Virtual Control Groups in Nonclinical Studies

Inaugural-Dissertation

to obtain the academic degree

Doctor of Philosophy in Natural Science (Ph.D. in Natural Science)

submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by

Alexander Gurjanov

Berlin, 2024

1st reviewer: Prof. Dr. Andrea Volkamer

2nd reviewer: Prof. Dr. Gerhard Wolber

Date of defense: 28.11.2024

Acknowledgements

My heartfelt gratitude to all who supported me in the completion of this work.
First, I would like to thank to my supervisors guiding me through the years. Thomas Steger-Hartmann, for his valuable toxicological support; Lea Vaas, whose statistical expertise was indispensable; Prof. Andrea Volkamer, for her supervision and the considerable opportunity in expressing my creativity in this work; Prof. Gerhard Wolber for giving me the possibility to perform the work at FU Berlin; and Joerg Wichard, for his supervision and initial onboarding at Bayer AG.

With this, my appreciation also goes to Bayer AG for providing the opportunity for the work, especially to my colleagues Annika Kreuchwig, Carlos Vieira-Vieira, and Adam Zalewski, whose diligent maintenance of the internal databases was crucial for my research. A special thanks to Julia Vienenkoetter and Alexius Freyberger for sharing their extensive knowledge as toxicity study directors, which greatly enriched my work.

I would be remiss if I did not acknowledge the never-ending support of my family, whose encouragement have been vital throughout the entire time. Last but not least, I am profoundly grateful to Yulia for her love and support, which have been my inspiration and motivation.

To all of you, I offer my deepest thanks.

Declaration of Authorship

I hereby declare that I alone am responsible for the content of my doctoral dissertation and that I have only used the sources or references cited in the dissertation.

Berlin, 10.05.2024 _____ Alexander Gurjanov

Table of Contents

Abstract

The concept of virtual control groups (VCGs) aims to replace concurrent control groups (CCGs) in nonclinical studies, potentially reducing the number of animals required in toxicity studies by up to 25%. To evaluate and develop this concept, a consortium comprising pharmaceutical companies, small and medium enterprises, and academic institutions was formed.

VCGs are generated from historical control data (HCD), i.e., control group data from past toxicity studies that complied with regulatory guidelines and are usually stored in a standardized data format. VCGs are created by a resampling approach meaning that animal data is randomly drawn from the HCD pool forming the VCGs. To assess the performance of VCGs, nonclinical toxicity studies in rodents (denoted as “legacy studies”) consisting of the CCG and three treatment groups served as a benchmark. “Well-performing” VCGs should reproduce the original results of legacy studies after substituting CCGs with VCGs—comprising statistical outcomes, test-substance related findings, and study conclusions on the test substance’s adverse effects.

This thesis shows that the reproducibility of original results with VCG significantly depends on careful selection of HCD in advance: HCD resembling the legacy-study animals in their endpoints’ value distribution leads to high reproducibility of statistical outcomes. Additionally, it was shown on three representative legacy studies that despite low to moderate reproducibility of statistical results, the overall conclusions of the toxicity studies still remained reproducible.

The results of the thesis highlight the need for rigorous selection and quality control of HCD including examination of distributions and time-dependent changes of endpoint values. This work introduces and discusses strategies to mitigate the effects of potential confounding factors in the data, aiming to create effective VCGs. Furthermore, the results highlight the need to judge the VCG performance beyond the reproducibility of statistical outcomes. Demonstrating the applicability of VCGs in nonclinical toxicity studies, this work provides workflows and procedures for the ongoing development of the VCG concept ultimately aiming for acceptance by regulatory bodies.

Zusammenfassung

Das Konzept der virtuellen Kontrollgruppen (VCGs) hat zum Ziel, gleichlaufende Kontrollgruppen (CCG) in nichtklinischen Studien zu ersetzen. Dadurch könnte die Zahl der in Studien benötigten Tiere um bis zu 25% reduziert werden. Um die Anwendbarkeit dieses Konzepts zu prüfen und weiterzuentwickeln, wurde ein Konsortium aus Mitgliedern der pharmazeutischen Industrie, kleinen und mittelständischen Unternehmen und akademischen Instituten gegründet.

VCGs werden aus historischen Kontrolldaten (HCD) erstellt, i.e., Kontrollgruppendaten aus vergangenen Toxizitätsstudien, welche sich an regulatorische Richtlinien hielten und in einem standardisierten Datenformat gespeichert werden. Um VCGs zu erstellen, wurde ein Resampling-Ansatz verwendet, i.e., Tierdaten werden zufällig aus dem HCD-pool gezogen und bilden die VCGs. Als Benchmark zur Bestimmung der VCG-Performance wurden nichtklinische Toxizitätsstudien in Nagern genutzt (genannt “Legacystudie“), bestehend aus einer CCG und drei Behandlungsgruppen. VCGs haben eine „gute Performance“ wenn sie die Originalergebnisse der Legacystudien reproduzieren nachdem die CCGs mit VCGs ersetzt worden sind—einschließlich statistischer Ergebnisse, testsubstanzbedingter Befunde sowie Schlussfolgerungen zu adversen Effekten durch die Testsubstanz.

Diese Arbeit zeigt, dass die Reproduzierbarkeit der Originalergebnisse durch die VCGs stark von der vorzeitigen Auswahl der HCD abhängt. HCD, welche der Legacystudie stark in der Verteilung ihrer Endpunktwerte ähneln, führen zu einer guten Reproduzierbarkeit der statistischen Ergebnisse. Weiterhin wurde am Beispiel drei repräsentativer Legacystudien wurde gezeigt, dass trotz niedriger bis moderater Reproduzierbarkeit statistischer Ergebnisse, die generellen Schlussfolgerungen der Toxizitätsstudien reproduziert werden konnten.

Die Ergebnisse dieser Arbeit verdeutlichen die Notwendigkeit einer gründlichen Vorauswahl und Qualitätskontrolle der HCD, was die Prüfung der Verteilungen und zeitabhängigen Veränderungen der Endpunkte beinhaltet. Die Arbeit präsentiert und diskutiert Strategien, um den Einfluss potenzieller Störfaktoren in den HCD zu begrenzen, alles zum Ziel, gut funktionierende VCGs zu erstellen. Außerdem heben die Ergebnisse die Notwendigkeit hervor, die Performance der VCGs nicht nur anhand der Reproduzierbarkeit statistischer Signifikanzen zu beurteilen. Diese Arbeit zeigt, dass VCGs in nichtklinischen Toxizitätsstudien verwendet werden können und stellt Workflows und Anleitungen zur Verfügung, die die Entwicklung des VCG-Konzepts voranbringen. Ziel ist es, dass das VCG-Konzept von Kontrollbehörden angenommen wird.

List of Abbreviations

3R	Replace, Reduce, Refine
BW	Body weight
CAT	Carnitine O-acetyltransferase
CCG	Concurrent control group
CDISC	Clinical Data Interchange Standards Consortium
CO	Clinical observations
DM	Demographics
ECOD	Ethoxycoumarin O-deethylase
EMA	European Medicines Agency
EPA	U.S. Environmental Protection Agency
EROD	Ethoxyresorufin-O-deethylase
FDA	U.S. Food and Drug Administration
FW	Food and water consumption
GLP	Good laboratory practice
GST	Glutathione S-transferase
HCD	Historical control data
HD	High dose
ICH	International Council for Harmonisation
IMI	Innovative Medicines Initiative
LB	Laboratory parameters
LD	Low dose
MA	Macroscopic findings
MD	Mid dose
MI	Microscopic findings
NO(A)EL	No observed (adverse) effect level
OECD	Organization for Economic Co-operation and Development
OM	Organ measurements
SEND	Standard for exchange of nonclinical data
SME	Subject matter expert
TS	Trial summary
VCG	Virtual control group

Abbreviations used outside of this thesis's main body relevant only to the supplementary material are listed in Table B1 within the supplementary materials of section 3.3.

1 Introduction

Bioassays are typically conducted with the inclusion of at least one treatment group receiving the test substance and a control group to correctly distinguish effects caused by the test substance from spontaneous effects (Kramer and Font, 2017). In early stages of drug development animals are used—most commonly rats and dogs (ICH, 2009; Prior et al., 2020)—for assessing the toxicity of drugs. There is a continuous strive for reducing the number of animals needed for studies due to ethical reasons as well as supply shortages of animals (Stephens and Mak, 2013; Steger-Hartmann and Clark, 2023), and to respond to this, the concept of virtual control groups (VCGs) in nonclinical studies was introduced (Steger-Hartmann et al., 2020). This concept aims to substitute the concurrent control group (CCG) in animal assays with VCGs, thus reducing the number of animals needed for a study by up to 25%. By that, VCGs may contribute to the 3R concept of Russel and Burch (1959), i.e., the strive to replace studies with alternative methods, reduce the number of animals needed, and refine procedures to minimize animal suffering.

The concept of using external control data instead of concurrent control arms comes from clinical studies (Pocock, 1976). In this field, external controls or virtual controls are in particular useful for studies where the usage of concurrent controls is unfeasible or unethical, for instance, if the outcome for an untreated patient—i.e., the control group—would be preventable death (Strayhorn, 2021). Regulatory authorities have already approved drugs on the basis of clinical trials that used external control groups in lieu of concurrent controls (Gökbuget et al., 2016).

The use of VCGs to contribute to the 3R concept in nonclinical animal studies is however rather novel. This concept was introduced by Steger-Harmann et al (2020). Originating from large data collection initiatives by the project eTOX (Sanz et al., 2017), and later eTRANSAFE (Sanz et al., 2023), both funded by the European Innovative Medicines Initiative (IMI), a consortium consisting of industry partners, small and medium enterprises (SMEs), and academic institutions was established. The aim of the VCG project is the collection and harmonization of cross-company data for creating VCGs, evaluate their performance and feasibility, and ultimately seek approval from regulatory authorities (Golden et al., 2023).

1.1 Nonclinical toxicity studies

To gain trust in the concept of VCGs, a key step is to examine whether the employment of VCGs yields similar results to using CCGs in nonclinical studies, denoted as “legacy studies”. Hence, the outcomes of toxicity studies following an established design serve as the benchmark, and replacing CCGs with VCGs must essentially replicate the study outcomes.

In the field of chemical and drug development, toxicology ensures the safety of a new product, accompanying the compound throughout the entire life cycle, from development, to market authorization, to distribution, to the eventual disposal (Horii, 2016). Regulatory authorities, such as US Food and Drug Administration (FDA) or the European Medicines Agency (EMA), control the authorization and registration of drugs (Murphy 2023). For a drug being approved, they must undergo highly regulated and standardized safety studies, following good laboratory practice (GLP) standards adhering to guidelines provided by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH, 2023) and the Organisation for Economic Co-operation and Development (OECD, 2023). There is a variety of studies in the field of regulatory toxicity, such as (sub)acute, (sub)chronic, pharmacokinetic, irritation and sensitization, immunotoxicity, carcinogenicity, and reproductive toxicity studies, each designed to address a specific question about the potential adverse effects—and reversibility of effects if recovery groups were included—of a compound on an organism (EMA, 2010; Avila et al., 2020; PBL, 2024). The specific aim of a 28-day sub-chronic toxicity study in rodents—the most frequently performed GLP study in nonclinical toxicology and therefore within the scope of this thesis (EPA, 2000; Chair, 2021)—is the establishment of a safe starting dose for first-in-human trials (ICH, 2009; Parasuraman, 2011; Steger-Hartmann and Clark, 2023). Ensuring the safety of a drug towards consumers necessitates the knowledge of the dose at which the compound starts showing toxicity, embodying Paracelsus’ principle that “only the dose makes the poison”.

Within the studies, a variety of endpoints are measured in the animals, as illustrated in Figure 1, including the following:

- "In life" parameters, i.e. observations on animal behavior (i.e. clinical observations), food and water consumption and body weight, measured daily during the dosing period to monitor the wellbeing of animals and to calculate the substance concentration for daily administration. If any of these parameters indicate a severe deterioration of the health status, animals are withdrawn from the study, or the dosage is reduced (Jourdan, 2013).
- Blood, serum, and urine parameters. In rodent 28-day toxicity studies, these parameters are taken once at the end of the drug-administration phase. If the study duration is longer or a larger species is used, these parameters are taken in regular intervals.
- Organ weights and histopathological parameters are taken post-mortem after the completion of the in-life phase.

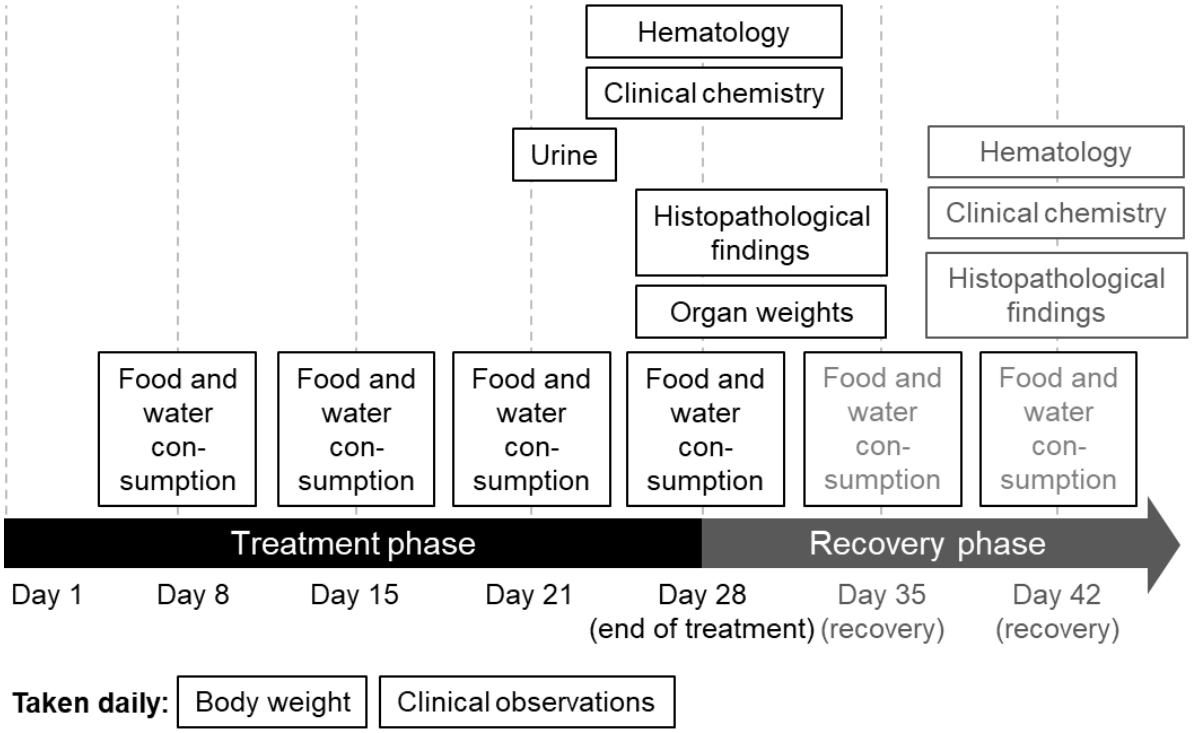

Figure 1: Measurements taken in an exemplary 28-day toxicity study in rodents followed by a 14-day recovery period.

Based on the collected animal data, study directors and subject matter experts (SMEs) can evaluate the evidence of the findings to determine the dosage at which adversity occurs. This evaluation procedure is outlined in Figure 2.

Regulatory guidelines require for that purpose that “when possible, numerical results should be evaluated by an appropriate and generally acceptable statistical method.” (OECD, 2008). Quantitative endpoints (e.g., body weight) are therefore assessed using statistical significance tests (Hothorn, 2014). For each measured endpoint one inferential test is assigned based on the presumed distribution of the respective endpoint’s values. The results are then classified as “statistically significant” if the *p*-value obtained from the significance tests has value of 0.05 or smaller.

However, conclusions about toxicity of a compound are not drawn on statistical outcomes alone (Steger-Hartmann and Clark, 2023). Study directors evaluate whether the resulting significant values are toxicologically relevant, i.e. determine whether the findings are related to the test substance or rather occurred spontaneously. For instance, if there is no dose-response relationship (the values or severities of the findings do not change monotonically with increasing dose), the effect might probably be of no toxicological relevance. Additionally, effects that fall within the range of variation of historical control data (see section 1.2), do not correlate with other findings, or transient findings are also likely to be considered as irrelevant by study directors.

Finally, the adversity of the test-substance related quantitative and qualitative findings is evaluated to determine the highest dose at which these effects are not seen, i.e., the NOAEL—the no observed adverse effect level (Palazzi et al., 2016; Baird et al., 2019). Adversity can be defined as the “impairment of functional capacity to maintain homeostasis and/or impairment of the capacity to respond to an additional challenge” (Palazzi et al., 2016). The NOAEL is determined by evaluating all findings related to the test substance and considering their evidence in a holistic manner. An isolated quantitative finding in liver parameters for instance might provide less evidence for liver damage than a pathological finding in liver tissue, and therefore be classified as non-adverse (Baird et al., 2019).

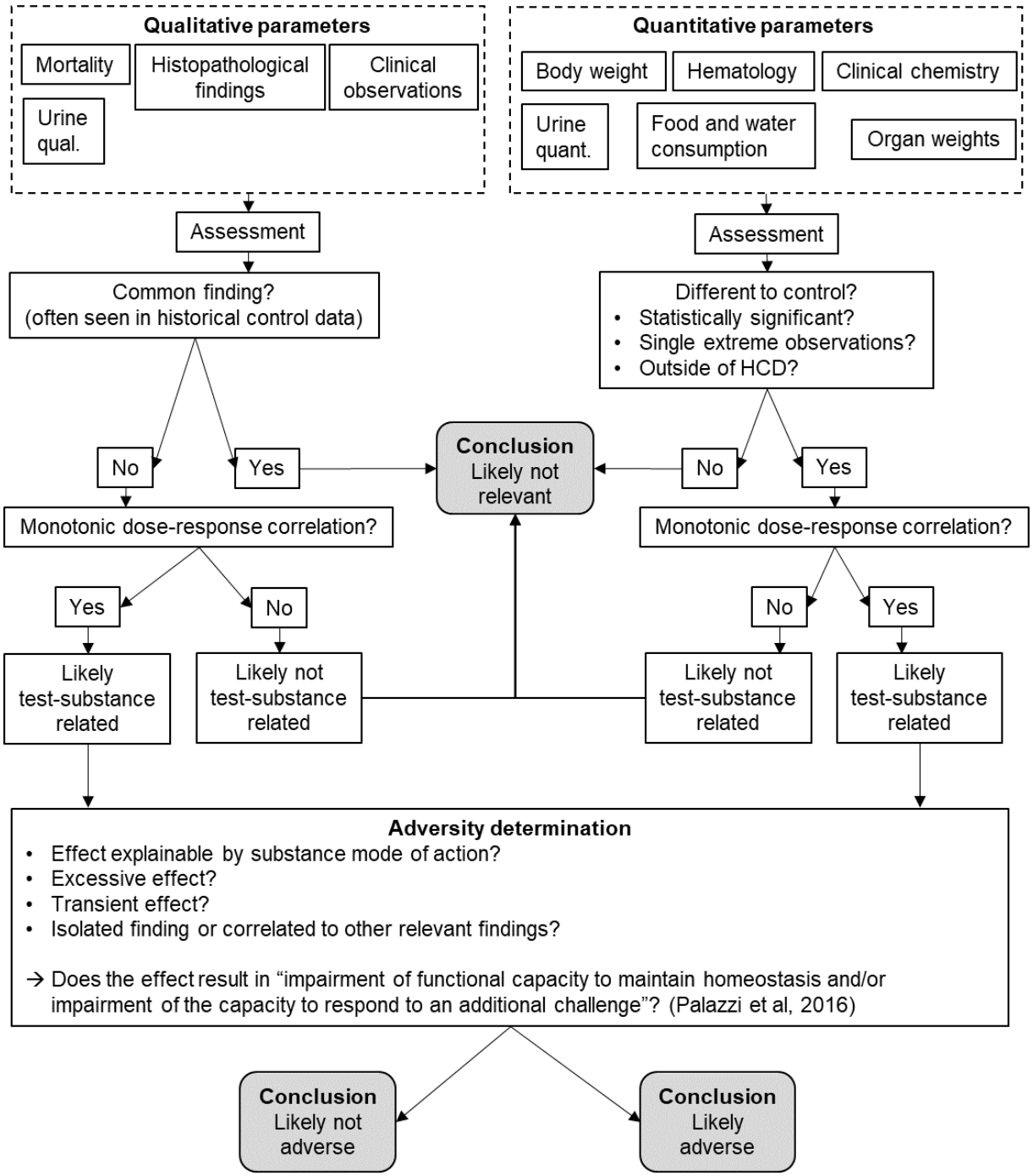

Figure 2: Workflow for adversity determination in a sub-chronic study in rodents.

