment-specific variables have been added now. These are the water flow rate (in ml per minute), the measured temperature (in C), the air pressure (in Pa), the carbon dioxide value (CO2), the corrected ammonia value (NH3), the hydrogen sulfide value (H2S) (each in ppm) and the relative humidity (in %), each in one compartment. Also evaluated was the floor condition (as new vs. worn).

Three of these supplemental factors show a statistical significant effect, which again indicates the overarching hierarchical necessity of fitting nested models.

So far, only the fixed effects have been studied. However, both models have also allowed random intercepts for pens and compartments. The values of the estimated variances (VAR), standard deviations (SD), the log-likelihood function and intraclass-correlation (conditional and adjusted), for the comparison of the models are shown in Table 5. Here we notice that the estimated variances and standard deviations for the pen effects are close to zero for all our models. Adding explanatory variables significantly reduces the estimated variance across compartments, suggesting that the distribution of one or more variables varies across compartments.

**Table 5: Measures for Model Specificity Comparing HFM 2, Model 1 and Model 2**

| | HFM 2 | Model 1 | Model 2 |
|---|---|---|---|
| log-Likelihood | -261.10 | -254.00 | -245.60 |
| number parameters | 3 | 13 | 21 |
| estimated VAR pens | <0.01 | <0.01 | <0.01 |
| estimated SD pens | <0.01 | <0.01 | <0.01 |
| estimated VAR compartments | 2.78 | 2.66 | 1.77 |
| estimated SD compartments | 1.67 | 1.63 | 1.33 |
| $ICC_{adj}$ | 0.46 | 0.45 | 0.35 |
| $ICC_c$ | 0.46 | 0.42 | 0.28 |

The adjusted intraclass-correlation ($ICC_{adj}$) considers only the random effects in the model. Here, for model 2, a total of 35 % of the variation (compared with the total variation) in the outcome variable cough can be explained by the clustering structure of the data in this model. Compared with HFM 2 (46 %), this is quite low, implying that the addition of explanatory variables must help explain the variation in the outcome variable cough. However, the conditional $ICC_c$ (considers fixed and random effects) is slightly lower here at 28 %.

### *3.3 Results of Bayesian models*

Starting the model selection process from a Bayesian point of view, we ran a non-hierarchical Bayes model with all our explanatory variables. This model claimed that the stocking density, high animal pollution, the floor condition, the $H_2S$-value and the temperature in the pens are statistically significant variables for the response.

Since we know of the hierarchical structure of the data, this non-hierarchical model does not account for the pen and compartment effects. Therefore, we ran an intercept only model with varying intercepts for the hierarchical levels pen and compartment. The ICC-value for this model is $ICC_{adj}$ = 0.53, meaning that 53 % of

the variation in the outcome variable can be accounted for by the clustering structure of the data. Splitting this measure into the two hierarchy levels, we get $ICC_{pen}$ = 0.03 for the pen level and $ICC_{compartment}$ = 0.50 for the compartment level, which drives the decision to do not take the pen-level into further consideration.

This leaves us with Bayesian models with random intercepts for the compartment level and all explanatory variables. Within these we accounted for different kinds of (non- and high informative) prior distributions to our models as outlined in Table 66.

**Table 6: Measures for Model Specificity Comparing Bayesian Model (BM) 1-5**

| | |
|---|---|
| BM 1<br>(non-hierarchical) | $\beta_0 \sim \mathcal{N}\ (0, 50)$<br>$\beta_1, \dots, \beta_{15} \sim \mathcal{N}\ (0, 100)$<br>-<br>- |
| BM 2<br>(hierarchical)<br>(non-informative) | $\beta_0 \sim \mathcal{N}\ (0, 50)$<br>$\beta_1, \dots, \beta_{15} \sim \mathcal{N}\ (0, 100)$<br>$\sigma_0 \sim InvG\ (0.01, 0.01)$<br>$\sigma_{0i} \sim InvG\ (0.01, 0.01)$ |
| BM 3<br>(hierarchical)<br>(non-informative) | $\beta_0 \sim \mathcal{N}\ (0, 100)$<br>$\beta_1, \dots, \beta_{15} \sim \mathcal{N}\ (0, 1000)$<br>$\sigma_0 \sim InvG\ (0.01, 0.01)$<br>$\sigma_{0i} \sim InvG\ (0.01, 0.01)$ |
| BM 4<br>(hierarchical)<br>(high-informative) | $\beta_0 \sim \mathcal{N}\ (0, 1)$<br>$\beta_1, \dots, \beta_{15} \sim \mathcal{N}\ (0, 10)$<br>$\sigma_0 \sim InvG\ (1, 1)$<br>$\sigma_{0i} \sim InvG\ (1, 1)$ |
| BM 5<br>(hierarchical)<br>(high-informative) | $\beta_0 \sim \mathcal{N}\ (0, 1)$<br>$\beta_1, \dots, \beta_{15} \sim \mathcal{N}\ (0, 1)$<br>$\sigma_0 \sim InvG\ (0.5, 0.5)$<br>$\sigma_0 \sim InvG\ (0.5, 0.5)$ |

The resulting estimated odds ratios and their associated 95% credibility intervals were as follows (**Fehler! Verweisquelle konnte nicht gefunden werden.**): For

stocking density in BM 1 (non-hierarchical), it was estimated at 65.07 with a credibility interval of [49.46, 85.61]. In contrast, BM 2 (hierarchical; non-informative) showed an odds ratio of 13.55 [9.23, 19.90], while BM 3 (hierarchical; non-informative) had an odds ratio of 14.11 [9.57, 20.80]. For BM 4 (hierarchical; highly informative), stocking density yielded an odds ratio of 13.71 [9.30, 20.21], and BM 5 (also hierarchical; highly informative) resulted in an odds ratio of 7.24 [5.07, 10.36].

In our analysis, floor condition emerged as a significant variable influencing the occurrence of cough in pigs. The assessment categorized floor condition into two levels: "new" and "worn." The data revealed that 4.36 odds ratio (OR) for worn floor conditions indicates that pigs housed in compartments with worn flooring have approximately four times higher odds of exhibiting coughing symptoms compared to those in pens with new flooring.

**Table 7: Odds Ratios and 95% credibility intervals for fitted Bayesian models 1 to 5 (for categorical factors reference is underlined; statistically significant effects are marked bold)**

| Factor | BM 1 non-hierarchical | BM 2 hierarchical non-informative | BM 3 hierarchical non-informative | BM 4 hierarchical highly-informative | BM 5 hierarchical highly-informative |
|---|---|---|---|---|---|
| **stocking density** (in animals/$m^2$) | **65.07 [49.46,85.61 ]** | **13.55 [9.23, 19.90]** | **14.11 [9.57, 20.80]** | **13.71 [9.30, 20.21]** | **7.24 [5.07, 10.36]** |
| **pen size** (in $m^2$) | 0.94 [0.70, 1.25] | 0.97 [0.53, 1.75] | 0.97 [0.53, 1.74] | 0.97 [0.52, 1.78] | 0.96 [0.56, 1.67] |
| **age** (in days) | 1.01 [0.76, 1.34] | 1.02 [0.64, 1.64] | 1.02 [0.64, 1.63] | 1.02 [0.64, 1.63] | 1.02 [0.67, 1.57] |
| **floor condition** (as new vs. worn) | **4.19 [2.34, 7.70]** | **6.30 [1.56, 28.81]** | **6.47 [1.56, 28.80]** | **6.11 [1.53, 26.06]** | 3.04 [0.98, 8.94] |
| **water flow rate** (in ml/min) | 1.00 [0.74, 1.34] | 1.00 [0.49, 2.02] | 1.00 [0.50, 2.00] | 1.00 [0.51, 1.95] | 1.00 [0.55, 1.83] |
| **air pressure** (in Pa) | 0.70 [0.74, 1.27] | 0.97 [0.46, 2.01] | 0.96 [0.45, 2.04] | 0.96 [0.47, 1.99] | 0.98 [0.52, 1.84] |
| **$CO_2$-value** (in ppm) | 1.00 [0.73, 1.38] | 1.00 [0.43, 2.33] | 1.00 [0.43, 2.34] | 1.00 [0.43, 2.33] | 1.00 [0.50, 2.00] |
| **$NH_3$-value** (in ppm) | 0.97 [0.75, 1.26] | 0.95 [0.49, 1.87] | 0.95 [0.49, 1.85] | 0.95 [0.50, 1.84] | 0.97 [0.55, 1.74] |
| **$H_2S$-value** (in ppm) | **0.58 [0.45, 0.74]** | 0.60 [0.31, 1.15] | 0.61 [0.31, 1.20] | 0.62 [0.32, 1.20] | 0.75 [0.42, 1.36] |
| **temperature** (in °C) | **5.02 [1.24, 24.56]** | 4.01 [0.18,102.86] | 4.01 [0.17,110.77] | 3.78 [0.15, 96.97] | 1.38 [0.27, 6.91] |