1.2 Historical control data

A pivotal step in the assessment of test-substance-relatedness is the comparison of the findings against control data of earlier studies, i.e. historical control data (HCD). This is used as a representation of the background variance of controls (Kluxen et al., 2021). If numerical findings are within the relevant ranges of historical control data (OECD, 2008; Menssen, 2023), or specific qualitative findings are frequently seen in HCD (Golden et al., 2023; Grevot et al., 2023), it is likely that they are of no toxicological relevance.

Apart from using HCD for decision making in toxicity studies, they are also an essential component for the development of VCGs, since VCGs are derived from HCD (Steger-Hartmann et al., 2020; Kluxen et al., 2021).
The requirement for using HCD for comparative purposes in the context of regulatory toxicology is that they are “originating from the same laboratory, species, strain, and collected under similar conditions” (OECD, 2018a). HCD collected from studies performed “under similar conditions” is also a key requirement for generating well-performing VCGs. It was shown that the performance of VCGs for reproducing statistical results improves if the study design parameters of the chosen HCD, such as species, strain, sex, route of administration, and treatment vehicle, are as similar as possible to the legacy study (Wright et al., 2023). Furthermore, to counter the effects of genetic drift, the selection of HCD should be from a specified period (Golden et al., 2023). Yet, it is still unknown which HCD study design parameters need to be aligned with the legacy study, and which can safely be combined (Golden et al., 2023).

1.3 SEND structure

A key component for ensuring the usability of HCD for comparing animal values across different studies is that HCD are recorded and stored in an identical structure. A harmonic data structure in accordance with specific formats is also required for submitting data to regulatory authorities. In the nonclinical toxicology field, data is stored under the standard for exchange of nonclinical data (SEND) structure (CDISC, 2022). The implementation of SEND in 2002 (Wood, 2011) facilitated data storage in a harmonized format, and since 2016, submission of all nonclinical data in SEND format to the FDA is a prerequisite for drug market approval (CDISC, 2022). The study data in the SEND format is divided into so-called domains, comprising:

- Quantitative parameters: such as body weight (BW); laboratory parameters, i.e. hematology, clinical chemistry, and urine (LB); organ measurements (OM); and food and water consumption (FW).

- Qualitative parameters: such as histopathological findings, i.e. macroscopic findings (MA); microscopic (MI); and clinical observations (CO).
- Demographics of the animal (DM): animal information on strain, species, supplier, birth data, etc.
- Trial summary (TS): study design information, such as study date, study duration, route of administration, etc.

The SEND format not only ensures the comparability of different studies within and across different companies, but the vast quantity of data that follows a uniform structure serves as the cornerstone of creating and assessing VCGs (Steger-Hartmann et al., 2020).

1.4 Aim of the study

Utilizing the internal HCD of Bayer AG, stored in the SEND format, VCGs can be created. Steger-Harmann et al. (2020) proposed two ways of creating VCGs: the resampling approach and the simulation approach. In the resampling approach, animals are sampled directly in a randomized fashion from the HCD, each animal encompassing all measured endpoints. The simulation approach in turn uses the distribution information of HCD endpoints for synthetic generation of the values. Currently, the emphasis is on the resampling method (Golden et al., 2023) as this maintains the GLP-status of the measurements and preserves potential correlations among them. The research hypothesis of this work is that the substitution of concurrent control groups (CCGs) with virtual control groups (VCGs), generated from historical control data (HCD) with a resampling approach, will allow for the reproduction of toxicity study outcomes, comprising statistical results, test-substance related findings, and study conclusions. The performance of VCGs can be evaluated by their ability to reproduce the results of nonclinical 28-day toxicity studies in rats. The objective of this thesis was to accurately replicate the evaluation process of a conventional 28-day toxicity study which involved maintaining the original study design and evaluation methods with the only deviation being the substitution of CCGs with an identical number of VCG animals. Consequently, the group sizes remained unchanged, univariate inferential statistics were applied, and HCD served for assessment for test-substance-relatedness. Diverging from this evaluation process through the incorporation of mixed-effect models, effect sizes (Schmidt et al., 2016), or Bayesian models (Kramer and Font, 2017)—while being viable options given the applicability of HCD—was outside of the scope of this work.

To test the research hypothesis this thesis:

- Developed a method for identification of suitable HCD by analysis of their value distribution and time-dependent trends. It presents the consequences of selecting

unsuitable HCD for VCG generation on the example of serum-electrolyte values in rats and presents ways to counter the effect of confounding factors.

- Presents different methods of sampling virtual controls from HCD to obtain the best performing VCG. It underlines the importance of matching HCD to the legacy study by animal age, or body weight as a surrogate for age, to obtain high-performing VCGs able to reproduce statistical outcomes for body weights.
- Assesses the performance of VCGs on entire studies comprising all results of qualitative and quantitative endpoints. It highlights the necessity of evaluating not only statistical reproducibility but also to include the assessments of test-substance-relatedness and study conclusions on adverse effects into the validation framework.

To the best of the author's knowledge, this thesis is the first to demonstrate the use of VCGs on entire toxicity studies, laying a foundation for future development of the VCG concept.

1.5 References

Avila, A.M., Bebenek, I., Bonzo, J.A., Bourcier, T., Bruno, K.L.D., Carlson, D.B., et al. (2020). An FDA/CDER perspective on nonclinical testing strategies: Classical toxicology approaches and new approach methodologies (NAMs). Regulatory Toxicology and Pharmacology 114, 104662. https://doi.org/10.1016/j.yrtph.2020.104662.

Baird, T.J., Caruso, M.J., Gauvin, D.V., Dalton, J.A., 2019. NOEL and NOAEL: a retrospective analysis of mention in a sample of recently conducted safety pharmacology studies. J. Pharmacol. Toxicol. Methods 99, 106597. https://doi.org/10.1016/j.vascn.2019.106597.

CDISC, C.D.I.S.C. (2022). SEND [Online]. CDISC: CDISC. Available: https://www.cdisc.org/standards/foundational/send (Accessed March 29th 2024 2024).

Chair, I. (2021). 4446 Clinical Signs of Pain and Disease in Laboratory Animals [Online]. Yale University: Yale Office of Animal Research Support (OARS). Available: https://your.yale.edu/policies-procedures/guides/4446-clinical-signs-pain-and-disease-laboratory-animals (Accessed August 19, 2023 2023).

EMA (2010). “CPMP/SWP/1042/99 Rev 1 Corr* - guideline on repeated dose toxicity,” in European Medicines agency. [Online]. Available: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-repeated-dose-toxicity-revision-1_en.pdf (Accessed September 23 2023).

EPA (2000). Health Effects Test Guidelines: OPPTS 870.3050 Repeated Dose 28–Day Oral Toxicity Study in Rodents [Online]. Washington D.C., U.S.: EPA. Available: https://ntp.niehs.nih.gov/sites/default/files/iccvam/suppdocs/feddocs/epa/epa_870_3050.pdf (Accessed Apr 6, 2024 2024).

Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., et al. (2016). Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood cancer journal 6(9), e473-e473. https://doi.org/10.1038/bcj.2016.84.

Golden, E., Allen, D., Amberg, A., Anger, L.T., Baker, E., Baran, S.W., et al., (2023). Toward implementing virtual control groups in nonclinical safety studies: workshop report and roadmap to implementation. ALTEX-Alternative Anim. Exp. https://doi.org/10.14573/altex.2310041.

Grevot, A., Boisclair, J., Guffroy, M., Hall, P., Pohlmeyer-Esch, G., Jacobsen, M., et al. (2023). Toxicologic Pathology Forum Opinion Piece: Use of Virtual Control Groups in Nonclinical Toxicity Studies: The Anatomic Pathology Perspective. Toxicologic Pathology, 01926233231224805. https://doi.org/10.1177/01926233231224805.

Horii, I. (2016). The principle of safety evaluation in medicinal drug-how can toxicology contribute to drug discovery and development as a multidisciplinary science? The Journal of Toxicological Sciences 41(Special), SP49-SP67. https://doi.org/10.2131/jts.41.sp49.

Hothorn, L.A. (2014). Statistical evaluation of toxicological bioassays–a review. Toxicology Research 3(6), 418-432. https://doi.org/10.1039/c4tx00047a.

ICH (2009). Guidance on nonclinical safety studies for the conduct of human clinical trials and marketing authorization for pharmaceuticals m3 (r2). International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-m3r2-non-clinical-safety-studies-conduct-human-clinical-trials-and-marketing-authorisation-pharmaceuticals-step-5_en.pdf (Accessed Apr 7 2024).

ICH (2023). Safety Guidelines [Online]. Geneva, Switzerland: ICH. Available: https://www.ich.org/page/safety-guidelines (Accessed Apr 2, 2024).

Jourdan, T. (2013). Empfehlungen der Berliner Tierschutzbeauftragten zu Score Sheets und Abbruchkriterien [Online]. Unterfranken: Regierung Unterfranken. Available: https://www.regierung.unterfranken.bayern.de/mam/aufgaben/bereich5/sg54/score_s

heet_empfehlungen-der-berliner-tschb-zu-abbruchkriterien.pdf (Accessed January 2 2024).
Kluxen, F. M., Weber, K., Strupp, C., Jensen, S. M., Hothorn, L. A., Garcin, J.-C., et al. (2021). Using historical control data in bioassays for regulatory toxicology. Regul. Toxicol. Pharmacol. 125, 105024. https://doi.org/10.1016/j.yrtph.2021.105024.
Kramer, M., and Font, E. (2017). Reducing sample size in experiments with animals: historical controls and related strategies. Biological Reviews 92(1), 431-445. https://doi.org/10.1111/brv.12237.
Menssen, M. (2023). The calculation of historical control limits in toxicology: Do's, don'ts and open issues from a statistical perspective. Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 503695. https://doi.org/10.1016/j.mrgentox.2023.503695.
Murphy , K. (2023). Regulating and Authorizing Medicines: A Comparison of the FDA and EMA [Online]. Xtelligent Healthcare Media, U.S.: PharmaNewsIntelligence. Available: https://pharmanewsintel.com/features/regulating-and-authorizing-medicines-a-comparison-of-the-fda-and-ema (Accessed Apr 3, 2024 2024).
OECD (2008). Test no. 407: Repeated dose 28-day oral toxicity study in rodents. https://doi.org/10.1787/9789264070684-en.
OECD (2018a). Test No. 408: Repeated Dose 90-Day Oral Toxicity Study in Rodents. https://doi.org/10.1787/20745788.
OECD (2023). OECD Test Guidelines for Chemicals [Online]. Paris, France: OECD. Available: https://www.oecd.org/chemicalsafety/testing/oecdguidelinesforthetestingofchemicals.htm (Accessed Aprl 2, 2024).
Palazzi, X., Burkhardt, J.E., Caplain, H., Dellarco, V., Fant, P., Foster, J.R., et al. (2016). Characterizing “adversity” of pathology findings in nonclinical toxicity studies: Results from the 4th ESTP international expert workshop. Toxicologic Pathology 44(6), 810-824. https://doi.org/10.1177/0192623316642527.
Parasuraman, S. (2011). Toxicological screening. Journal of pharmacology & pharmacotherapeutics 2(2), 74. https://doi.org/10.4103%2F0976-500X.81895.
PBL, P.B. (2024). Regulatory Toxicology Studies [Online]. San Francisco Bay Area 551 Linus Pauling Drive, Hercules CA 94547: PBL, Pacific Biolabs. Available: https://pacificbiolabs.com/regulatory-tox (Accessed Apr 1, 2024 2024).
Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. J. chronic Dis. 29 (3), 175–188. https://doi.org/10.1016/0021-9681(76)90044-8.
Prior, H., Haworth, R., Labram, B., Roberts, R., Wolfreys, A., and Sewell, F. (2020). Justification for species selection for pharmaceutical toxicity studies. Toxicology Research 9(6), 758-770. https://doi.org/10.1093/toxres/tfaa081.
Russel, W. M. S. and Burch, R. L. . The principles of humane experimental technique. Methuen, (1959). http://117.239.25.194:7000/jspui/bitstream/123456789/1342/1/PRILIMINERY%20%20AND%20%20CONTENTS.pdf.
Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., Asakura, S., Amberg, A., et al., (2023). eTRANSAFE: data science to empower translational safety assessment. Nat. Rev. Drug Discov. https://doi.org/10.1038/d41573-023-00099-5.
Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., Cases, M., et al., (2017). Legacy data sharing to improve drug safety assessment: the eTOX project. Nat. Rev. Drug Discov. 16 (12), 811–812. https://doi.org/10.1038/nrd.2017.177.
Schmidt, K., Schmidtke, J., Kohl, C., Wilhelm, R., Schiemann, J., van der Voet, H., et al., (2016). Enhancing the interpretation of statistical P values in toxicology studies: implementation of linear mixed models (LMMs) and standardized effect sizes (SESs). Arch. Toxicol. 90, 731–751. https://doi.org/10.1177/0192623313517771.
Steger-Hartmann, T. and Clark, M. (2023). Can historical control group data be used to replace concurrent controls in animal studies? Toxicologic Pathology 01926233231208987. https://doi.org/10.1177/01926233231208987.
Steger-Hartmann, T., Kreuchwig, A., Vaas, L., Wichard, J., Bringezu, F., Amberg, A., et al. (2020). Introducing the concept of virtual control groups into preclinical toxicology

testing. ALTEX-Alternatives to animal experimentation 37(3), 343-349. https://doi.org/10.14573/altex.2001311.

Stephens, M.L., and Mak, N.S. (2013). History of the 3Rs in toxicity testing: From Russell and Burch to 21st century toxicology.

Strayhorn, J.M. (2021). Virtual controls as an alternative to randomized controlled trials for assessing efficacy of interventions. BMC Medical Research Methodology 21(1), 1-14.

Wood, F. K., and Lou, A. (2011). The standard for the Exchange of nonclinical data (SEND): History and basics. https://www.pharmasug.org/proceedings/2011/CD/PharmaSUG-2011-CD14.pdf. (Accessed Janury 3, 2022).

Wright, P. S., Smith, G. F., Briggs, K. A., Thomas, R., Maglennon, G., Mikulskis, P., et al. (2023). Retrospective analysis of the potential use of virtual control groups in preclinical toxicity assessment using the eTOX database. Regul. Toxicol. Pharmacol. 138, 105309. https://doi.org/10.1016/j.yrtph.2022.105309.

2 Materials and Methods

This chapter provides an overview of the VCG creation and evaluation process. Detailed procedures, including in-depth description of steps respective to each experiment, can be found in the corresponding publications presented in the Results section. The general process is illustrated in Figure 3. Throughout this thesis, the following terms are frequently used:

- *Legacy study*: toxicity study conducted in the past. Studies comprise treated groups and a concurrent control group (CCG). The legacy study results serve as the benchmark for assessing VCG performance.
- *Study design parameters*: parameters describing the methodology of the toxicity study, such as the selection of species, strain, route of administration, treatment vehicle, etc.
- *Endpoint*: values or observations measured in animals to assess the outcome of a test substance in a study.
- *Historical control data* (HCD): control data from toxicity studies performed in the past. HCD is used for comparative purposes in toxicology. In this thesis, HCD is also used to generate virtual control groups.
- *Virtual control groups* (VCGs): control-group animal data generated from HCD.
- *Initial body weight*: animal body weight measured on day 1, i.e., the beginning of the study prior to first application of the test substance.
- *Sentinel animals*: CCG animals which were not replaced by VCGs and remained in the set of the legacy study.

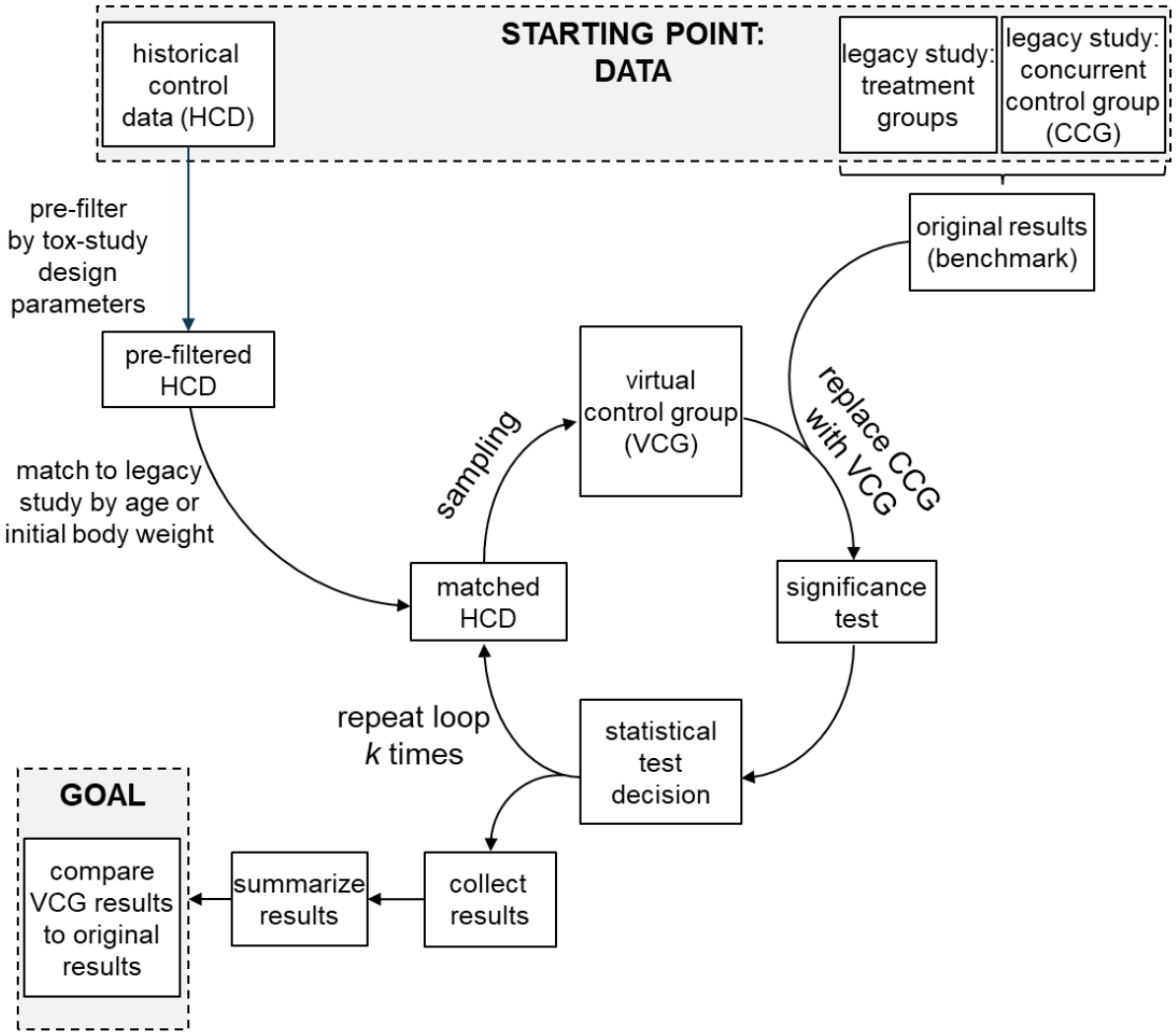

Figure 3: Overview of the VCG generation and performance evaluation.

2.1 Data and code

Internal data was collected and evaluated using laboratory information management systems (LIMS) provided by Xybion Digital Inc (Xybion, 2024). All animal studies stored in SEND (standard for exchange of nonclinical data) format. Data gathering, harmonization, (statistical) evaluation, and visualization was performed with the statistical software R (R Core Team, 2021). A list of libraries used can be found in the respective within the results of this thesis. The used R-code can be downloaded from Bayer's open-source GitHub repositories.

2.2 Selection of HCD

HCD is an essential part of creating VCGs. HCD from sub-chronic toxicity studies in rats was extracted from internal data repositories. The studies in the HCD were conducted in a maximum similar design, i.e. they followed these specific design parameters:

- Dosing duration of 4 weeks.