**Table 7: Odds Ratios and 95% credibility intervals for fitted Bayesian models 1 to 5 (for categorical factors reference is underlined; statistically significant effects are marked bold)**

| Factor | BM 1 non-hierarchical | BM 2 hierarchical non-informative | BM 3 hierarchical non-informative | BM 4 hierarchical highly-informative | BM 5 hierarchical highly-informative |
|---|---|---|---|---|---|
| **relative humidity** (in %) | 0.97 [0.72, 1.30] | 0.92 [0.43, 1.97] | 0.92 [0.43, 1.99] | 0.92 [0.43, 1.96] | 0.94 [0.50, 1.79] |
| **skin lesions 1** (none vs. low) | 1.67 [1.00, 2.88] | 1.52 [0.65, 3.44] | 1.54 [0.66, 3.51] | 1.54 [0.68, 3.47] | 1.36 [0.68, 2.73] |
| **skin lesions 1** (none vs. high) | 1.22 [0.61, 2.45] | 0.90 [0.31, 2.61] | 0.93 [0.32, 2.62] | 0.92 [0.33, 2.58] | 0.87 [0.37, 2.09] |
| **animal pollution 1** (none vs. low) | 1.25 [0.64, 2.45] | 2.35 [0.94, 6.30] | 2.33 [0.94, 6.29] | 2.33 [0.90, 6.23] | 1.79 [0.81, 3.93] |
| **animal pollution 2** (none vs. high) | 2.28 [1.00, 5.32] | 3.22 [0.88, 12.18] | 3.21 [0.86, 12.46] | 3.25 [0.92, 12.26] | 2.16 [0.80, 5.99] |

This finding suggests that the quality of the flooring has a direct impact on respiratory health. Worn or degraded flooring can contribute to increased dust and pathogen exposure, leading to higher instances of respiratory issues among livestock. In this context, it is crucial for farm management practices to prioritize maintaining good floor conditions within pig housing facilities as part of overall biosecurity and animal welfare strategies. The results highlight the importance of addressing environmental factors such as floor condition when evaluating animal health outcomes.

In terms of model specificity measures comparing Bayesian Models (BM) 2 to 5, Table 8 summarizes key statistics including estimated random effects variance ($\hat{\sigma}_{0\ compartments}$) which was estimated at 1.87, 1.89, 1.83, and 1.74 across models BM 2 through BM 5 respectively.

**Table 8: Measures for Model Specificity Comparing BM 2-5**

| | BM 2 | BM 3 | BM 4 | BM 5 |
|---|---|---|---|---|
| $\hat{\sigma}_{0\text{compartments}}$ | 1.87 | 1.89 | 1.83 | 1.74 |
| $ICC_{adj}$ | 0.51 | 0.52 | 0.51 | 0.48 |
| $ICC_c$ | 0.41 | 0.42 | 0.41 | 0.41 |

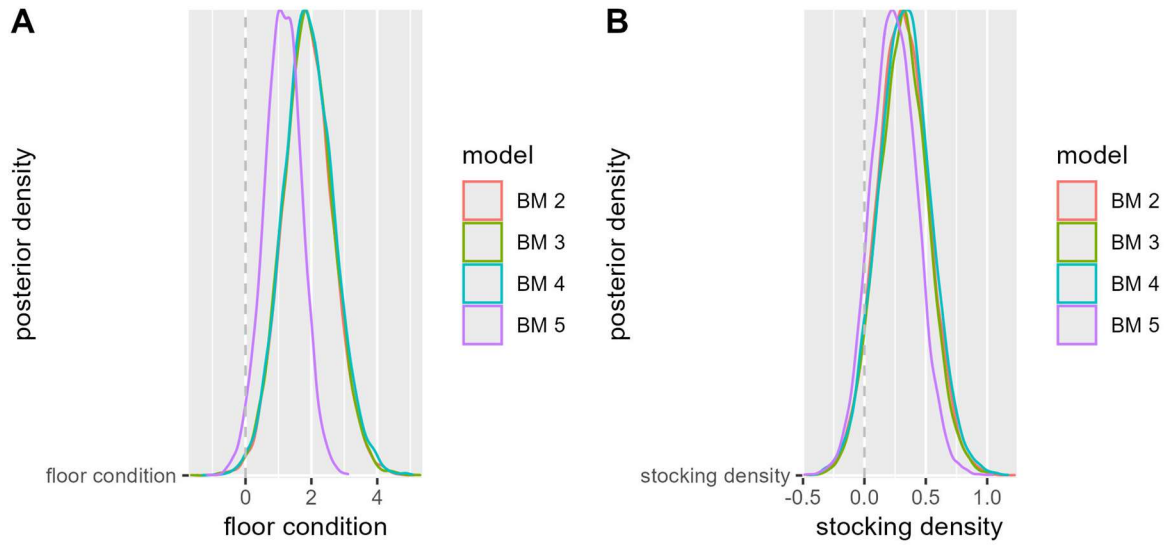



**Figure 3: Posterior density distribution for different models (BM 2 – BM 5) in relation to floor condition (A) and stocking density (B)**

Figure 3 shows the posterior densities for different models (BM 2 – BM 5) in relation to floor condition. Most models are very similar in their density distribution, with the exception of BM 5, which has a higher peak concentration. All models show a main distribution around a positive effect area, which indicates that floor properties have an overall positive influence on the target variable under consideration. The dashed zero reference line marks the boundary between positive and negative, i.e. here preventive effects. None of the models has a substantial density in the negative range, which indicates that restrictions in floor quality always result in increased respiratory problems.