- Wistar HAN rats were used.
- Animals were randomized into four groups: control, low dose, mid dose, and high dose.
- Test substance was administered daily orally by gavage.
- All studies were conducted in the same test facility of Bayer AG, Wuppertal, Germany.
- The treatment vehicle was a mixture of ethanol, water, and either Kolliphor® HS 15 or polyethylene glycol 400.
- Animals were supplied by either Charles River Laboratories, Germany, or Harlan Netherlands.
- Animals were housed in group cages with 2–3 animals per cage.
- Animals had an ad libitum supply of food and water.
- Body weight was measured daily.
- Food and water consumption was measured weekly.
- Clinical pathology and histopathology parameters were measured on day 28 ± 7.

The distribution and time-dependent trends of the endpoint values in HCD were inspected via the use of histograms and time-series charts. HCD was then further used as a pool of animals where VCGs were extracted from.

2.3 Resampling

HCD is essentially a pool of animal data, with each animal representing a collection of endpoint values corresponding to that specific animal. Animals were randomly sampled from the HCD pool and formed a VCG. The goal was to match the number of VCG animals to the CCG animals used in the legacy study. For instance, if the original study used 10 animals per sex, the combined number of VCG animals and sentinel animals (if used) aimed to be 10. In cases where the available HCD did not provide enough animals to sample from, all available animals were used.

2.4 Benchmarking the VCG performance

The respective statistical tests assigned to each endpoint were extracted from study reports and protocols provided by Xybion Digital Inc. The statistical tests can be found in the supplementary material of section 3.3. The p value obtained from the statistical tests was then used to categorize the differences between control and dose groups as either “statistically significant” ($p \leq 0.05$) and “not statistically significant” ($p > 0.05$). The statistical tests were performed in two rounds: initially using CCGs and then after substituting CCGs with VCGs. The statistical outcomes of both were then compared to evaluate the reproducibility.

The procedure of animal sampling and subsequent recalculation of statistical tests was repeated 100-1000 times, depending on the context and the complexity of the data to increase the robustness of the VCG assessment.

Within each repetition, the original result was compared to the result following the VCG substitution. The number of successfully replicated results was then summarized as “reproducibility percentage”.

Apart from statistical outcomes, the number of test-substance related findings and the study conclusions were compared between the legacy study and the results after replacing CCGs with VCGs. This formed the basis for the VCG performance evaluation.

2.5 References

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Xybion (2024). *XYBION All-In-One Cloud LIMS for Life Sciences, R&D, and Labs with a Full QMS Suite* [Online]. 105 College Road East, Princeton, New Jersey 08540: XYBION. Available: https://www.xybion.com/ (Accessed Apr. 1, 2024).

3 Results

This chapter presents the outcomes of applying virtual control groups (VCGs) in nonclinical studies and is divided into three sections. Each section features a publication that highlights specific findings with regard to the performance and applicability of VCGs. The supplementary material to each corresponding publication is provided directly below the respective article.

3.1 Hurdles and signposts on the road to virtual control groups—A case study illustrating the influence of anesthesia protocols on electrolyte levels in rats

The initial and vital step in creating well-performing VCGs is the selection of suitable HCD. It is necessary to inspect the HCD values for any unusual distributions or time-dependent trends that could result from undetected confounding factors. This study illustrates the consequences of ignoring such confounders on the example of serum electrolyte values in rats. Furthermore, this study presents the approach of generating VCGs by resampling in the context of nonclinical studies. It shows that the performance of VCGs to reproduce statistical results of legacy studies can be significantly improved by careful selection and analysis of HCD and the use of sentinel animals.

Authors: A. Gurjanov, A. Kreuchwig, T. Steger-Hartmann, L. A. I. Vaas

CRediT author statement: *Conceptualization*: A. Gurjanov, A. Kreuchwig, T. Steger-Hartmann, L. A. I. Vaas; *Methodology*: A. Gurjanov, T. Steger-Hartmann, L. A. I. Vaas, Software: A. Gurjanov; *Validation*: A. Gurjanov, T. Steger-Hartmann, L. A. I. Vaas; *Formal analysis*: A. Gurjanov; *Investigation*: A. Gurjanov; *Resources*: A. Kreuchwig; *Data curation*: A. Gurjanov, A. Kreuchwig; *Writing – original draft*: A. Gurjanov; *Writing – review and editing*: A. Gurjanov, A. Kreuchwig, T. Steger-Hartmann, L. A. I. Vaas; *Visualization*: A. Gurjanov; *Supervision*: T. Steger-Hartmann, L. A. I. Vaas; *Project administration*: T. Steger-Hartmann; *Funding acquisition*: T. Steger-Hartmann

Citation: Gurjanov A, Kreuchwig A, Steger-Hartmann T and Vaas LAI (2023), Hurdles and signposts on the road to virtual control groups—A case study illustrating the influence of anesthesia protocols on electrolyte levels in rats. Front. Pharmacol. 14:1142534. doi:10.3389/fphar.2023.1142534.

frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 20 April 2023
DOI 10.3389/fphar.2023.1142534

OPEN ACCESS

EDITED BY
Dirk Steinritz,
Ludwig Maximilian University of Munich, Germany

REVIEWED BY
Niko Amend,
Bundeswehr Institute of Pharmacology and Toxicology, Germany
Paul Moser,
Independent Researcher, Albi, France

*CORRESPONDENCE
A. Gurjanov,
alexander.gurjanov@bayer.com

SPECIALTY SECTION
This article was submitted to Predictive Toxicology, a section of the journal Frontiers in Pharmacology

RECEIVED 11 January 2023
ACCEPTED 31 March 2023
PUBLISHED 20 April 2023

CITATION
Gurjanov A, Kreuchwig A, Steger-Hartmann T and Vaas LAI (2023), Hurdles and signposts on the road to virtual control groups—A case study illustrating the influence of anesthesia protocols on electrolyte levels in rats. *Front. Pharmacol.* 14:1142534. doi: 10.3389/fphar.2023.1142534

Hurdles and signposts on the road to virtual control groups—A case study illustrating the influence of anesthesia protocols on electrolyte levels in rats

A. Gurjanov[1]*, A. Kreuchwig[1], T. Steger-Hartmann[1] and L. A. I. Vaas[2]

[1]Bayer AG, Pharmaceuticals, Investigational Toxicology, Berlin, Germany, [2]Bayer AG, Pharmaceuticals, Research and Pre-Clinical Statistics Group, Berlin, Germany

Introduction: Virtual Control Groups (VCGs) represent the concept of using historical control data from legacy animal studies to replace concurrent control group (CCG) animals. Based on the data curation and sharing activities of the Innovative Medicine Initiatives project eTRANSAFE (enhancing TRANSlational SAFEty Assessment through Integrative Knowledge Management) the ViCoG working group was established with the objectives of i) collecting suitable historical control data sets from preclinical toxicity studies, ii) evaluating statistical methodologies for building adequate and regulatory acceptable VCGs from historical control data, and iii) sharing those control-group data across multiple pharmaceutical companies. During the qualification process of VCGs a particular focus was put on the identification of hidden confounders in the data sets, which might impair the adequate matching of VCGs with the CCG.

Methods: During our analyses we identified such a hidden confounder, namely, the choice of the anesthetic procedure used in animal experiments before blood withdrawal. Anesthesia using CO_2 may elevate the levels of some electrolytes such as calcium in blood, while the use of isoflurane is known to lower these values. Identification of such hidden confounders is particularly important if the underlying experimental information (e.g., on the anesthetic procedure) is not routinely recorded in the standard raw data files, such as SEND (Standard for Exchange of Non-clinical Data). We therefore analyzed how the replacement of CCGs with VCGs would affect the reproducibility of treatment-related findings regarding electrolyte values (potassium, calcium, sodium, and phosphate). The analyses were performed using a legacy rat systemic toxicity study consisting of a control and three treatment groups conducted according to pertinent OECD guidelines. In the report of this study treatment-related hypercalcemia was reported. The rats in this study were anesthetized with isoflurane.

Results: Replacing the CCGs with VCGs derived from studies comprising both anesthetics resulted in a shift of control electrolyte parameters. Instead of the originally reported hypercalcemia the use of VCG led to fallacious conclusions of no observed effect or hypocalcemia.

Discussion: Our study highlights the importance of a rigorous statistical analysis including the detection and elimination of hidden confounders prior to the implementation of the VCG concept.

KEYWORDS

historical control data, virtual control groups, systemic toxicity study, replacement, clinical chemistry, 3R

1 Introduction

In vivo toxicity studies continue to play a central role in regulatory toxicology. The design of these studies is well harmonized: first, groups of animals are exposed to a test substance in different doses. This is followed by measurements, analyses and microscopic assessments of selected parameters (e.g., body weight, organ weights, electrolytes, protein levels) in the blood, serum, urine, tissues and organs (OECD, 2008; EMA, 2010; EMA, 2013; OECD, 2018a). To subsequently determine whether observed effects are treatment-related, the endpoints of the test substance-treated animals (i.e., treatment groups or dose groups) are compared to those of the control group. Statistical tests (such as Dunnett's test (Dunnett, 1955)) are performed to determine the statistical significance of a deviation from the control data (Hamada, 2018). All measured data are then stored in the archives of the test facility, allowing the reuse of control group animal data for a historical control data (HCD) collection (Kluxen et al., 2021). So far, HCDs are mainly used as a reference base in toxicity studies (OECD, 2018b). Expert toxicologists use HCD to determine whether the measured endpoints in ongoing studies are within the range of HCDs. This helps in the assessment of biological relevance, i.e., if a certain measured value of dose group animals is outside the limits of the HCD values, this may indicate a biologically relevant effect (Kluxen et al., 2021). In most test facilities the HCD selection is only based on a fixed retrospective time interval but without any further statistical quality control.

The bioassays of regulatory toxicology are highly standardized, compliant to the guidelines of good laboratory practice (GLP) and conducted according to regulations of the US Food and Drug Administration (FDA), the European Medicines Agency (EMA), the Organization for Economic Cooperation and Development (OECD) and the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH). Given the highly controlled environment of the studies, it can be expected that the variations of physiological parameters between HCD and a concurrent control group (CCG) are limited and detectable in statistical analysis. Based on these considerations, the idea of the Virtual Control Groups (VCGs) was introduced some years ago (Steger-Hartmann et al., 2020). This concept aims to reduce the number of concurrent control group animals by using VCGs—which are generated from HCD—and thus to contribute to the 3R concept of (Russel and Burch, 1959).

The ViCoG (Virtual Control Group) was established as part of the Innovative Medicine Initiatives project eTRANSAFE (Pognan et al., 2021) with the aim of collecting, analyzing, and sharing HCD across multiple pharmaceutical companies and to start a qualification process for the VCGs. The collected HCD consist of animal data from regulatory toxicity studies conducted according to internationally highly standardized research practices (Bode, 2020) and their collection is often mandatory or strongly recommended (OECD, 2008; OECD, 2018a; OECD, 2018b; ICH, 2020). To use HCD for creating VCGs in the future, a common database meeting the following essential requirements has been set up: i) the collected data should have a harmonized format and ii) metadata must be recorded, such as information on the study design, animal suppliers, food type, and analytical methods.

Similar terminology and diagnostic criteria should be used when pooling data (Greim et al., 2003). Historically, a lack of harmonization has been a major obstacle to almost all data-collection efforts in Life Sciences (Kolker et al., 2012). But with the introduction of SEND (Standard for Exchange of Non-clinical Data) in 2002 (Wood and Lou, 2011) the harmonization of data from systemic toxicity studies has been greatly facilitated. SEND provides a set of harmonized terminologies and a framework for storing data of studies with various designs. Since 2016, SEND has been the mandatory format for submitting toxicity data to the FDA (CDISC, 2022), making it a well-suited framework for a database with a uniform architecture. With these SEND guidelines at hand, the members of the ViCoG can provide a large amount of historical control data in a harmonious format, which has been collected over the last decade. Beyond data harmonization, it is critical to obtain a thorough understanding of the characteristics and variability of the collected data itself before starting to replace CCGs with VCGs. Even minor changes in individual parameters, resulting from genetic drift or changes in methodological or analytical procedures, may induce significant differences between VCGs and concurrent controls impacting the outcome of statistical analyses and consequently have effects on the identification of treatment-related findings. In case of unknown or undocumented differences of design parameters influencing thus both the dependent and the independent variable, a classical confounder is present. To avoid misleading impact of those lurking variables, adequately matching of HCD for generation of VCGs is key. The importance has recently been shown for clinical pathology parameters (Wright et al., 2023) where the error rate in the recognition of treatment-related findings increased while loosening the selection criteria for HCD. In certain study types, such as carcinogenicity bioassays or the rat bone marrow micronucleus assays, requirements for the selection process of HCD are defined (Greim et al., 2003; Keenan et al., 2009; Igl et al., 2019).

OECD and ICH regulatory guidelines are less stringent when it comes to using HCD for comparison purposes but emphasize that data must be "collected from the same laboratory, species, strain and under similar conditions" (OECD, 2008; OECD, 2018b; OECD, 2018a), ideally also from "recent time" (ICH, 2020).

To summarize the requirements from literature and regulatory authorities, for *in vivo* toxicity studies discussed here, the following parameters are considered to be essential and therefore need to be controlled when comparing concurrent controls and HCD:

- sex,
- strain,
- supplier,
- age,
- housing conditions,
- route of administration,
- diet,
- tissue collection and processing procedures,
- treatment vehicle

and due to potential changes in analytical methods, one should also consider staying within a timeframe between 2 and 7 years.

These requirements provide a good starting point for the VCG approach in toxicity studies but given the huge number of measurements plus the complexity of employed bioassays, potential confounders remain highly likely. Constant and careful exploration shall be part of any activity creating robust and reliable VCGs.

Interestingly, the concept of using external controls instead of concurrent controls exists for clinical trials since years (Pocock, 1976) and has led to various applications to reduce the time- and resource-intensive recruitment of patients without risking the loss of statistical power or declining quality of the results. External controls are of particular interest for rare (orphan) diseases or for those where recruiting control subjects would be cumbersome or unethical (Pocock, 1976; Strayhorn, 2021). In the clinical setting, various methods have been introduced to derive external control groups from historical data that match subjects in a treatment arm of a clinical trial (Lim et al., 2018). Propensity score methods (Rosenbaum and Rubin, 1985), Bayesian methods (Lim et al., 2018; Zhan et al., 2022), or a mixture of both (Sawamoto et al., 2022) are used to construct external controls. Clinical trials with external controls generated by propensity scores have even successfully resulted in a drug approval (Gökbuget et al., 2016).

However, the methods for creating the VCGs from clinical studies cannot be directly transferred to non-clinical conditions. Clinical studies and preclinical animal experiments differ too much in terms of design and the homogeneity of their subjects. Control groups in clinical trials can show significantly higher variance in key characteristics such as age, body mass index, comorbidities, and especially genetic variance. In comparison, preclinical *in vivo* toxicity studies are usually designed as randomized case-control studies (e.g., control group and low, medium, and high dose group) and use treatment-naïve animals of similar, well-defined age, weight and low genetic variation. This is achieved by breeding under highly standardized conditions, strict inclusion criteria, and sourcing from the same supplier (White and Cham, 1998).

In the non-clinical settings, several methods for generating virtual control groups were proposed and include Bayesian approaches to either replace concurrent controls with inclusion of historical data (Kramer and Font, 2017; Wright et al., 2023)—which is the goal of the VCG approach—or to increase the statistical power of studies (Bonapersona et al., 2021). In regulatory toxicology the studies are conducted according to standardized guidelines (FDA, 2000; OECD, 2008; EMA, 2010), which provide general recommendations for group sizes for each study type. In short-term repeated-dose toxicity studies in rodents, a sample size of 10 animals per sex per group is normally recommended (FDA, 2000). Although statistical power is an important component for the design of meaningful assays (Charan and Kantharia, 2013), studies for preclinical safety assessment are carried out with those standardized designs and pre-specified group sizes. Considerations of individual statistical power or sample-size estimations play a minor role in preclinical safety assessment. This article therefore focuses on creation of virtual control groups with the goal of reducing the size of concurrent control group animals while maintaining the given design of the studies.

Aside from Bayesian approaches, simulation-based approaches that artificially generate control group values from aggregated historical data have also been proposed (Hothorn et al., 2019; Steger-Hartmann et al., 2020). However, the resulting VCG values would be purely synthetic and potential, but so far unknown, correlations between various endpoints might be difficult to reproduce. Therefore, a simple and straightforward resampling method (Steger-Hartmann et al., 2020), in which VCG data are randomly drawn directly from HCD build a meaningful first step here. The method is directly applicable and does not require complex mathematical and statistical background knowledge. In addition, the data shall come from historical studies conducted under tightly regulated GLP conditions and each value is directly traceable to the individual animal. Further, the outcomes of the resampling method are easy to interpret and allow for an in-depth analysis of the underlying data.

Before sampling the control group data, the HCD itself should be pre-filtered according to the key factors listed above to ensure concordance between concurrent controls and VCGs. It remains to determine how to validate the performance of the VCGs concept in general. Since regulatory toxicology works with strictly standardized studies, it is considered appropriate that VCG performance can initially be quantified by how well VCGs can reproduce the results of these studies. Thus, the VCGs need to reproduce results of a given historical study with respect to identifying treatment-related findings. In regulatory toxicology, significance tests are commonly performed to assess whether an observed difference is statistically significant or not (Hamada, 2018). When significance is detected, expert toxicologists determine whether the finding indicates a treatment-related effect by consulting HCD and observing interactions between various endpoints measured in the bioassay. HCD can be also used to calculate an effect size supporting the toxicologists in making this decision (Schmidt et al., 2016; Kluxen et al., 2021). However, because effect sizes were not calculated in the study reports reviewed in this article, we do not consider them here either and rather focus on VCGs' ability to reproduce originally reported statistical significances of the studies.

In this article, a historical study of systemic toxicity in rats (legacy study) serves as a test case to evaluate and illustrate the performance of virtual control groups generated by a straightforward resampling method. A treatment-related increase in serum calcium was reported and VCGs should be able to reproduce this finding. Calcium belongs to the group of electrolytes where changes found in animal studies after administration of a drug candidate can help identify toxicities potentially leading to adverse events during clinical trials. Calcium is, apart from its role in bone formation, essential for the proper functioning of muscles, nerves, and the heart. Calcium imbalances can lead to severe effects such as bone pain, hypertension, seizures, tetany, paresthesia, laryngospasm, and cardiac conduction abnormalities (Schenck et al., 2006; Tinawi, 2021). The range considered normal is relatively small due to the strict physiological regulation of electrolytes. Therefore, small decreases or increases in serum concentrations are likely to result in significant differences between control and treatment groups.

We present VCGs generated by a resampling approach and validate their performance on the parameter calcium. The statistical significance of the selected legacy study serves as a reference for the performance and the aim is to reproduce the statistical results of this study, after replacing the CCGs with VCGs. During data quality assessment of the HCD, it was found that calcium, potassium, sodium, and phosphate values of control animals showed time-dependent changes that proved to be of critical importance in terms of proper data selection for VCGs. The presence of a confounder in

the electrolyte values distorts their variability. This article demonstrates the consequences of an unreflective application of VCGs. Creating VCGs based on insufficiently prefiltered HCD results in a poor ability to reproduce statistical results which might in turn lead to erroneous toxicological decisions. We then describe strategies to counter the impact of such hidden confounders—and thus improving the performance of VCGs—by using adequate statistical control mechanisms. We developed a procedure for selecting data to be used for the generation of VCGs in toxicity tests and recommendations for an in-depth analysis that reveals previously unknown or unrecognized confounding variables.

Such a procedure would follow a stepwise approach: first, HCD needs to be selected to match common parameters such as study year, sex, strain, route of administration, treatment vehicle, supplier, age, and initial body weight. Afterwards, the quality of the resulting data must be assessed. This article shows that visualizing study data over time may reveal various phenomena in the data which offers a good starting point for identifying abnormalities, such as atypical shifts in values or unexplained increases or decreases in these values over time. Assuming that an atypical shift is detected in the data, an in-depth analysis is recommended to identify the root cause and underlying confounder in the data. In the event that the confounder cannot be identified, access to the original study reports and expert knowledge are critical for interpreting and tracing the confounder variables.