All models (BM 2 – BM 5) relating to the variable “stocking density” show a similar distribution with a central maximum around a value close to zero. This indicates that the overall influence of stock density on the target variable under investigation is small or has no clear direction. The dashed zero reference line indicates where the effect would be neutral. As the distributions are closely grouped around this line, this could indicate that stock density in this collective of farms plays only a minor role for the outcome coughing.

It is noticeable that BM 5 (purple) has a slightly narrower distribution and a greater maximum value than the other models, which could indicate less uncertainty in this estimate. The remaining models (BM 2 – BM4) show a wider distribution, which could indicate greater uncertainty or variability in the estimate of the effect.

Overall findings indicate that while both hierarchical and non-hierarchical models provided insights into factors affecting cough incidence in pigs, modeling approaches that incorporate random effects offer more robust estimates by accounting for underlying data structures.

### *3.4 Evaluation Results*

At the beginning, three hierarchical frequentist models (HFMs) without explanatory variables were fitted to assess the impact of the data's clustering structure. The model estimates indicated that, in an "average" pen, pigs have a 38% probability of coughing without accounting for other influencing factors. Subsequently, models incorporating explanatory variables identified significant predictors - namely, stocking density and floor condition - as influential on the incidence of coughing among pigs.

Next, our approach was extended using Bayesian methods. First, all explanatory variables were included in non-hierarchical models, which revealed that stocking density and floor condition significantly influenced coughing incidence. Recognizing the hierarchical structure of the data, intercept-only models with varying intercepts across pen and compartment levels were estimated. These models showed an intra-class correlation (ICC) indicating that 53% of the variation was attributable solely to clustering effects, with the compartment level being dominant. The resulting estimated odds ratios across different Bayesian models highlighted significant variations, particularly with regard to stocking density; inflated estimates were observed when the hierarchical structure was not accounted for, compared to models that did include random effects.

The *loo* package in R was used to perform Leave-One-Out Cross-Validation (LOO), thereby assessing the out-of-sample predictive performance of the hierarchical

Bayesian models. Using the loo() function, expected log pointwise predictive density (elpd) scores were computed. The model comparison analysis based on LOO resulted in elpd differences for five Bayesian models. The results are summarized in Table 9.

**Table 9: Model comparison with the Bayesian models**

| Model | elpd_diff | se_diff |
|---|---|---|
| BM5 | 0.0 | 0.0 |
| BM3 | -1.6 | 1.6 |
| BM2 | -2.0 | 1.6 |
| BM4 | -2.7 | 1.5 |
| BM1 | -42.5 | 9.1 |

Generally model BM5 served as the reference with the best predictive performance, while BM1 showed a substantial drop in performance. Convergence diagnostics were monitored using the Rhat statistic derived from MCMC samples generated with tools like Stan (via the rstan package in R). All chains converged well (Rhat < 1.01), confirming the reliability of the parameter estimates.

Additionally, Bayes $R^2$ values were computed to assess the proportion of variance explained by the models. Table 10 summarizes the Bayes $R^2$ estimates for various models with and without the inclusion of random effects:

**Table 10: Bayes $R^2$ estimates for different models with and without random effects**

| Model | Estimate | Est.Error | Q2.5 | Q97.5 |
|---|---|---|---|---|
| **Random Effects Included** | | | | |
| **HFM1** | 0.314 | 0.048 | 0.214 | 0.400 |
| **HFM2** | 0.366 | 0.033 | 0.297 | 0.427 |
| **HFM3** | 0.367 | 0.033 | 0.300 | 0.428 |
| **BM1** | 0.173 | 0.023 | 0.126 | 0.218 |
| **BM2** | 0.401 | 0.027 | 0.345 | 0.453 |
| **BM3** | 0.401 | 0.028 | 0.343 | 0.452 |
| **BM4** | 0.413 | 0.028 | 0.355 | 0.464 |
| **BM5** | 0.395 | 0.029 | 0.333 | 0.448 |
| **Random Effects Not Included** | | | | |
| **HFM1** | 0.000 | 0.000 | 0.000 | 0.000 |
| **HFM2** | 0.000 | 0.000 | 0.000 | 0.000 |
| **HFM3** | 0.000 | 0.000 | 0.000 | 0.000 |
| **BM1** | 0.173 | 0.023 | 0.126 | 0.218 |
| **BM2** | 0.225 | 0.037 | 0.150 | 0.292 |
| **BM3** | 0.225 | 0.036 | 0.151 | 0.292 |
| **BM4** | 0.229 | 0.037 | 0.153 | 0.296 |
| **BM5** | 0.187 | 0.035 | 0.117 | 0.250 |