2 Methods

2.1 Data selection

All data were collected from previous animal studies performed by Bayer, Wuppertal, Germany. For these historical studies, the animals were kept and treated in accordance with the German Animal Welfare Act and approved by the competent state authorities. All data from animal studies were recorded in SEND format (Standard for Exchange of Non-Clinical Data) (CDISC, 2022). The data processing, statistical evaluations and visualizations were carried out with the software R, version 4.1.0. The R code used along with the data can be extracted from GitHub https://github.com/bayer-group/VCG-resampling.git. A detailed description of the origin of the data and the software used for data analysis can be found in the Supplementary Material S1, chapter 1.1 and 1.2. For the construction of the VCG database, control data sets for the different toxicological endpoints were extracted, including metadata describing the design of the study. The HCD were filtered with the aim of obtaining the largest possible amount of animal data while minimizing the potential impact of genetic variability. 28-day repeated dose toxicity studies are the dominant type of *in vivo* studies in regulatory toxicology (Baldrick, 2008; EMA, 2013). Therefore, data were selected from both 28-day studies and studies longer than 28 days in duration (in this case, only measured endpoints measured between study days 1 and 35 were extracted). Additional filter steps for studies for the VCG collection were selected based on the following criteria:

- Study initiation between 2011 and 2021.
- Usage of Wistar HAN rats.
- Age of rats between 6 and 9 weeks (at the beginning of the respective study).
- Animals obtained from the supplier Charles River, Germany, or Harlan, Netherlands.
- The route of administration was "oral gavage".
- A mixture of Ethanol, Kolliphor®HS15, and water served as treatment vehicle.
- The initial body weight was between 100 g and 250 g.
- All endpoints were measured in Bayer's laboratory in Wuppertal, Germany.
- Only male rats were used.

From a total of 114 rat studies of the Bayer VCG data set, the data selection process reduced the set to 30 studies for calcium, 31 for potassium and sodium, and 26 studies for inorganic phosphate. The data selection process is summarized in Figure 1.

2.2 VCG performance assessment

To understand how replacing concurrent controls with VCGs influences the outcome of a study with respect to the so-called "treatment-relatedness" (Wright et al., 2023), a study with a reported treatment-related change in calcium in male rats was selected. This study is denoted as "legacy study" in this article. In the performance assessment, the aim was to test whether the statistical results of the legacy study were reproducible after replacing the concurrent control group (CCG) with VCGs while preserving the study design parameters of the legacy study. In other words, animals from the concurrent control group (CCG) of this legacy study were replaced by the same number of VCGs constructed by different selection criteria. Afterwards, the effect on the outcome was analyzed focusing on whether changes between control groups and dose groups were statistically significant (= evidence for a treatment-related effect) or not (= no evidence for a treatment-related effect). In addition to the calcium value, two other parameters were examined to gain an understanding of whether significant findings for correlated parameters can also be reproduced with the VCGs. Apart from the statistically significantly increased values for calcium, the legacy study also showed a significantly increased value in the highest dose group for the parameter inorganic phosphate in the blood serum. This parameter is strongly correlated with the calcium parameter, and it is of interest to test whether the VCGs can also reproduce this significance. Additionally, body weight on day 28 was taken into account. This parameter should not show strong correlations to the electrolyte levels in the rat and should not be affected by the used anesthetic.

The resampling was performed in the following procedure which is illustrated in Figure 2:

1) The individual values for calcium in blood serum of the rats from the legacy study were retrieved from the database.
2) A Dunnett's test was performed identical to the procedure used to analyze the original data with concurrent controls. The Dunnett's test output was classified into two categories: if the resulting p-value of the Dunnett's test was smaller than or equal to 0.05, the result was classified as "significant", otherwise as "not significant".
3) A sample population was created with respect to predefined filtering criteria.

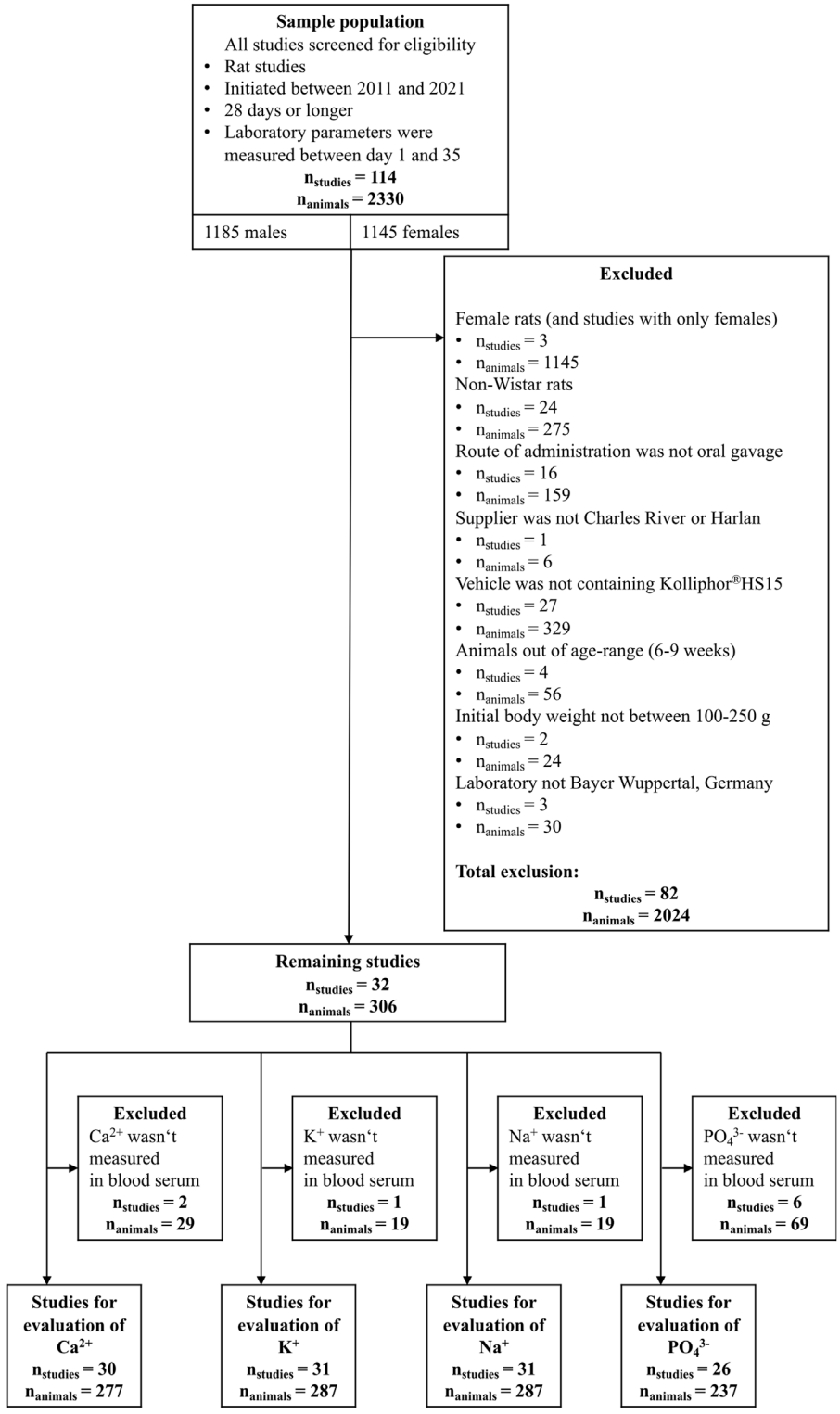

FIGURE 1
Data selection flow diagram.

4) *n* values were picked by random sampling without replacement from the sample population set where *n* is the number of removed CCG animals. The removed CCG-animal values were then replaced by these drawn VCG values.
5) The Dunnett's test was calculated, now using the VCG instead of the CCG as the reference group.
6) The result of the Dunnett's test was compared with the original outcome of the legacy study. If the VCG result was consistent

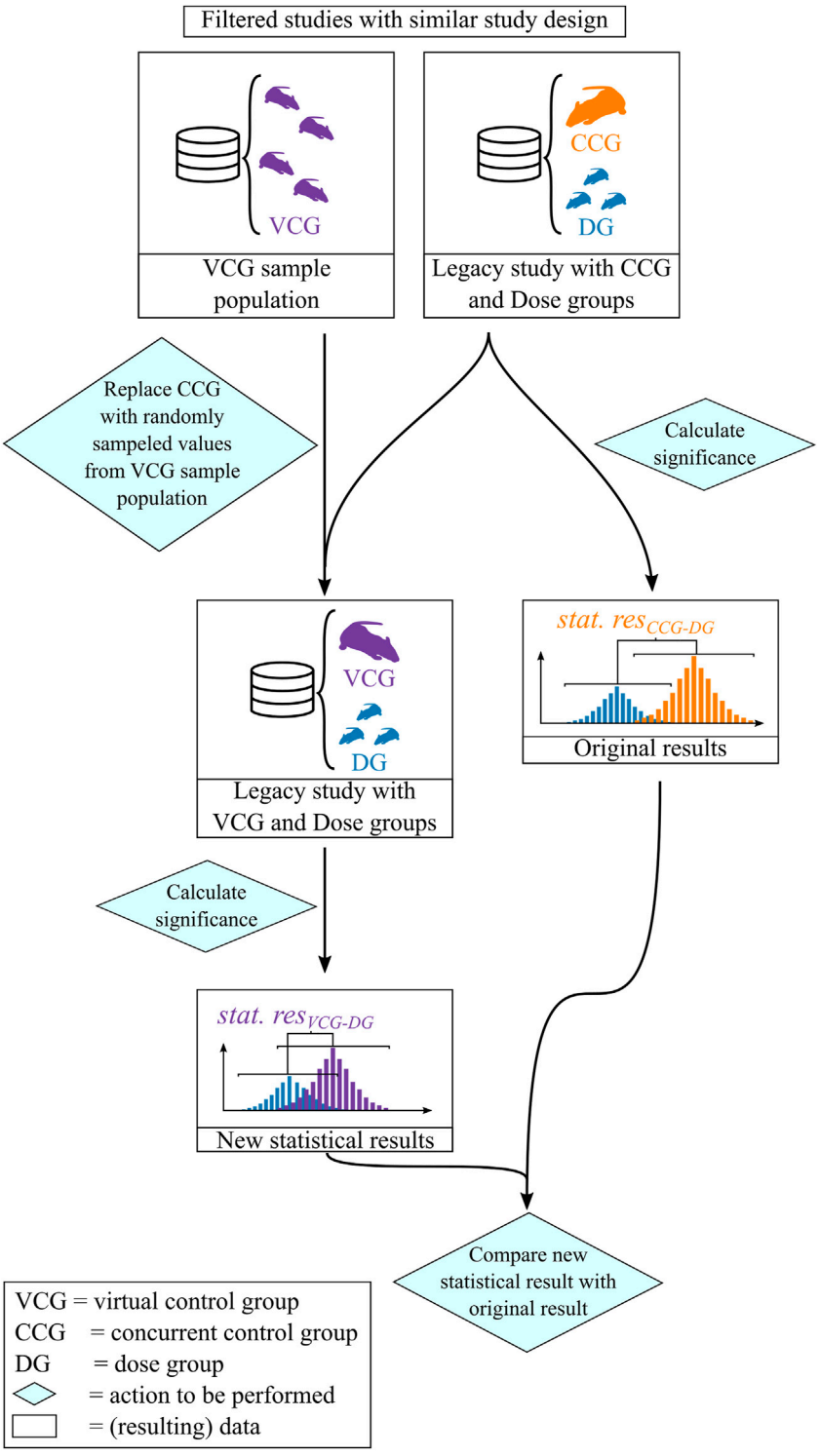

FIGURE 2
Workflow of assessing the performance of VCGs.

with the CCG one, this sample was classified as "consistent", otherwise as "inconsistent". Inconsistent results were further classified into following categories:

- When the statistical outcome of the VCG leads to a significant result while the original distance was not significant, the result was classified as "inconsistently significant".

- When the statistical outcome of the VCG leads to a non-significant result while the original distance was significant, the result was classified as "inconsistently non-significant".
- When the statistical outcome of the VCG leads to a significant result and the original result does so too, but the direction of the distance is inverted (e.g., a significant decrease is observed while there was a significant increase in the original data), the result was classified as "inverted significant".

7) The resampling was repeated 500 times and the percentage of consistent results per dose group was summarized. This percentage is termed here as the "reproducibility [%]" and is used as the parameter to validate the performance of the VCGs.

To examine the correlations of the VCGs with other parameters, the VCGs animals were selected based on their unique animal subject ID. This means that at each iteration, a certain number of animals were randomly sampled and the endpoints for all parameters of interest (i.e., calcium, phosphate, and body weight) were extracted from each selected animal.

2.3 Performance-improvement methods

2.3.1 Search for confounders

The selected dataset was statistically analyzed for each parameter to gain an understanding of variance and distribution of the data over time. Electrolyte parameters were plotted as histograms and boxplots relative to the year the studies began. To gain further insight into how the underlying parameters (i.e., confounders) might affect electrolyte levels, the plots have been colored in relation to these confounders. Because procedural data for the identified confounder "anesthetic" was not captured in the SEND dataset, information was manually extracted from the selected study reports. In addition to the plots, a Welch-adapted *t*-test was performed to statistically describe the effects on the two groups separated by the confounder (the normality of the distribution with unequal variances was confirmed visually by the respective histograms). The VCG data was then filtered further: the VCG sample population was divided into animals anesthetized with isoflurane and animals anesthetized with CO_2. After controlling for the confounder (i.e., using animals from only one of the anesthetic subgroups), the performance of the VCGs was re-evaluated by the same procedure as mentioned above.

2.3.2 Keeping sentinel animals

Instead of replacing all CCG animals, the performance of the VCGs was additionally evaluated after only a fraction of the CCG animals were replaced. The CCG animals to remain (i.e., the sentinel animals) were selected using the initial body weight values of the CCGs, i.e., the body weight of the animals before the start of the study. The animals were selected with the aim of obtaining the original mean values and standard deviation values of the body weight distribution. Of n sentinel animals, the $n/2$ heaviest and the $n/2$ lightest animals remained in the control group, while the remaining animals were replaced with VCG data. In addition, when n was an odd number, an animal was selected from the middle of the initial body weight distribution, e.g., if it has been decided to include five out of ten animals, animal 1, 2, 5, 9 and 10 (sorted by body weight) were selected. Three different sub-scenarios were performed: i) all CCG animals were replaced by VCGs, ii) two animals were retained as sentinel animals, iii) half of the concurrent control animals remained in the group. To use all available information when recruiting VCGs, the measured calcium values of the sentinel animals were used as an additional filter narrowing down the VCG sample population: VCG data were filtered within the mean ± 2·SD range of sentinel animals' calcium values. The performance of the resulting VCGs was then evaluated using the same design as described above and shown in Figure 2.

3 Results

This section is divided into the following parts: first, the statistical results of the original study with the concurrent control group are shown in Table 1 along with the distribution of the VCG sample population in Figure 3. The methods and the resulting performance of the VCGs are further separated in six scenarios: first, the performance of the VCGs (i.e., the reproducibility [%]) of the "agnostic scenario" is shown. Afterwards, the results for the approaches to improve the performance of the VCGs by two methods are presented: the "search for confounders" method where data was removed from the VCG sample population which was affected by a confounder; the "keeping sentinel animals" method where instead of completely replacing the CCG animal data either 80% or 50% of the animal data was replaced respectively; and finally, a combination of both methods where data affected by the confounder was removed and only 80% or 50% of all CCG animals were replaced. The results of all scenarios are summarized in Table 2. The ranges of the VCG data are illustrated in the Supplemetary Material S2, Supplemetary Figure S5.

3.1 Original statistical results of the legacy study

The legacy study consists of a concurrent control group (CCG) and three dose groups, denoted as Dose group 1, Dose group 2, and Dose group 3. All groups consisted of 10 male rats. The mean and standard deviation values as well as the population of the serum calcium values are shown in Table 1 and are illustrated in Figure 3. Statistical analysis employing Dunnett's tests leads to no significant difference between CCG and Dose group 1 but revealed significant differences between the CCG and Dose groups 2 and 3 respectively. The corresponding historical control data for serum calcium values are shown as a histogram in Figure 3A: a bimodal distribution is present with two peaks at 2.87 mmol/L and 2.55 mmol/L. This renders the mean value of the VCG sample population to be at 2.68 ± 0.19 mmol/L (Figure 3B), i.e., considerably higher in mean value and standard deviation compared to the concurrent control with 2.57 ± 0.06 mmol/L (orange cross in Figure 3C).

3.2 VCG performance: The agnostic scenario

In the first scenario, all control animals (n =10) were replaced and the VCG sample population was not further filtered ignoring both the bimodal distribution and the fact, that assays lacking controls would be invalid (Figure 4A). The consistencies for this

TABLE 1 Selected legacy study with the mean calcium levels in blood serum in each sex, the standard deviation, and the population.

Dose group	Mean Ca^{2+} values in serum [mmol/L]	Population
Concurrent control	2.57 ± 0.06	10
Dose group 1	2.57 ± 0.05	10
Dose group 2	**2.64 ± 0.04***	10
Dose group 3	**2.69 ± 0.04***	10

*Significant difference with Dunnett *p*-value of <0.05.
Statistically significant differences between the respective dose groups and the control are highlighted in bold and marked with an asterisk.

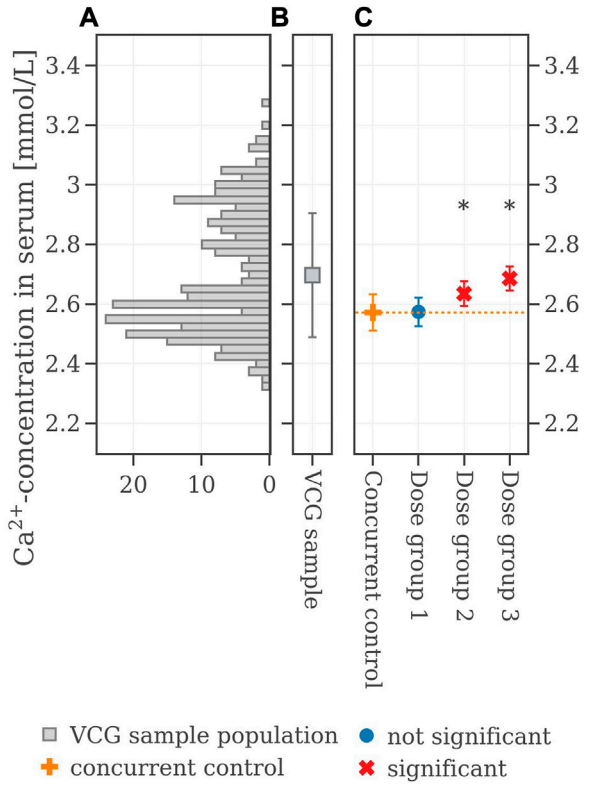

FIGURE 3
(A) Histogram of the calcium values of the VCG sample population. **(B)** Boxplot of the VCG sample population. **(C)** Extracted original results of serum calcium concentration from the legacy study. If the mean difference between a dose group and the concurrent control exceeded the Dunnett-critical distance, the group was marked with an asterisk.

approach were generally poor, with 52% for Dose group 1, 2% for Dose group 2, and 5% for Dose group 3 (see Table 2 for more details). One example iteration is shown in Figure 4B below: the mean value of the VCG (2.79 mmol/L) is considerably higher than the one of the CCG (2.57 mmol/L). Furthermore, while the mean values increased in comparison to the CCG with rising substance-level, here, the difference of mean values between the VCG and the dose groups is diminishing with increasing dose. Also, the standard deviation of the VCG (0.22 mmol/L) is considerably larger for the CCG (0.06 mmol/L) and the dose groups (0.05 mmol/L for Dose group 1, and 0.04 mmol/L for Dose groups 2 and 3 each). The mean values of the sampled VCGs from all iterations with respect to the reproducibility of the original results can be found in the Supplementary Material S2, Supplementary Figure S6.

3.3 Improvement of the performance: Search for confounders

In order to understand the reason for the poor performance of the VCGs in the agnostic scenario, the very first step was to gain a deeper understanding of the underlying data and the factor(s) causing a bimodal distribution in the electrolytes. For a general overview of the control data pool, the electrolyte values were illustrated as histograms and as box plots with respect to the year when the study was initiated. Figure 5 shows the values for the electrolyte calcium. The electrolyte values for sodium, potassium, and phosphate can be found in the Supplementary Material S3, Supplementary Figures S12–S15. As mentioned before, a bimodal distribution in these electrolyte values was observed. The time-controlled graph revealed a sharp drop from 2016 to 2017. Before 2017, the mean values of calcium in serum were at (2.87 ± 0.14) mmol/L (95% CI [2.85, 2.90]) and dropped afterwards to (2.55 ± 0.07) mmol/L (95% CI [2.53, 2.56]).

An in-depth analysis revealed that before 2017 animals were anesthetized with a different procedure compared to animals in studies after 2017. As this information is not part of the SEND-data, the anesthetic procedure had to be extracted manually for each study. Implementing the information of the anesthetics revealed two normally distributed subsets shown in Figure 5: the distribution of the isoflurane subset (violet) has lower calcium levels and shows a smaller standard deviation (2.55 ± 0.07) mmol/L, while the distribution of the CO_2 subset (grey) has higher calcium values and shows a higher/larger standard deviation (2.87 ± 0.14) mmol/L.

Finally, a Welch-adapted *t*-test was performed to assess the difference between these two groups. For calcium, the 95% CI for the difference in means resulted in [0.31, 0.37] with a *p*-value of <2.2 e−16.

In the chosen legacy study, which was performed in 2018, the animals were anesthetized with isoflurane and subsequentially, the VCG sample population was selected accordingly by excluding data from animals anesthetized with CO_2. The new range from which VCG animals were derived is shown in Figure 6A. Using this filter, the performance of the VCGs improved. Now, the consistencies were at 74% for Dose group 1, 97% for Dose group 2, and 100% for Dose group 3 (Table 2 for more details). Figure 6B shows the results from one of the 500 iterations. The CCG mean is (2.57 ± 0.06) mmol/L while the mean of the picked VCG is very close at (2.54 ± 0.08) mmol/L. The mean values of all iterations and with respect to the reproducibility of the original results can be found in the Supplementary Material S2, Supplementary Figure S9.

TABLE 2 Resampling results of the legacy study on the parameter calcium after replacing the concurrent control group with virtual control groups (VCG) sampled from the respective subgroups. The sampling was performed 500 times and the percentage of consistent statistical results are given for each sex and each dose group (DG).

Mean value of the CCG [mmol/L]	Scenario	Mean value of the VCG sample population [mmol/L]	Sub-scenario	DG 1 consistency	DG 2 consistency	DG 3 consistency
2.57 ± 0.06	1: Confounder is unknown	2.69 ± 0.20	1a: Replace all CCG animals	**52% consistently non-significant**	**2% consistently significant**	**5% consistently significant**
					83% inconsistently non-significant	92% inconsistently non-significant
				48% inconsistently significant	15% inverted significant	3% inverted significant
			1b: Keep 2 sentinel animals	**90% consistently non-significant**	**100% consistently non-significant**	**100% consistently non-significant**
				10% inconsistently significant		
			1c: Replace half of the CCG animals	**100% consistently non-significant**	**100% consistently significant**	**100% consistently significant**
	2: Confounder is known	2 54 ± 0.08	2a: Replace all CCG animals	**74% consistently non-significant**	**97% consistently significant**	**100% consistently significant**
				26% inconsistently significant	3% inconsistently non-significant	
			2b: Keep 2 sentinel animals	**87% consistently non-significant**	**100% consistently significant**	**100% consistently significant**
				13% inconsistently significant		
			2c: Replace half of the CCG animals	**100% consistently non-significant**	**100% consistently significant**	**100% consistently significant**

The bold text represents the reproducibility percentage, i.e., the ability of the VCGs to reproduce the original statistical results of the legacy study and is thus a measure of performance of the VCGs.

3.4 Improvement of the performance: Keeping sentinel animals

As a second option to improve the performance of the VCGs several concurrent animals of the legacy study were kept as so-called sentinel animals. After filtering for heaviest and lightest animals, their calcium values' mean ± 2·SD range was used to narrow down the VCG sample population. Two scenarios were examined here: keeping the values from two CCG animals and keeping the values of half of the CCG population (shown in Figure 7A). If two animals were kept and VCGs were only derived within the calculated range, the consistency of the VCGs was at 90% for Dose group 1, and 100% for Dose group 2 and 3 respectively. Keeping half of the CCG animals in the set improved the performance to 100% for all dose groups respectively (see Table 2 for more details). Figure 7B shows one example iteration of the performed 500 where half of the animals were kept as sentinel animals. The CCG mean value is at 2.57 ± 0.06 mmol/L while the mean value of the VCG is again very close to the CCG at 2.56 ± 0.06 mmol/L. The mean values of the VCGs from all iterations with respect to the reproducibility of the original results can be found in the Supplementary Material S2, Supplementary Figure S7, S8.

3.5 Improvement of the performance: Control confounder and keep sentinel animals

Combining both methods did not considerably improve the performance. Keeping two sentinel animals and controlling for the confounder had a similar performance as the one where only the confounder was controlled: 87% of all iterations led to reproducible results for Dose group 1, 100%, and 100% for Dose group 2 and 3. Keeping half of the CCG animals while controlling for the confounder led to a high consistency of 100% for all dose groups respectively (Table 2 for more details). The mean values of all sampled VCGs with respect to the reproducibility of the original results for all iterations can be found in the Supplementary Material S2, Supplementary Figures S10, S11.

3.6 The performance of the VCGs on the parameter inorganic phosphate and body weight

The same performance patterns could be observed for phosphate: in the "agnostic scenario", a generally poor

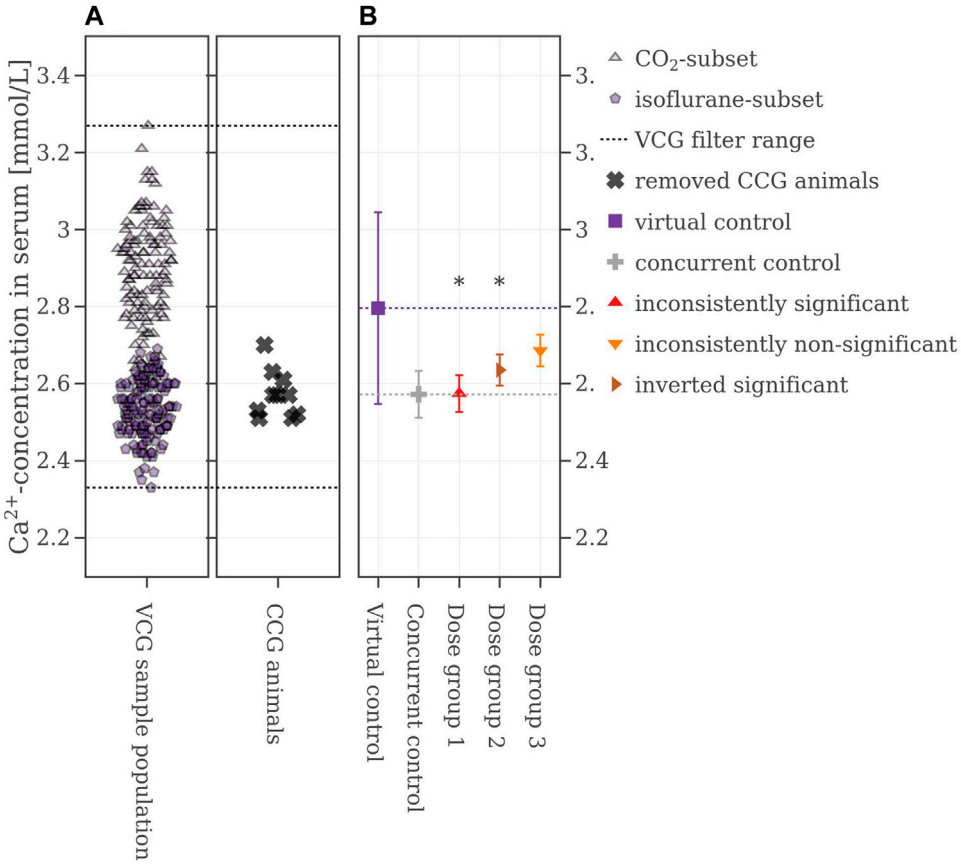

FIGURE 4
(A) Selection range of virtual control values and concurrent control values which are removed. **(B)** Mean values of the legacy study and the virtual control, colored with respect to whether the original statistical results were reproduced or not. If the mean difference between a dose group and the virtual control exceeded the Dunnett-critical distance, the group was marked with an asterisk.

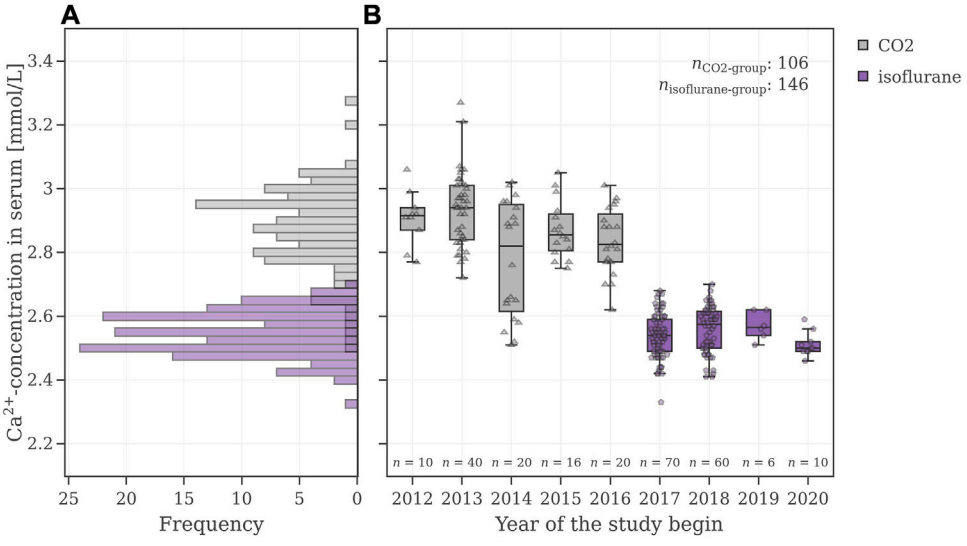

FIGURE 5
(A) Calcium value distributions of male Wistar-rats **(B)** Box plots of these calcium levels with respect to the study year. The CO_2-group is colored grey, and the isoflurane-group is colored violet.

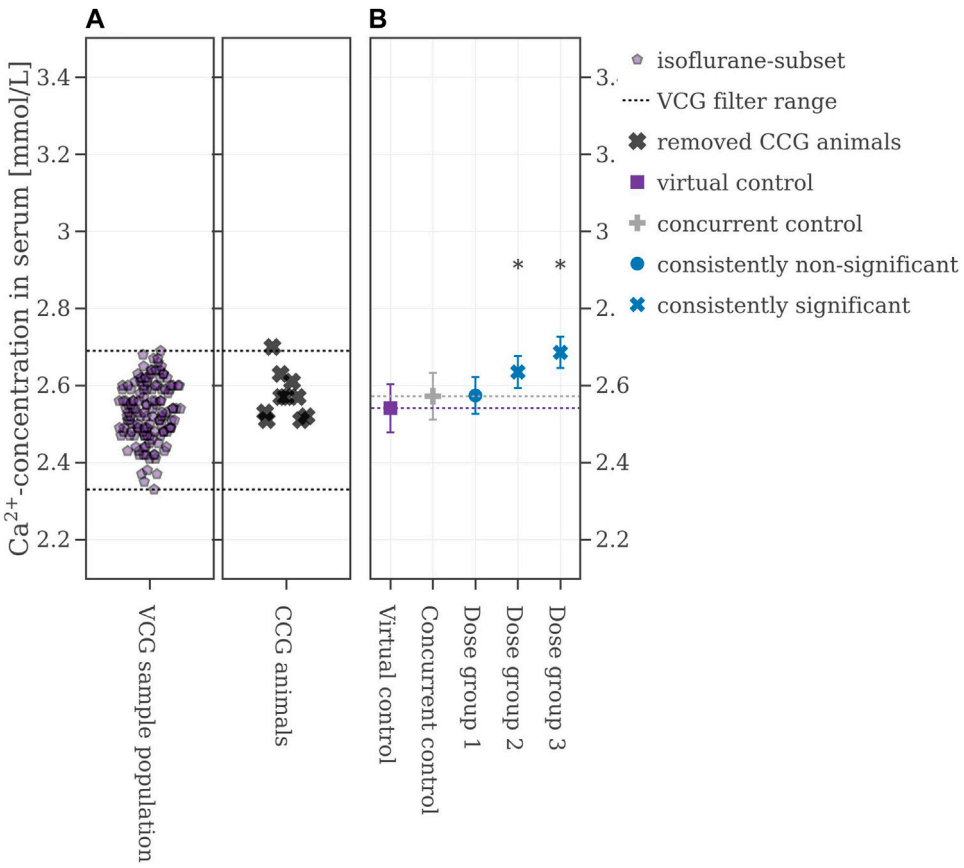

FIGURE 6
(A) Selection range of virtual control values along with concurrent control values which are removed. **(B)** Mean values of the legacy study and the virtual control, colored with respect to whether the original statistical results were reproduced or not. If the mean difference between a dose group and the virtual control exceeded the Dunnett-critical distance, the group was marked with an asterisk.

performance is recorded, with 11% for Dose group 1, 18% for Dose group 2, and 4% for Dose group 3. However, the performance was significantly increased by both improvement methods. Using only isoflurane data improved the performance considerably: the statistical results from all iterations were reproduced in 62% for Dose group 1, 80% for Dose group 2, and 83% for Dose group 3. Sentinel animals selected based on the calcium parameter also increased performance on the phosphate parameter. Keeping two sentinel animals in the control group revealed a reproducibility of 95% in Dose group 1, 99% in Dose group 2, and 66% in Dose group 3. With five sentinel animals retained, a reproducibility of 100% was found for Dose groups 1% and 2, and 82% for Dose group 3 was observed. Further details can be found in the Supplementary Material S4, Supplementary Table S1.

The parameter body weight on day 28 could generally be reproduced well. In the agnostic scenario, the statistical results were reproduced with 94% for Dose group 1% and 100% for Dose group 2 and 3. After removal of CO_2 data from the VCG sample population, the reproducibility even decreased to 76% for Dose group 1, and 95% for Dose group 2. Dose group 3 remained at 100%. The presence of sentinel animals left the original performance unchanged. The statistical results for Dose group 1 could be reproduced in 95% of all iterations when two sentinel animals were kept; Dose groups 2 and 3 in 100% of all cases. Five sentinel animals resulted in a reproducibility of 100% in all dose groups. Further details on the performance of the VCGs towards the body weight parameter can be found in the Supplementary Material S4, Supplementary Table S2.

4 Discussion

In this article, we describe key requirements for statistical characterization of HCD prior to the use of historical data for the implementation of VCGs. Through time-control plots of electrolyte values we identified a sudden drop from 1 year to the other and were able to identify changes in the anesthetic procedure as cause of this drop. We subsequently analyzed how such a hidden confounder might influence the replacement of CCG with VCGs with regard to identification of treatment-related effects using a legacy study, in which treatment-related findings for calcium were reported. This was demonstrated by the performance of the virtual controls, which was assessed by their ability to reproduce the statistical significance of the increased calcium values from the legacy study. The performance of the VCGs is poor when using a insufficiently filtered data set and increased impressively after

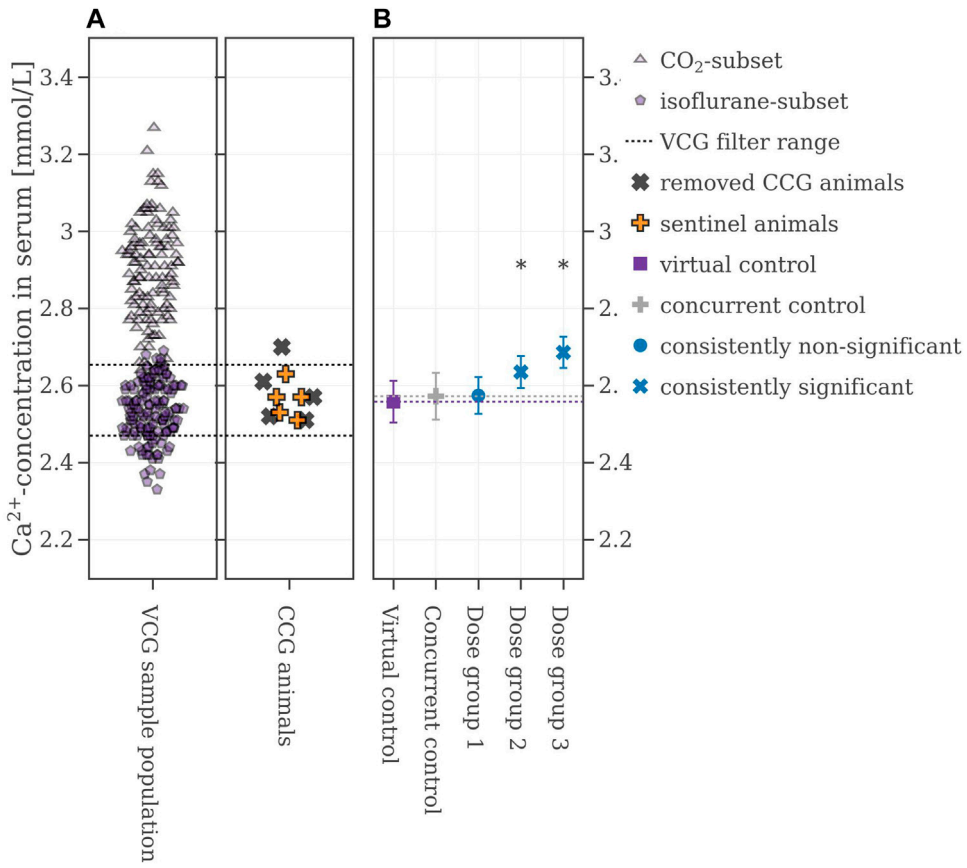

FIGURE 7
(A) Selection range of virtual control values along with concurrent control values which are removed and kept respectively. **(B)** Mean values of the legacy study and the virtual control, colored with respect to whether the original statistical results were reproduced or not. If the mean difference between a dose group and the virtual control exceeded the Dunnett-critical distance, the group was marked with an asterisk.

removing the confounding factor and any data affected by that confounder.