In summary, the Bayes $R^2$ results indicate that models including random effects provide significantly higher explained variance compared to models without them, especially evident in the HFMs which exhibit no explanatory power without random effects.

## 4 Discussion

The investigation presented here shows an extended version of a model building process for respiratory health in pig production. Therefore, in the discussion section we want to reflect on the implications of our findings in two dimensions by comparing different modelling approaches and their suitability for analyzing complex, hierarchical data first, and by discussing the findings from the viewpoint of veterinary advice to the farmers.

### *4.1 The Model Selection Process and its Characteristics*

The model selection was based on multiple criteria, including model fit, convergence issues, and the interpretability of results as recommended in Burnham & Anderson,2004 (Burnham and Anderson 2004) . Frequentist models offer a straightforward interpretation with clear estimates and confidence intervals, making them widely used in applied research (Agresti 2018). However, they can struggle with complex hierarchical structures and small sample sizes, potentially leading to biased estimates when the assumption of independence is violated (Snijders and Bosker 2012).

In contrast, Bayesian models integrate prior knowledge into the analysis, allowing for a more understanding of the data and the ability to handle hierarchical structures effectively (Gelman et al. 2013). They provide credible intervals that more accurately reflect uncertainty, especially in small samples or complex models (Gelman et al. 2013). The trade-offs, however, include a need for careful prior selection, which can be subjective and context-dependent (Kass and and Wasserman 1996), as well as increased computational intensity due to iterative simulation methods such as MCMC (Robert and Casella 2004). Hierarchical models, which account for the nested structure of data, further improve precision by modelling variability across different levels (Gelman and Hill 2007). While these models yield more stable and realistic estimates, they also introduce additional complexity in both model fitting and interpretation, requiring advanced diagnostics and greater computational resources (McElreath 2020).

The study compared four basic model types, summarized in Table 11.

**Table 11: Comparison overview of the four underlying models**

| Model | Model Characteristics | | | | |
|---|---|---|---|---|---|
| | Flexibility | Handles Clustering | Handles Small Data | Includes Prior Knowledge | Complexity |
| **Frequentistic Hierarchical** | Moderate | Yes | No | No | Low |
| **Bayesian Non-Hierarchical Non-Informative** | Moderate | No | Yes | No | Moderate |
| **Bayesian Hierarchical Non-Informative** | High | Yes | Yes | No | High |
| **Bayesian Hierarchical Highly Informative** | Very High | Yes | Yes | Yes | Very High |

Convergence issues were observed in models with a high number of explanatory variables, underlining the inherent complexity of hierarchical modeling. Notably, the Bayesian hierarchical model with highly informative priors (BM5) outperformed its counterparts by delivering the highest predictive accuracy, as evaluated using Leave-One-Out Cross-Validation, and by achieving higher Bayesian $R^2$ values that underscored the explained variance. This model provided robust and realistic estimates by appropriately accounting for the hierarchical structure and avoiding the overestimation of effects observed in non-hierarchical models such as BM1.

The goodness-of-fit measures employed in our analysis - namely the marginal and conditional $R^2$, the Intraclass Correlation Coefficient (ICC), and the AIC/AICC - provided valuable insights into the performance and appropriateness of our models. The marginal $R^2$ ($R^2_m$) quantified the proportion of variance explained solely by the fixed effects, while the conditional $R^2$ ($R^2c$) captured the overall explanatory power when both fixed and random effects were considered. This distinction underscored the importance of incorporating random effects to account for the hierarchical

structure inherent in our data. Similarly, the ICC offered a direct measure of the variability attributable to clustering, highlighting the degree of within-cluster similarity and reinforcing the necessity for hierarchical modeling.

Our results revealed that the Bayesian hierarchical model with highly informative priors (BM5) demonstrated superior performance compared to its Frequentist counterparts. Notably, BM5 achieved higher conditional $R^2$ and ICC values, suggesting that it more effectively captured both the systematic (fixed) and the random variability in the data. In contrast, the Frequentist models, while easier to interpret, tended to produce lower $R^2$ estimates and were more prone to inflated effect estimates when clustering was inadequately addressed. Furthermore, model comparison through AIC and AICC consistently favored the Bayesian approach, albeit with the caveat that its increased computational complexity and sensitivity to prior specification require careful management.