4.1 Assessment of VCG performance

The performance is a percentage resulting from the comparison of *p*-values obtained from the Dunnett's test—once between the CCG and the dose groups of the legacy study and once between the VCGs and the dose groups of the same study. This procedure was repeated 500 times and at each iteration animals were drawn at random from the VCG sample population. Though the statistical significance is not the only decisive factor for toxicologists to speak of a treatment-related effect, it is nevertheless a cornerstone for further decision-making. It is therefore the first sensible step to test whether the statistical significance of the legacy study can be reproduced with the VCGs. The aim was to keep the design of this study unchanged, except for the CCG values which have been replaced by VCGs. The design of the legacy study itself was carried out in compliance with the guidelines for toxicity studies (FDA, 2000; OECD, 2008; EMA, 2010). The FDA guideline states that "for short-term toxicity studies of 30 days duration or less, experimental and control groups should have at least 10 rodents per sex per group" which the legacy study adhered to. To maintain the study design, any removed CCG value was replaced with exactly one value drawn from the VCG sample population, resulting in a constant number of $n = 10$ control animals. Although not considered in this article, increasing the number of control animals by introducing historical data might be an effective way to improve statistical power (Bonapersona et al., 2021). Scenarios with different control group sizes could be examined for potential changes in statistical power, and finally, for discussion of how increasing the power could be beneficial for decision-making in regulatory toxicology. However, a prerequisite would be strict and carefully chosen specifications for the selection of appropriate data together with expert knowledge input from both statisticians and regulatory toxicologists. Recommendation here is a selection of HCD from studies which are as similar as possible in design to the current study. A list of design parameters is proposed (Supplementary Material S5, Supplementary Table S3) and should be expanded in the future with ongoing research on definition of clear inclusion and exclusion criteria. The inclusion of VCGs from animals from different study designs might not necessarily increase statistical power or help toxicologists deciding whether observed statistical significances are treatment related. For example, an article on the curation and analysis of histopathological parameters in a

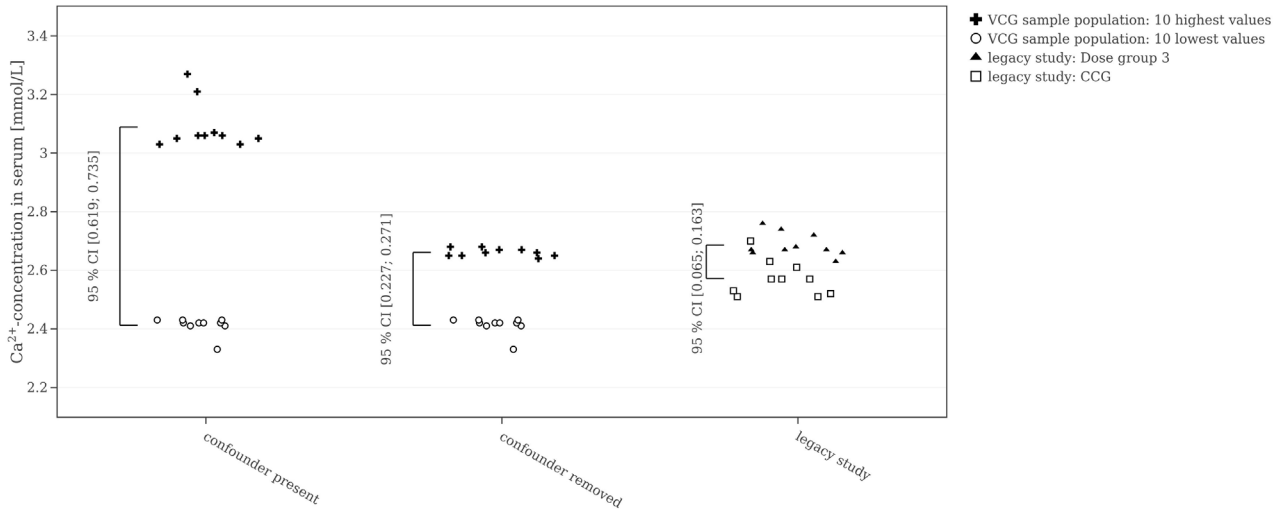

FIGURE 8
10 highest and 10 lowest calcium values of the Virtual Control Group (VCG) sample population (once with and once without the values affected by the confounder) compared to the Concurrent Control Group (CCG) and Dose group 3 of the legacy study. The 95% confidence interval for the difference in means is shown for each group.

large database showed that the proportions of pathological findings may change significantly with increasing sample size (Pinches et al., 2019). Although histopathological data are qualitative parameters, a similar conclusion could be drawn for quantitative parameters: a large HCD population could increase data variability and alter statistical results. There is currently no minimum number of *in vivo* toxicity studies recommended by the OECD to generate meaningful HCDs. However, for *in vitro* studies such as the mammalian micronucleus test, a minimum of ten studies is required (OECD, 2016). It might be worth to assess in future what the minimum number of studies would be for generation of meaningful VCGs. However, relying on HCD alone in a toxicological evaluation might be problematic since, compared to HCD, the concurrent data for bioassay interpretation are generally considered to be the most relevant as they ensure assay validity and help to identify infections in cages (Keenan et al., 2009; Kluxen et al., 2021).

A further suggestion is to check how representative HCD is to simulate concurrent controls. Here, we focused on reproducing the statistical significance of the chosen legacy study and monitor only one endpoint. Regulatory toxicology studies look for both target and off-target effects of a substance screening a comparably large number of endpoints without exerting an *a priori* defined endpoint hierarchy. When it comes to decision-making, an important aspect are relevance limits (Schmidt et al., 2016; Kluxen et al., 2021). To assess the size of potentially detectable effects in a given scenario with HCD, the confidence interval for the largest possible difference in the means of the HCD as the relevance limit is shown in Figure 8. Based on the ten largest and ten smallest calcium values of the HCD sample population, a 95% CI for the difference of means using a two-sample *t*-test is calculated. The legacy study presented here reports evidence of hypercalcemia determined by expert toxicologists. For reproduction of this treatment-related finding, the difference between the mean values of CCG and Dose group 3 would have to be greater than the one between the extreme values from HCD. In the legacy study, the 95% confidence interval for the difference of means between the CCG and Dose group 3 (i.e., the high dose group) obtained by a two-sample *t*-test was [0.065; 0.163] mmol/L. Calculating Cohen's D between these two groups resulted in an effect size of 2.2. For the VCG sample population, the difference in means between the 10 highest calcium levels and the 10 lowest calcium levels resulted in a mean difference of [0.619; 0.735] mmol/L when the confounder was present, i.e., about six times higher compared to the difference between CCG and Dose group 3. Cohen's D-calculation resulted in a large effect size of 11.0, also, five times higher than the effect observed in the legacy study. Removing all values affected by the confounder, selecting the 10 highest and 10 lowest calcium values, and recalculating the *t*-test reduced the difference in means to [0.227; 0.271] mmol/L, still two to three times greater than the difference between CCG and Dose group 3. However, the effect size was still large at 10.8.

Consequently, considering the differences in means or Cohen's D as the limit of biological relevance for estimating a biological effect is not straightforwardly applicable. Communication between toxicology experts and regulatory authorities is still needed to decide on acceptable limits and selection criteria of parameters to determine a standardized effect size (EFSA Scientific Committee, 2011).

4.2 Identifying hidden confounders and retaining sentinel animals dramatically increases VCG performance

Replacing all concurrent control animals with virtual control data from an insufficiently pre-filtered sample population (i.e., the "agnostic scenario") resulted in poor overall performance. In this

scenario, a confounder was present that elevated the mean value and the variance of the VCG sample population. This strongly increased variance led to a non-significant statistical result in most cases and this means in turn that the significant differences in Dose groups 2 and 3 from the legacy study cannot be reproduced. To improve the performance of the VCGs compared to the "agnostic scenario", two methods were presented: finding the parameter in the study design that influences the outcome of the study, and keeping a certain number of sentinel animals in the concurrent control group.

To find the confounder, the data of the VCG population was plotted against the study year and searched for atypical shifts. A marked drop in electrolyte levels was observed in the VCG sample population in 2017 (Figure 5), which ultimately compromised the performance of the VCGs. A confounder was present but could not be identified immediately. After excluding known stratification parameters—such as strain (Kacew, 1996), route of administration (Gad, 1994), sex, age (Wolford et al., 1987; McCutcheon and Marinelli, 2009), initial body weight (Wolford et al., 1987), vehicle (de Kort et al., 2020) and the laboratory carrying out the test (Igl et al., 2019)—other parameters were checked as possible reasons, such as a change in animal supplier and a potential change in the analytical method. However, none of these parameters provided an explanation for the observed decline in electrolytes in 2017. Discussing the data with the study director led to the identification of an additional parameter, which is not recorded in SEND, namely, the anesthetics used before blood collection. Since blood collection is a stressful procedure for animals, anesthesia with CO_2 or isoflurane is required by animal welfare ordinance (Parasuraman et al., 2010). While CO_2 is a cheap and easy to use gas that does not require resource intensive disposal, isoflurane is considered to be less stressful for the animals (Altholtz et al., 2006; Traslavina et al., 2010; Wong et al., 2013; Turner et al., 2020). But apart from that, it is also documented that CO_2 as an anesthetic can artificially increase blood serum electrolyte levels, as high levels of CO_2 cause blood acidosis (Langford, 2005; Traslavina et al., 2010). Isoflurane, on the other hand, is known to lower electrolyte levels (Hotchkiss et al., 1998; Deckardt et al., 2007). Manual extraction of this information from the original reports and subsequent addition to the database finally confirmed that the change from CO_2 to isoflurane caused the observed changes in electrolytes. Afterwards, the common VCG database was enriched with this parameter in order to generate meaningful VCGs for the simulation of electrolyte parameters.

Aside from finding and controlling confounders, the performance of VCGs has been increased after keeping a certain number of sentinel animals in the CCG set. Concurrent control animals are generally essential to ensure the quality and technical validity of the bioassay (Kluxen et al., 2021). For example, in a study without control animals, a possible infection would go unnoticed and a resulting increase in hematological parameters could be incorrectly attributed to the administered test item (Nicklas et al., 1999; Steger-Hartmann et al., 2020). Already keeping two sentinel animals significantly improved the performance of the VCGs compared to the "agnostic scenario" and keeping half of the CCG animals as sentinel animals resulted in a reproducibility percentage of 100% in all dose groups (see Table 2).

The combination of both methods, i.e., controlling the stratification parameter and keeping sentinel animals, did not improve the performance of the VCGs with respect to the method of "keeping sentinel animals". Sentinel animals attenuated the influence of the confounder in the presented study. This emphasizes their usefulness beyond ensuring the technical validity of a bioassay.

This article demonstrates that rigorous control of the data increases the performance of the VCGs. However, future VCG selection and matching should not be based on solely one parameter and should use several parameters instead. An long-term goal is to create VCGs from historical data that ideally are able cover all endpoints of a toxicological study well. In regulatory toxicology, around 75 quantitative clinical pathology parameters (and other qualitative ones, e.g., histopathology) are examined. Examination of each of these parameters for quality, as has been presented for the electrolyte values in this article and a comprehensive analysis of the parameter distribution at a particular time of measurement would make it possible to identify further confounders and thus continuously improve the quality of the HCD and generate more meaningful VCGs. Another factor to consider is the correlation or interdependence of several parameters. Calcium is known to correlate with phosphate, urea and potassium (Howard et al., 2011; Verzicco et al., 2020), and parathyroid hormone (PTH) (Lecoq et al., 2021), among others. If a parameter was found to differ significantly, it might also be of interest to check whether a correlating endpoint also shows a significant change in the same direction. If so, these findings could be flagged accordingly, which could ultimately help expert toxicologists and study directors differentiate true treatment-related effects from artifacts. In the legacy study, the above-mentioned correlation of calcium and inorganic phosphate (both are increased) were present. Both electrolytes are regulated by Vitamin D3, PTH, and calcitonin (Shrimanker and Bhattarai, 2016). The VCGs—selected to match the parameter calcium—show the same behavior for phosphate as for calcium: a poor performance when the confounder is present and an increase in performance after either leaving sentinel animals in the set or removing the data affected by the confounder (Supplementary Material S4, Supplementary Table S1). Thereby, in this case, we were able to reproduce the statistical results of the study well. In addition, VCGs were examined for a parameter that does not primarily correlate with blood serum electrolyte levels: the body weight, measured on day 28 at the end of the dosing period (Supplementary Material S4, Supplementary Table S2). No significant changes in body weight concentration were noted in the legacy study between the dose groups and the corresponding CCG. Unsurprisingly, the VCGs were able to reproduce these results well as they were selected leaving the original location parameters of the initial body weights of the rats unchanged. Neither removing the data affected by the confounder nor considering sentinel animals did affect the good performance of the VCGs.

5 Conclusion

Our study illustrates the importance of proper data analysis and selection and proposes strategies to generate virtual controls. The aim of this study was to generate VCGs to replace concurrent controls in future experiments, thereby contributing to the 3R concept. VCGs were generated by a resampling approach that proved to be easy to implement and straightforward, yielding results that were easy to interpret. In addition, each individual VCG value can be traced back to each individual animal in a historical study allowing for quality assurance. The performance

of the VCGs was shown to be highly dependent on the quality of the underlying HCD and can be improved by using a small number of remaining concurrent control animals (i.e., sentinel animals). A well maintained and constantly improved database together with thorough statistical characterization of the data will be the main requirements for the implementation of VCG.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: GitHub: https://github.com/bayer-group/VCG-resampling.git.

Ethics statement

Animals were treated in accordance with the German Law on the Protection of Animals and with permission of the state animal welfare committee.

Author contributions

AG: main body first authorship, supplementary authorship, figures, data analysis, statistical analysis LV: main body co-authorship and review AK: data gathering and cleaning, main body co-authorship and review TS-H: research lead, main body co-authorship and review.

Funding

The described research has been performed under the IMI eTRANSAFE project. eTRANSAFE has received support from IMI2 Joint Undertaking under Grant Agreement No. 777365. This Joint Undertaking received support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations (EFPIA).

Acknowledgments

A big thanks to Alexius Freyberger and Julia Vienenkoetter from Bayer for helping in detecting the hidden confounder and interpreting toxicological outcomes, and Joerg Wichard for supporting and reviewing during the development of this article.

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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1142534/full#supplementary-material

References

Altholtz, L. Y., Fowler, K. A., Badura, L. L., and Kovacs, M. S. (2006). Comparison of the stress response in rats to repeated isoflurane or CO_2: O_2 anesthesia used for restraint during serial blood collection via the jugular vein. *J. Am. Assoc. Laboratory Animal Sci.* 45 (3), 17–22.

Baldrick, P. (2008). Safety evaluation to support first-in-man investigations II: Toxicology studies. *Regul. Toxicol. Pharmacol.* 51 (2), 237–243. doi:10.1016/j.yrtph.2008.04.006

Bode, G. (2020). "Regulatory guidance: ICH, EMA, FDA," in *Drug discovery and evaluation: Methods in clinical Pharmacology*, 1085–1138.

Bonapersona, V., Hoijtink, H., Sarabdjitsingh, R., Joëls, M., and Joels, M. (2021). Increasing the statistical power of animal experiments with historical control data. *Nat. Neurosci.* 24 (4), 470–477. doi:10.1038/s41593-020-00792-3

Cdisc, C. D. I. S. C. (2022). SEND [online]. CDISC: Cdisc. Available: https://www.cdisc.org/standards/foundational/send (Accessed August 9, 2022).

Charan, J., and Kantharia, N. (2013). How to calculate sample size in animal studies? *J. Pharmacol. Pharmacother.* 4 (4), 303–306. doi:10.4103/0976-500X.119726

de Kort, M., Weber, K., Wimmer, B., Wilutzky, K., Neuenhahn, P., Allingham, P., et al. (2020). Historical control data for hematology parameters obtained from toxicity studies performed on different wistar rat strains: Acceptable value ranges, definition of severity degrees, and vehicle effects. *Toxicol. Res. Appl.* 4, 239784732093148. doi:10.1177/2397847320931484

Deckardt, K., Weber, I., Kaspers, U., Hellwig, J., Tennekes, H., and van Ravenzwaay, B. (2007). The effects of inhalation anaesthetics on common clinical pathology parameters in laboratory rats. *Food Chem. Toxicol.* 45 (9), 1709–1718. doi:10.1016/j.fct.2007.03.005

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.* 50 (272), 1096–1121. doi:10.1080/01621459.1955.10501294

EFSA Scientific Committee (2011). Guidance on conducting repeated-dose 90-day oral toxicity study in rodents on whole food/feed. *EFSA J.* 9 (12), 2438. doi:10.2903/j.efsa.2011.2438

EMA (2010). "CPMP/SWP/1042/99 Rev 1 Corr* - guideline on repeated dose toxicity," in *European Medicines agency*. [Online]. Available: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-repeated-dose-toxicity-revision-1_en.pdf (Accessed August 9, 2022).

EMA (2013). *ICH guideline M3(R2) on non-clinical safety studies for the conduct of human clinical trials and marketing authorisation for pharmaceuticals*. 5 ed. European Union: European Medicines Agency EMA.

FDA (2000). "Fda – US food and drug administration," in *Toxicological principles for the safety assessment of food ingredients. Redbook 2000. Chapter IV.C.3.a. Short-term toxicity studies with rodents*. https://www.fda.gov/files/food/published/Toxicological-Principles-for-the-Safety-Assessment-of-Food-Ingredients.pdf.

Gad, S. C. (1994). Routes in toxicology: An overview. *J. Am. Coll. Toxicol.* 13 (1), 34–39. doi:10.3109/10915819409140653

Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., et al. (2016). Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood cancer J.* 6 (9), e473. doi:10.1038/bcj.2016.84

Greim, H., Gelbke, H., Reuter, U., Thielmann, H., and Edler, L. (2003). Evaluation of historical control data in carcinogenicity studies. *Hum. Exp. Toxicol.* 22 (10), 541–549. doi:10.1191/0960327103ht394oa

Hamada, C. (2018). Statistical analysis for toxicity studies. *J. Toxicol. pathology* 31 (1), 15–22. doi:10.1293/tox.2017-0050

Hotchkiss, C., Brommage, R., Du, M., and Jerome, C. (1998). The anesthetic isoflurane decreases ionized calcium and increases parathyroid hormone and osteocalcin in cynomolgus monkeys. *Bone* 23 (5), 479–484. doi:10.1016/s8756-3282(98)00124-0

Hothorn, L. A., Kluxen, F. M., and Hasler, M. (2019). Pseudo-data generation allows the statistical re-evaluation of toxicological bioassays based on summary statistics. bioRxiv, Preprint. doi:10.1101/810408

Howard, S. C., Jones, D. P., and Pui, C.-H. (2011). The tumor lysis syndrome. *New England J. Med.* 364 (19), 1844–1854. doi:10.1056/NEJMra0904569

ICH (2020). "Harmonised guideline: Detection of reproductive and developmental toxicity for human pharmaceuticals S5 (R3)," in *International conference on the harmonisation of technical requirements for registration of pharmaceuticals for human use*.

Igl, B.-W., Bitsch, A., Bringezu, F., Chang, S., Dammann, M., Frötschl, R., et al. (2019). The rat bone marrow micronucleus test: Statistical considerations on historical negative control data. *Regul. Toxicol. Pharmacol.* 102, 13–22. doi:10.1016/j.yrtph.2018.12.009

Kacew, S. (1996). Invited review: role of rat strain in the differential sensitivity to pharmaceutical agents and naturally occurring substances. *J. Toxicol. Environ. Health Part A* 47 (1), 1–30.

Keenan, C., Elmore, S., Francke-Carroll, S., Kemp, R., Kerlin, R., Peddada, S., et al. (2009). Best practices for use of historical control data of proliferative rodent lesions. *Toxicol. Pathol.* 37 (5), 679–693. doi:10.1177/0192623309336154

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

Kolker, E., Stewart, E., and Ozdemir, V. (2012). Opportunities and challenges for the life sciences community. *OMICS A J. Integr. Biol.* 16 (3), 138–147. doi:10.1089/omi.2011.0152

Kramer, M., and Font, E. (2017). Reducing sample size in experiments with animals: Historical controls and related strategies. *Biol. Rev.* 92 (1), 431–445. doi:10.1111/brv.12237

Langford, N. J. (2005). Carbon dioxide poisoning. *Toxicol. Rev.* 24 (4), 229–235. doi:10.2165/00139709-200524040-00003

Lecoq, A.-L., Livrozet, M., Blanchard, A., and Kamenický, P. (2021). Drug-related hypercalcemia. *Endocrinol. Metabolism Clin.* 50 (4), 743–752. doi:10.1016/j.ecl.2021.08.001

Lim, J., Walley, R., Yuan, J., Liu, J., Dabral, A., Best, N., et al. (2018). Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities. *Ther. innovation Regul. Sci.* 52 (5), 546–559. doi:10.1177/2168479018778282

McCutcheon, J. E., and Marinelli, M. (2009). Age matters. *Eur. J. Neurosci.* 29 (5), 997–1014. doi:10.1111/j.1460-9568.2009.06648.x

Nicklas, W., Homberger, F. R., Illgen-Wilcke, B., Jacobi, K., Kraft, V., Kunstyr, I., et al. (1999). Implications of infectious agents on results of animal experiments: Report of the Working Group on Hygiene of the Gesellschaft für Versuchstierkunde-Society for Laboratory Animal Science (GV-SOLAS). *Lab. Anim.* 33 (1), 39–87.

OECD (2008). *Test No. 407: Repeated dose 28-day oral toxicity study in rodents*.

OECD (2018a). *Test No. 408: Repeated dose 90-day oral toxicity study in rodents*.

OECD (2018b). *Test No. 453: Combined chronic toxicity/carcinogenicity studies*.

OECD (2016). *Test No. 487*: In vitro *mammalian cell micronucleus test*.

Parasuraman, S., Raveendran, R., and Kesavan, R. (2010). Blood sample collection in small laboratory animals. *J. Pharmacol. Pharmacother.* 1 (2), 87–93. doi:10.4103/0976-500X.72350

Pinches, M. D., Thomas, R., Porter, R., Camidge, L., and Briggs, K. (2019). Curation and analysis of clinical pathology parameters and histopathologic findings from eTOXsys, a large database project (eTOX) for toxicologic studies. *Regul. Toxicol. Pharmacol.* 107, 104396.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *J. chronic Dis.* 29 (3), 175–188. doi:10.1016/0021-9681(76)90044-8

Pognan, F., Steger-Hartmann, T., Díaz, C., Blomberg, N., Bringezu, F., Briggs, K., et al. (2021). The eTRANSAFE project on translational safety assessment through integrative knowledge management: Achievements and perspectives. *Pharmaceuticals* 14 (3), 237. doi:10.3390/ph14030237

Rosenbaum, P. R., and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Statistician* 39 (1), 33–38. doi:10.2307/2683903

Russel, W., and Burch, R. (1959). The principles of humane experimental technique. *ALTEX* 3, 94.