Some of these recommendations support the findings of our analysis, while others offer alternative perspectives that enrich the discussion. For example, Burnham and Anderson (2002) emphasize that model selection should be based on multiple criteria - such as parsimony, explanatory power, and convergence behavior - rather than relying solely on fit indices like AIC or BIC. This aligns with our approach of balancing interpretability and model performance. Similarly, McNeish and Stapleton (2016) caution against using complex hierarchical models in small-sample contexts without careful consideration, echoing our observation that hierarchical modeling can lead to instability if not adequately supported by the data. Conversely, other studies highlight the value of Bayesian approaches in sparse or nested data scenarios. For instance, Gelman and Hill (2007) and McElreath (2020) advocate for the use of multilevel Bayesian models, particularly when dealing with complex data structures and uncertainty across levels. These perspectives confirm that there is no one-size-fits-all solution in model selection; instead, the choice depends on the structure of the data, the research questions, and practical considerations such as computational cost and interpretability.

In summary, the integration of these goodness-of-fit measures into our evaluation not only validated our model selection but also highlighted the trade-offs between

the Bayesian and Frequentist paradigms. While Bayesian models offer enhanced flexibility and robustness in capturing complex hierarchical structures, they demand rigorous prior selection and greater computational resources. Conversely, Frequentist models provide simplicity and ease of interpretation but may fall short in accurately reflecting the underlying data structure, particularly in the presence of significant clustering effects.

### *4.2 Risk Factors for Coughing in Selected Farms with Sustainable Respiratory Problems*

The analysis of various models revealed that both frequentist and Bayesian approaches were able to identify key factors influencing the occurrence of coughing in pigs, including stocking density, floor condition, and water flow. Specifically, an increase in pen size by one square meter was associated with an odds ratio (OR) of 0.79, suggesting a preventive effect, although this result did not reach statistical significance. In contrast, pigs housed in areas with worn floor experienced more than five times the odds of coughing compared to those on new floor, highlighting the crucial environmental impact as an important surrogate for biosecurity. Additionally, an increased water flow rate demonstrated a protective influence (OR = 0.57), emphasizing the importance of adequate hydration.

These general results are of importance due to the farm population studied here. It has to be pointed out that the VASIB project was not a representative cross-sectional study of German pork production rather than s highly selected collection of farms with sustainable problems in respiratory health. It can therefore be assumed that the usual farm management measures and the continuous supervision by the herd veterinarian have already exhausted significant factors for improving animal health. Against this background, it is particularly remarkable that even in this collective, factors still appear to be significant which, from the point of view of animal hygiene and the associated biosecurity measures, can actually already be assumed to be known.

It can be concluded that this may indicate that certain influencing factors are either ignored in agricultural practice or cannot be implemented at all. For example, the

factor of floor condition, which is consistently considered to be conspicuous, is a factor that cannot be continuously improved, as this requires structural measures. By taking the hierarchical structure into account, however, there were indications of specific compartments with an increased impact, so that this can ultimately also be understood as an indication for the development of alternative hygiene concepts.

## 5 Conclusion and Outlook

This study demonstrates generally that hierarchical regression models are essential for accurately assessing respiratory health risks in pig farms. Frequentist models highlight the importance of clustering effects, while Bayesian approaches refine estimates using prior knowledge. Here, the Bayesian hierarchical model with highly informative priors (BM5) demonstrated the best performance by achieving the highest predictive accuracy, effectively accounting for hierarchical data structures, and reliably identifying significant predictors of coughing, such as stocking density, floor condition, and water flow rate. However, for other data structures these results on model selection may be different.

From the pig health perspective, the findings suggest that optimizing stocking density, maintaining flooring conditions, and monitoring air quality are key strategies for reducing sustainable respiratory disease in piglet production with sustainable problems in animal health. These aspects should be carefully examined in veterinary and farming practice.

## Ethics statement

The veterinary examinations described here were carried out as accompanying measures to therapeutic measures in the care of farms. Against the background of German animal welfare legislation, there was therefore no obligation to obtain formal authorization for the study.

However, participating farmer give written consent to all investigations and general pseudonymisation was performed for all data.

## Data Availability Statement

The data were collected on an individual basis from farmers and veterinary practitioners. Each participant provided written consent with the understanding that data would not be transferred to a third party. Therefore, any data transfer to interested persons is not allowed without an additional formal contract. Data are available to qualified researchers who sign a contract with the University of Veterinary Medicine Hannover. This contract will include guarantees of the obligation to maintain data confidentiality in accordance with the provisions of the German data protection law. Currently, there is no data access committee or another body who could be contacted for the data. However, for this purpose, a committee will be founded. This future committee will consist of the authors as well as members of the University of Veterinary Medicine Hannover. Interested cooperative partners who are able to sign a contract as described above may contact

Prof. Dr. Lothar Kreienbrock

Institute of Biometry, Epidemiology and Information Processing

University of Veterinary Medicine, Hannover

Bünteweg 2, 30559 Hannover

Email: lothar.kreienbrock@tiho-hannover.de

## Author's contributions

TT: Conceptualization, Formal statistical analysis, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing.

FM: Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing.

FS: Development study protocol, Data management, Data description, plausibility check,

DH: Development study protocol, Veterinary consulting, Writing – review and editing

LK: Funding acquisition, Study coordination, Development study protocol, Conceptualization, Supervision, Writing – review and editing

KI: Conceptualization, Supervision, Writing – review and editing.

## Funding

The project VASIB was supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme (Project No. 2817702014). We acknowledge financial support by the Open Access Publication Fund of the University of Veterinary Medicine Hannover, Foundation.

T.T. was supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, Project I1) funded by the German Research Foundation (DFG, Project Number 427806116).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

Agresti, Alan. 2018. *Statistical Methods for the Social Sciences (Fifth Edition)*. Harlow: Person.

Amenu, Kebede, K. Marie McIntyre, Nebyou Moje, Theodore Knight-Jones, Jonathan Rushton, and Delia Grace. 2023. “Approaches for Disease Prioritization and Decision-Making in Animal Health, 2000–2021: A Structured Scoping Review.” *Frontiers in Veterinary Science* 10 (October). https://doi.org/10.3389/fvets.2023.1231711.

Austin, Peter C. 2010. “Estimating Multilevel Logistic Regression Models When the Number of Clusters Is Low: A Comparison of Different Statistical Software Procedures.” *The International Journal of Biostatistics* 6 (1). https://doi.org/10.2202/1557-4679.1195.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using Lme4.” *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Bürkner, Paul-Christian. 2017. “Brms: An R Package for Bayesian Multilevel Models Using Stan.” *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

———. 2018. “Advanced Bayesian Multilevel Modeling with the R Package Brms.” *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

———. 2021. “Bayesian Item Response Modeling in R with Brms and Stan.” *Journal of Statistical Software* 100 (5): 1–54. https://doi.org/10.18637/jss.v100.i05.

Burnham, Kenneth P., and David R. Anderson, eds. 2004. *Model Selection and Multimodel Inference*. New York, NY: Springer New York. https://doi.org/10.1007/b97636.

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. “**Mice** : Multivariate Imputation by Chained Equations in *R*.” *Journal of Statistical Software* 45 (3). https://doi.org/10.18637/jss.v045.i03.

Dohoo, Ian R., Wayne Martin, and Henrik E. Stryhn. 2003. *Veterinary Epidemiologic Research*. University of Prince Edward Island.

Fernández-i-Marín, Xavier. 2016. “Ggmcmc: Analysis of MCMC Samples and Bayesian Inference.” *Journal of Statistical Software* 70 (9): 1–20. https://doi.org/10.18637/jss.v070.i09.

Finch, W. Holmes, Jocelyn E. Bolin, and Ken Kelley. 2019. *Multilevel Modeling Using R*. Vol. 2. CRC Press.

Gabry, Jonah, and Tristan Mahr. 2024. “Bayesplot: Plotting for Bayesian Models.” https://mc-stan.org/bayesplot/.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. “Visualization in Bayesian Workflow.” *J. R. Stat. Soc. A* 182 (2): 389–402. https://doi.org/10.1111/rssa.12378.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Akti Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. Third. Chapman & Hall/CRC Texts in Statistical Science Series. Boca Raton, Florida: CRC. https://stat.columbia.edu/~gelman/book/.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Grosse-Kleimann, Julia, Heiko Plate, Henning Meyer, Hubert Gerhardy, Corinna Elisabeth Heucke, and Lothar Kreienbrock. 2021. “Health Monitoring of Finishing Pigs by Secondary Data Use – a Longitudinal Analysis.” *Porcine Health Management* 7 (1): 20. https://doi.org/10.1186/s40813-021-00197-z.

Grund, Simon, Oliver Lüdtke, and Alexander Robitzsch. 2023. “Handling Missing Data in Cross-Classified Multilevel Analyses: An Evaluation of Different Multiple Imputation Approaches.” *Journal of Educational and Behavioral Statistics* 48 (4): 454–89. https://doi.org/10.3102/10769986231151224.