Sawamoto, R., Oba, K., and Matsuyama, Y. (2022). *Bayesian adaptive randomization design incorporating propensity score-matched historical controls*. Pharmaceutical Statistics.

Schenck, P. A., Chew, D. J., Nagode, L. A., and Rosol, T. J. (2006). Disorders of calcium: Hypercalcemia and hypocalcemia. *Fluid, electrolyte, acid-base Disord. small animal Pract.* 4, 120–194.

Schmidt, K., Schmidtke, J., Kohl, C., Wilhelm, R., Schiemann, J., van der Voet, H., et al. (2016). Enhancing the interpretation of statistical P values in toxicology studies: Implementation of linear mixed models (LMMs) and standardized effect sizes (SESs). *Archives Toxicol.* 90, 731–751. doi:10.1007/s00204-015-1487-8

Shrimanker, I., and Bhattarai, S. (2016). "Electrolytes," in *StatPearls*. Treasure Island (FL); StatPearls Publishing. 2022. PMID: 31082167.

Steger-Hartmann, T., Kreuchwig, A., Vaas, L., Wichard, J., Bringezu, F., Amberg, A., et al. (2020). Introducing the concept of virtual control groups into preclinical toxicology testing. *ALTEX-Alternatives animal Exp.* 37 (3), 343–349. doi:10.14573/altex.2001311

Stokes, A. H., Kemp, D. C., Faiola, B., Jordan, H. L., Merrill, C. L., Hailey, J. R., et al. (2013). Effects of Solutol (Kolliphor) and cremophor in polyethylene glycol 400 vehicle formulations in Sprague-Dawley rats and beagle dogs. *Int. J. Toxicol.* 32 (3), 189–197. doi:10.1177/1091581813485452

Strayhorn, J. M. (2021). Virtual controls as an alternative to randomized controlled trials for assessing efficacy of interventions. *BMC Med. Res. Methodol.* 21 (1), 3–14. doi:10.1186/s12874-020-01191-9

Tinawi, M. (2021). Disorders of calcium metabolism: Hypocalcemia and hypercalcemia. *Cureus* 13 (1), e14619. doi:10.7759/cureus.14619

Traslavina, R. P., King, E. J., Loar, A. S., Riedel, E. R., Garvey, M. S., Ricart-Arbona, R., et al. (2010). Euthanasia by CO_2 inhalation affects potassium levels in mice. *J. Am. Assoc. Laboratory Animal Sci.* 49 (3), 316–322.

Turner, P., Hickman, D., van Luijk, J., Ritskes-Hoitinga, M., Sargeant, J., Kurosawa, T., et al. (2020). Welfare impact of carbon dioxide euthanasia on laboratory mice and rats: A systematic review. *Front. Vet. Sci.* 7, 411. doi:10.3389/fvets.2020.00411

Verzicco, I., Regolisti, G., Quaini, F., Bocchi, P., Brusasco, I., Ferrari, M., et al. (2020). Electrolyte disorders induced by antineoplastic drugs. *Front. Oncol.* 10, 779. doi:10.3389/fonc.2020.00779

White, W. J., and Cham, S. (1998). The development and maintenance of the crl: CH!! J (SD) IGSBR rat breeding System. *Biol. Ref. Data CD (SD) IGS Rats.*, 8–14.

Wolford, S., Schroer, R., Gallo, P., Gohs, F., Brodeck, M., Falk, H., et al. (1987). Age-related changes in serum chemistry and hematology values in normal Sprague-Dawley rats. *Fundam. Appl. Toxicol.* 8 (1), 80–88. doi:10.1016/0272-0590(87)90102-3

Wong, D., Makowska, I. J., and Weary, D. M. (2013). Rat aversion to isoflurane versus carbon dioxide. *Biol. Lett.* 9 (1), 20121000. doi:10.1098/rsbl.2012.1000

Wood, F. K., and Lou, A. (2011). *The standard for the Exchange of nonclinical data (SEND): History and basics*.

Wright, P. S., Smith, G. F., Briggs, K. A., Thomas, R., Maglennon, G., Mikulskis, P., et al. (2023). Retrospective analysis of the potential use of virtual control groups in preclinical toxicity assessment using the eTOX database. *Regul. Toxicol. Pharmacol.* 138, 105309. doi:10.1016/j.yrtph.2022.105309

Zhan, T., Zhou, Y., Geng, Z., Gu, Y., Kang, J., Wang, L., et al. (2022). Deep historical borrowing framework to prospectively and simultaneously synthesize control information in confirmatory clinical trials with multiple endpoints. *J. Biopharm. Statistics* 32 (1), 90–106. doi:10.1080/10543406.2021.1975128

3.1 Supplementary Material

Table of Contents

1 Methods

1.1 The data

All animal studies performed after 2011 were recorded using Pristima (Xybion) Laboratory Information Management System (LIMS). Pristima's SEND solution, called Savante was utilized for the conversion of the collected raw data into the SEND (Standard for Exchange of Non-clinical Data) data model. Additionally, external SEND studies which were conducted at CROs (Contract Research Organizations) are accessible. Harmonization of the study data was performed according to the SEND controlled terminology for enabling data analysis. The data is being stored in AWS S3 and easily accessible for data scientists at Bayer.

1.2 Used software for data selection, curation, and visualization

Data gathering was performed with a version of the statistical software R, version 3.5.3. The processing, and cleaning of the data was done with R, version 4.0.1, using the tidyverse package (Wickham, 2017) and the data.table package (Dowle and Srinivasan, 2021). Statistical evaluation was performed using the DescTools package (Signorell and al., 2019). Data visualization was performed with, using the plotly package (Sievert, 2018) along with the webshot package for exporting images (Winston, 2019). The used R-code can be downloaded from the GitHub repository https://github.com/bayer-group/VCG-resampling.git.

1.3 Difference between potassium and calcium levels of rats anesthetized with CO_2/air and CO_2/O_2 in the data set

During the procedure of blood drawing in rats, anesthetics are used. One possible anesthetic is CO_2 which is however not administered purely but as a mixture. In this data set from Bayer, in most cases, CO_2 was mixed with room air in a 6:4 ratio (60 % CO_2 and 40 % O_2). In some studies however, a mixture of 80 % CO_2 and 20 % O_2 was used insted. In order to ensure that these additives to the CO_2 were not influencing the outcome on the electrolyte values, the values were compared to each other. For the potassium values, a histogram (Panel A of Figure S1) and a box plot (Panel B of Figure S1) is illustrated, showing no visible differences between these two subgroups in the potassium values. Same was done for calcium (Figure S2), sodium (Figure S3) and inorganic phosphate (Figure S4).

It was therefore concluded that both sub-sets can be combined into one group as they have no difference in the potassium and calcium values-in relation to the serum values of the rats which were anesthetized with isoflurane.

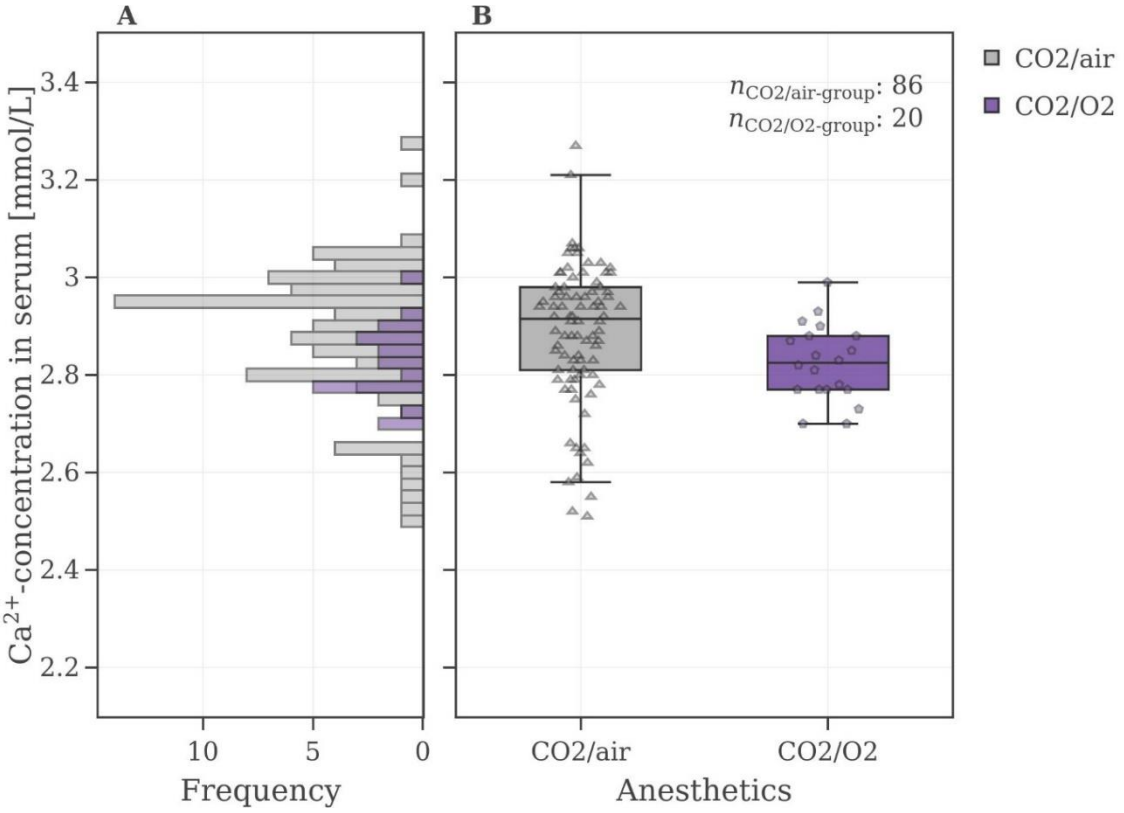

Figure S 1: (A) Calcium value distributions of male Wistar-rats (B) Box plots of these calcium levels with respect to the anesthetic. The CO_2/air-group is colored grey, and the CO_2/O_2-group is colored violet.

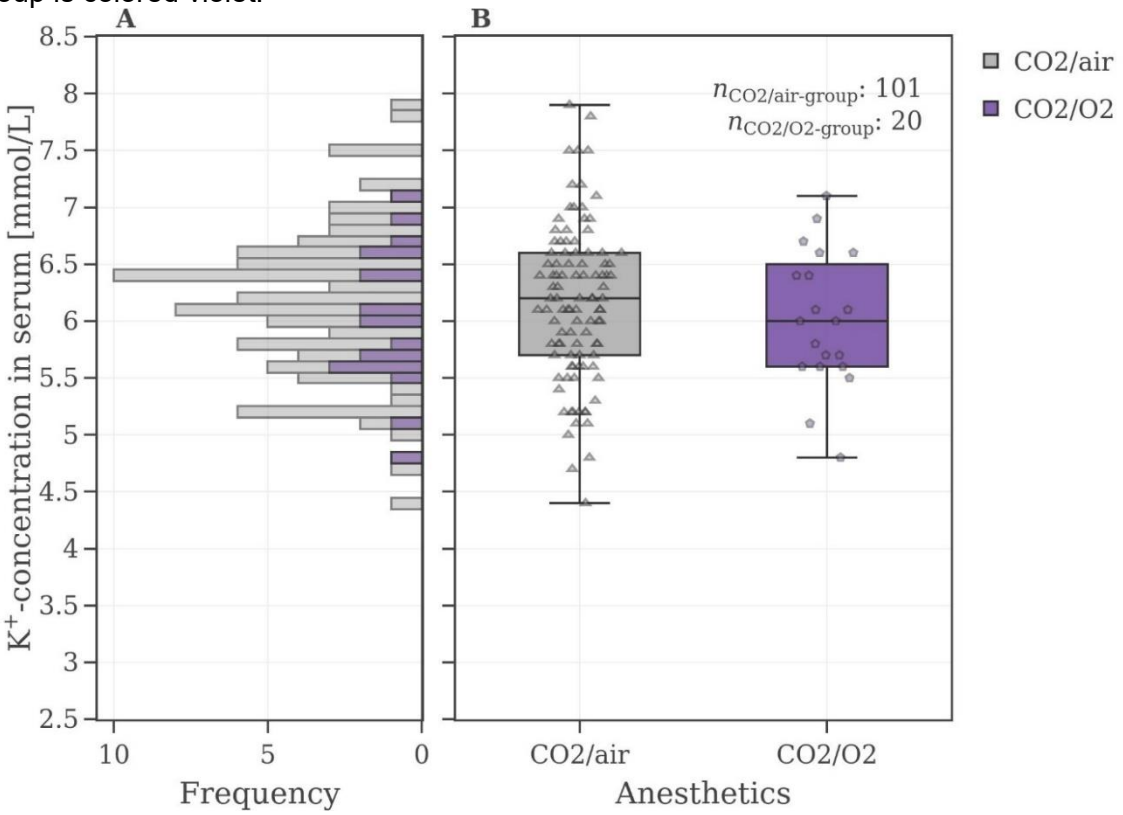

Figure S 2: (A) Potassium value distributions of male Wistar-rats (B) Box plots of these potassium levels with respect to the anesthetic. The CO_2/air-group is colored grey, and the CO_2/ O_2-group is colored violet.

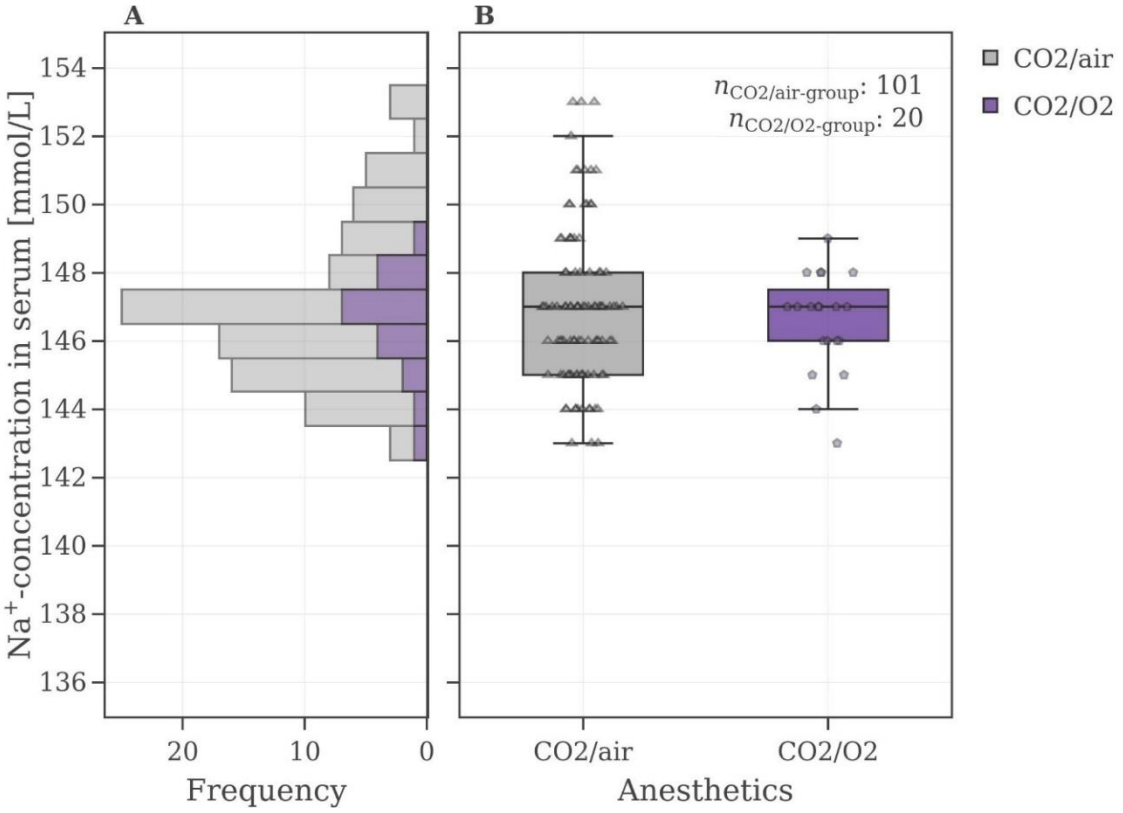

Figure S 3: (A) Sodium value distributions of male Wistar-rats (B) Box plots of these sodium levels with respect to the anesthetic. The CO_2/air-group is colored grey, and the CO_2/ O_2-group is colored violet.

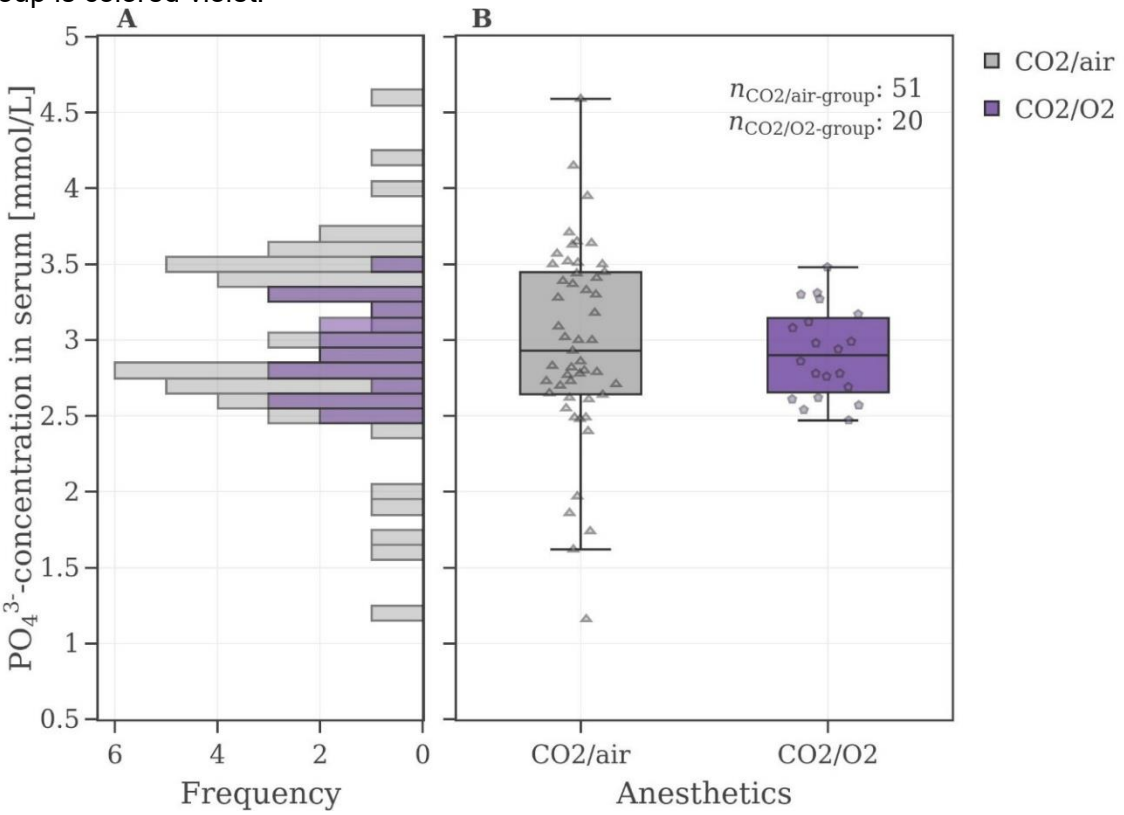

Figure S 4: (A) Phosphate value distributions of male Wistar-rats (B) Box plots of these phosphate levels with respect to the anesthetic. The CO_2/air-group is colored grey, and the CO_2/ O_2-group is colored violet.

2 Results

2.1 Legacy study graphical results

This section shows the overall results of all 500 iterations of all six scenarios used to test the VCG performance. All recruitment scenarios are shown in Figure S9. For the numerical results of the performances for each scenario please refer to Table 2 of the main body of this article. The graphs illustrate the mean values of the VCGs in each iteration and the outcoming result for each dose group. In the first scenario (1A), *i.e.*, the “agnostic scenario”, which is shown in Figure S10, a high variance of the mean values of the VCGs can be seen. As a result, virtual controls with a mean value too far away from the concurrent one tends to lead to results inconsistent to the one of the legacy study. The performance of the VCGs was improved by two approaches. One approach was to keep a number of values from the legacy study of satellite animals in the set. Keeping two satellite animals (Figure S11) reduced the variance of the VCG means and improved the performance considerably; keeping five animals (*i.e.*, half of the original control group population) (Figure S12) improved the performance even further. Another strategy was to control the confounding factor (*i.e.*, the anesthetic used) as shown in Figure S13 which too, improved the performance, though not as good as keeping sentinel animals. Combining both methods, *i.e.*, controlling the confounder and keeping two sentinel animals (Figure S14) or five sentinel animals (Figure S15) did not further improve the performance of the virtual controls.