Kass, Robert E, and Larry and Wasserman. 1996. “The Selection of Prior Distributions by Formal Rules.” *Journal of the American Statistical Association* 91 (435): 1343–70. https://doi.org/10.1080/01621459.1996.10477003.

Kay, Matthew. 2024. “Tidybayes: Tidy Data and Geoms for Bayesian Models.” https://doi.org/10.5281/zenodo.1308151.

Lüdecke, Daniel, Mattan S Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. “Performance: An R Package for Assessment, Comparison and Testing of Statistical Models.” *Journal of Open Source Software* 6 (60): 3139. https://doi.org/10.21105/joss.03139.

Maes, D., M. Sibila, P. Kuhnert, J. Segalés, F. Haesebrouck, and M. Pieters. 2018. “Update on Mycoplasma Hyopneumoniae Infections in Pigs: Knowledge Gaps for Improved Disease Control.” *Transboundary and Emerging Diseases* 65 (May):110–24. https://doi.org/10.1111/tbed.12677.

McElreath, Richard. 2020. *Statistical Rethinking*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608.

Müller, Petra; Brauns, Jasmin; Kreienbrock, Lothar; Nathaus, Rolf; Höltig, Doris; Wendt, Michael; Kietzmann, Manfred; Meißner, Jessica. 2020. “Arzneimitteltherapie in der Ferkelaufzucht – wie sich Praxis und Wissenschaft die Hand reichen.” *Der praktische Tierarzt* 101 (10): 1006–15. https://doi.org/10.2376/0032-681X-2035.

Nakagawa, Shinichi, and Holger Schielzeth. 2013. “A General and Simple Method for Obtaining $R^2$ from Generalized Linear Mixed-effects Models.” *Methods in Ecology and Evolution* 4 (2): 133–42. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Pessoa, Joana, Jordi Camp Montoro, Telmo Pina Nunes, Tomas Norton, Conor McAloon, Edgar Garcia Manzanilla, and Laura Boyle. 2022. “Environmental Risk Factors Influence the Frequency of Coughing and Sneezing Episodes in Finisher Pigs on a Farm Free of Respiratory Disease.” *Animals* 12 (8): 982. https://doi.org/10.3390/ani12080982.

R Core Team. 2021. “R: A Language and Environment for Statistical Computing.” Vienna, Austria. https://www.R-project.org/.

Robert, Christian P., and George Casella. 2004. *Monte Carlo Statistical Methods*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-4145-2.

Robertson, Ian D. 2020. “Disease Control, Prevention and On-Farm Biosecurity: The Role of Veterinary Epidemiology.” *Engineering* 6 (1): 20–25. https://doi.org/10.1016/j.eng.2019.10.004.

Robitzsch, Alexander, and Simon Grund. 2025. “Miceadds: Some Additional Multiple Imputation Functions, Especially for ‘Mice.’” https://CRAN.R-project.org/package=miceadds.

Rubin, Donald B. 1986. “Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations.” *Journal of Business & Economic Statistics* 4 (1): 87. https://doi.org/10.2307/1391390.

Sanchez-Vazquez, M. J., W. D. Strachan, D. Armstrong, M. Nielen, and G. J. Gunn. 2011. “The British Pig Health Schemes: Integrated Systems for Large-scale Pig Abattoir Lesion Monitoring.” *Veterinary Record* 169 (16): 413–413. https://doi.org/10.1136/vr.d4814.

Snijders, T A B, and R J Bosker. 2012. *Multilevel Analysis : An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. Los Angeles ; London : SAGE. http://lib.ugent.be/catalog/rug01:001698339.

Stan Development Team. 2024. “RStan: The R Interface to Stan.” https://mc-stan.org/.

White, Ian R., Rhian Daniel, and Patrick Royston. 2010. “Avoiding Bias Due to Perfect Prediction in Multiple Imputation of Incomplete Categorical Variables.” *Computational Statistics & Data Analysis* 54 (10): 2267–75. https://doi.org/10.1016/j.csda.2010.04.005.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wilson, Jeffrey R., and Kent A. Lorenz. 2015. *Modeling Binary Correlated Responses Using SAS, SPSS and R*. Springer International Publishing.

# Eidesstattliche Versicherung (Affidavit)

Tug, Timur
Name, Vorname
(Surname, first name)

148066
Matrikel-Nr.
(Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Dortmund,15.04.2025
Ort, Datum
(Place, date)

_____
Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

"Integrative Statistical Methods for Analyzing Biomedical Data:

Applications in Health and Disease”

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Dortmund,15.04.2025
Ort, Datum
(Place, date)

_____
Unterschrift
(Signature)