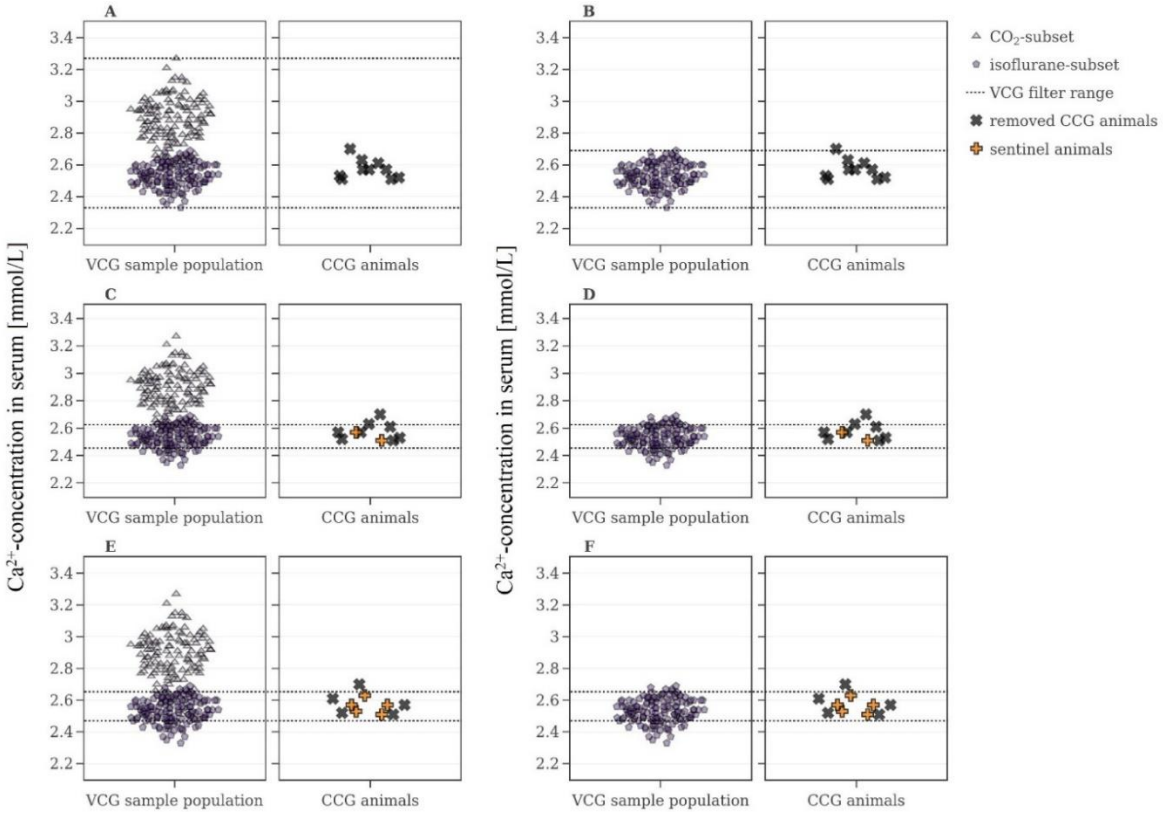

Figure S5: All scenarios of the VCG selection process. (A) The “agnostic-scenario”) No sentinel animals kept; confounder not controlled. (B) No sentinel animals kept; confounder controlled. (C) 2 sentinel animals kept; confounder not controlled. (D) 2 sentinel animals kept; confounder controlled. (E) Half of the CCG-animals kept; confounder not controlled. (F) Half of the CCG animals kept; confounder controlled.

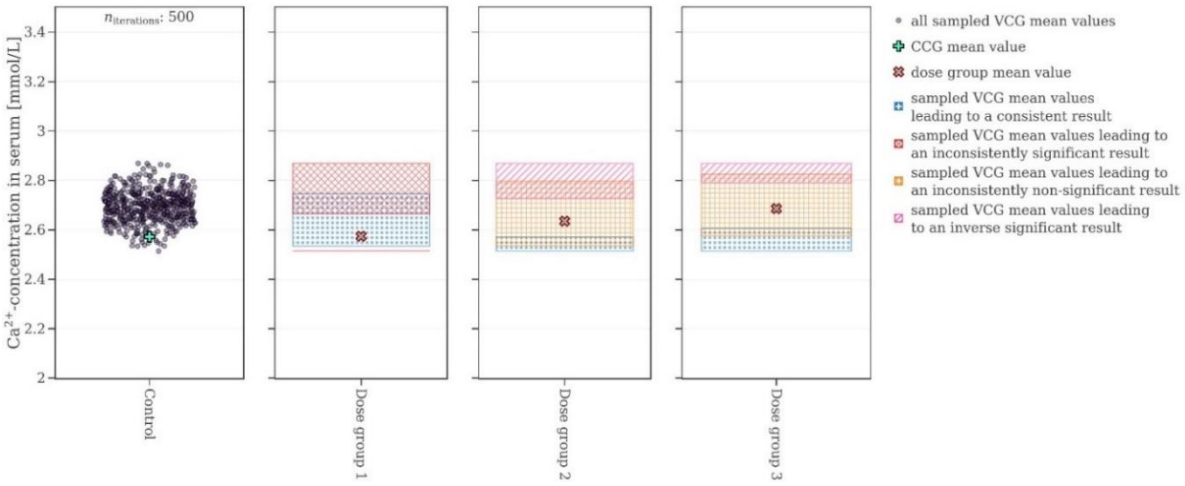

Figure S6: Scenario 1a: confounder not controlled; no sentinel animals kept. Resampling results of the legacy study in male rats in each dose group. Virtual control groups (VCGs) were generated using the calcium sample population of the respective subset The mean value of concurrent control group (CCG) (green cross) and the dose group (grey X) are shown for each dose group. Additionally, all mean values of the sampled VCGs of 500 iterations are shown as a scattered violet cloud in the control-group panel. On the dose groups, the areas for the VCG from each iteration are shown as “zones”. If a VCG led to a result consistent with the one using the CCG, the zone is blue with a dotted pattern. VCGs leading to an inconsistently significant results are red with a crossed pattern. VCGs leading to an inconsistently non-significant results are orange with a checkered pattern. And finally, VCGs leading to an inverse significant result are magenta with a with a diagonally stroked pattern.

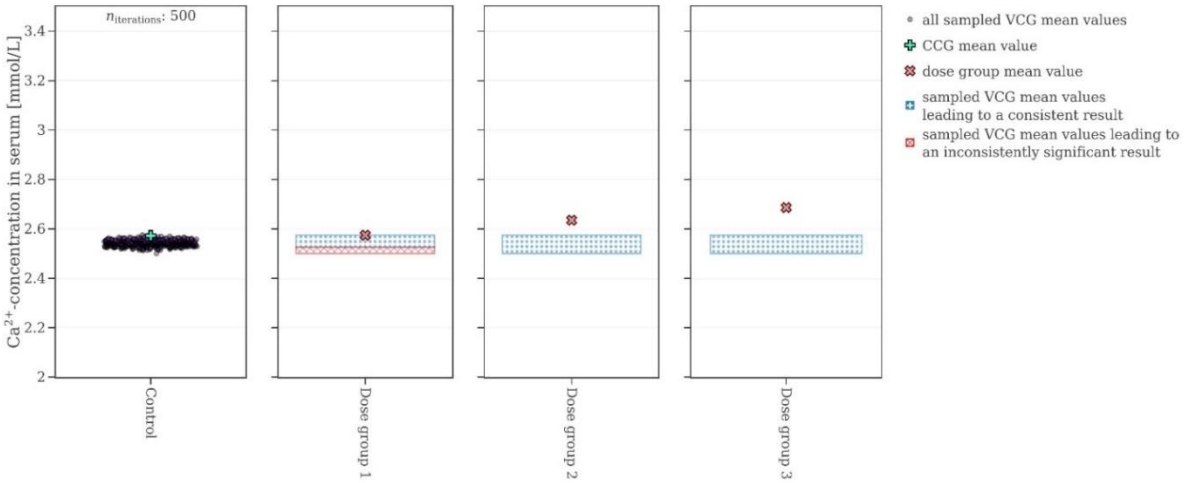

Figure S7: Scenario 1b: confounder not controlled; two sentinel animals kept. Resampling results of the legacy study in male rats in each dose group. Virtual control groups (VCGs) were generated using the calcium sample population of the respective subset The mean value of concurrent control group (CCG) (green cross) and the dose group (grey X) are shown for each dose group. Additionally, all mean values of the sampled VCGs of 500 iterations are shown as a scattered violet cloud in the control-group panel. On the dose groups, the areas for the VCG from each iteration are shown as “zones”. If a VCG led to a result consistent with the one using the CCG, the zone is blue with a dotted pattern. VCGs leading to an inconsistently significant results are red with a crossed pattern. VCGs leading to an inconsistently non-significant results are orange with a checkered pattern. And finally, VCGs leading to an inverse significant result are magenta with a with a diagonally stroked pattern.

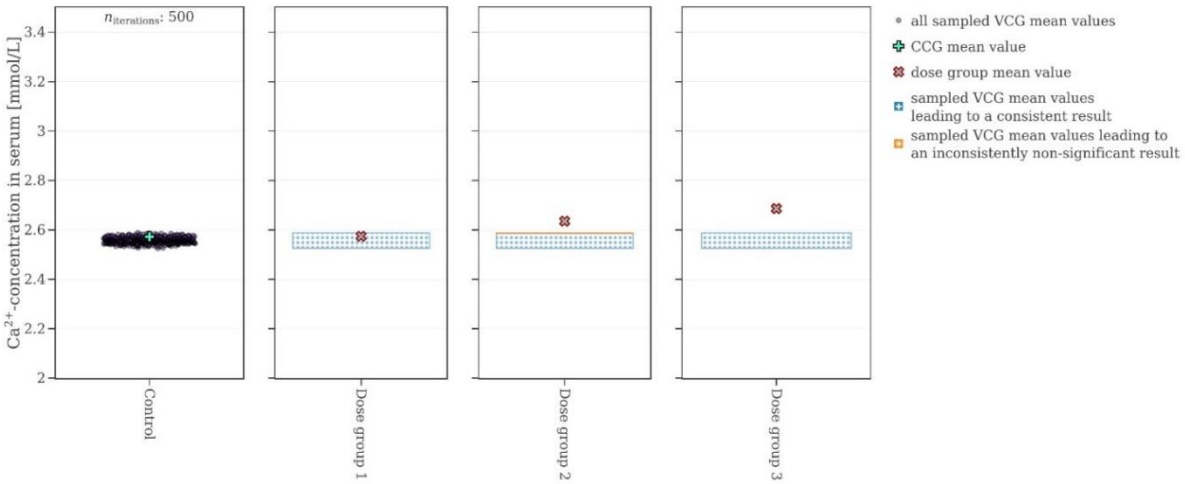

Figure S8: Scenario 1c: confounder not controlled; half of the CCG animals are kept as sentinel animals. Resampling results of the legacy study in male rats in each dose group. Virtual control groups (VCGs) were generated using the calcium sample population of the respective subset The mean value of concurrent control group (CCG) (green cross) and the dose group (grey X) are shown for each dose group. Additionally, all mean values of the sampled VCGs of 500 iterations are shown as a scattered violet cloud in the control-group panel. On the dose groups, the areas for the VCG from each iteration are shown as “zones”. If a VCG led to a result consistent with the one using the CCG, the zone is blue with a dotted pattern. VCGs leading to an inconsistently significant results are red with a crossed pattern. VCGs leading to an inconsistently non-significant results are orange with a checkered pattern. And finally, VCGs leading to an inverse significant result are magenta with a with a diagonally stroked pattern.

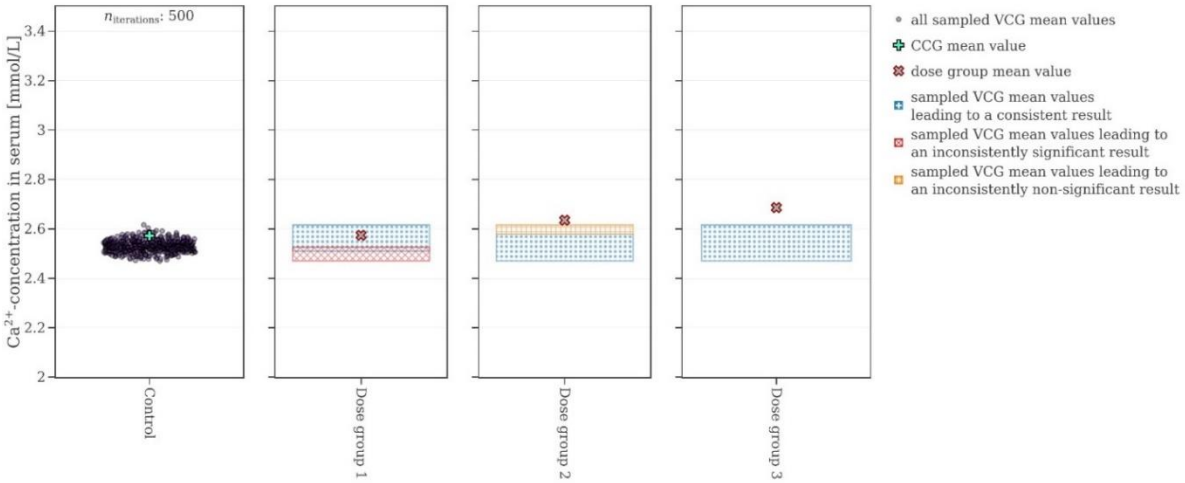

Figure S9: Scenario 2a: confounder controlled; no sentinel animals kept. Resampling results of the legacy study in male rats in each dose group. Virtual control groups (VCGs) were generated using the calcium sample population of the respective subset The mean value of concurrent control group (CCG) (green cross) and the dose group (grey X) are shown for each dose group. Additionally, all mean values of the sampled VCGs of 500 iterations are shown as a scattered violet cloud in the control-group panel. On the dose groups, the areas for the VCG from each iteration are shown as “zones”. If a VCG led to a result consistent with the one using the CCG, the zone is blue with a dotted pattern. VCGs leading to an inconsistently significant results are red with a crossed pattern. VCGs leading to an inconsistently non-significant results are orange with a checkered pattern. And finally, VCGs leading to an inverse significant result are magenta with a with a diagonally stroked pattern.

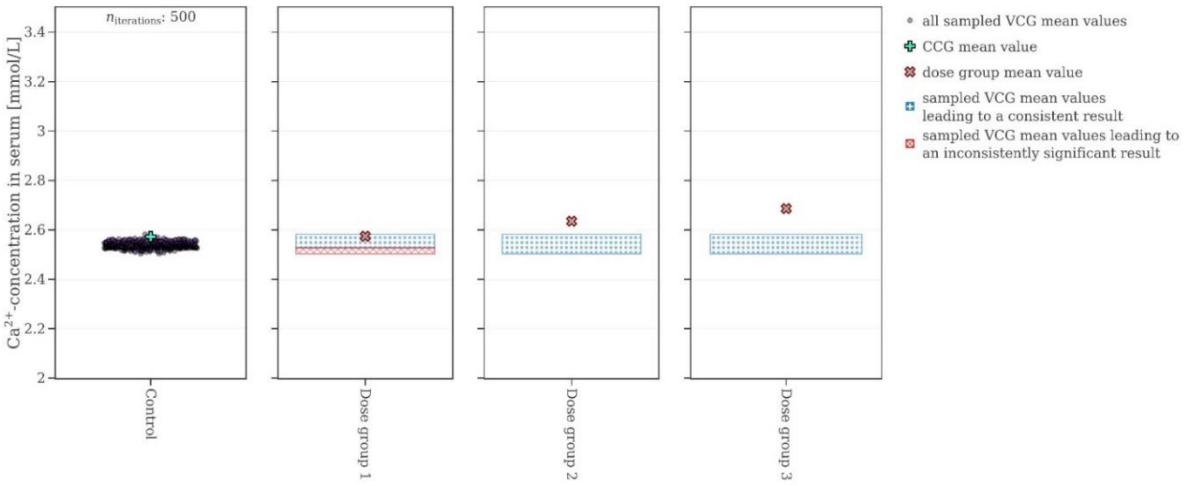

Figure S10: Scenario 2b: confounder controlled; two sentinel animals kept. Resampling results of the legacy study in male rats in each dose group. Virtual control groups (VCGs) were generated using the calcium sample population of the respective subset The mean value of concurrent control group (CCG) (green cross) and the dose group (grey X) are shown for each dose group. Additionally, all mean values of the sampled VCGs of 500 iterations are shown as a scattered violet cloud in the control-group panel. On the dose groups, the areas for the VCG from each iteration are shown as “zones”. If a VCG led to a result consistent with the one using the CCG, the zone is blue with a dotted pattern. VCGs leading to an inconsistently significant results are red with a crossed pattern. VCGs leading to an inconsistently non-significant results are orange with a checkered pattern. And finally, VCGs leading to an inverse significant result are magenta with a with a diagonally stroked pattern.

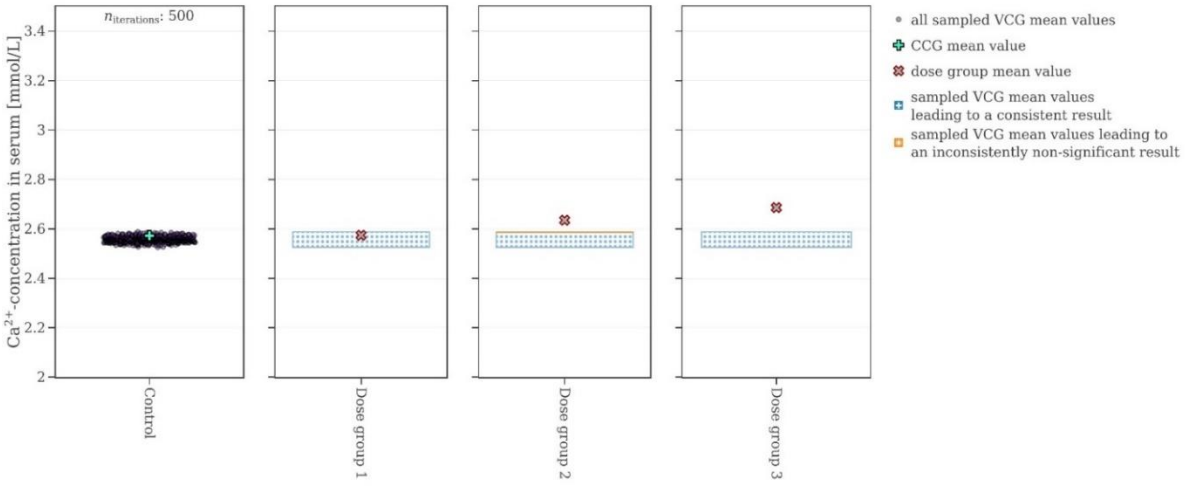

Figure S11: Scenario 2c: confounder controlled; half of the CCG animals are kept as sentinel animals. Resampling results of the legacy study in male rats in each dose group. Virtual control groups (VCGs) were generated using the calcium sample population of the respective subset The mean value of concurrent control group (CCG) (green cross) and the dose group (grey X) are shown for each dose group. Additionally, all mean values of the sampled VCGs of 500 iterations are shown as a scattered violet cloud in the control-group panel. On the dose groups, the areas for the VCG from each iteration are shown as “zones”. If a VCG led to a result consistent with the one using the CCG, the zone is blue with a dotted pattern. VCGs leading to an inconsistently significant results are red with a crossed pattern. VCGs leading to an inconsistently non-significant results are orange with a checkered pattern. And finally, VCGs leading to an inverse significant result are magenta with a with a diagonally stroked pattern.

2.2 Electrolyte values with respect to the used anesthetic

This section illustrates the differences in the electrolyte values of control-group animals from studies with as a histogram and as box plots with respect to the study year. The graphs are separated by color to illustrate the different anesthetics used in the studies. In each electrolyte value, namely calcium (Figure S12), potassium (Figure S13), sodium (Figure S14), and inorganic phosphate (Figure S15) there is a bimodal distribution visible in the histogram as well as a drop in the box plots from 2016 to 2017, *i.e.*, the year when the anesthetic procedure was changed from CO_2 to isoflurane.

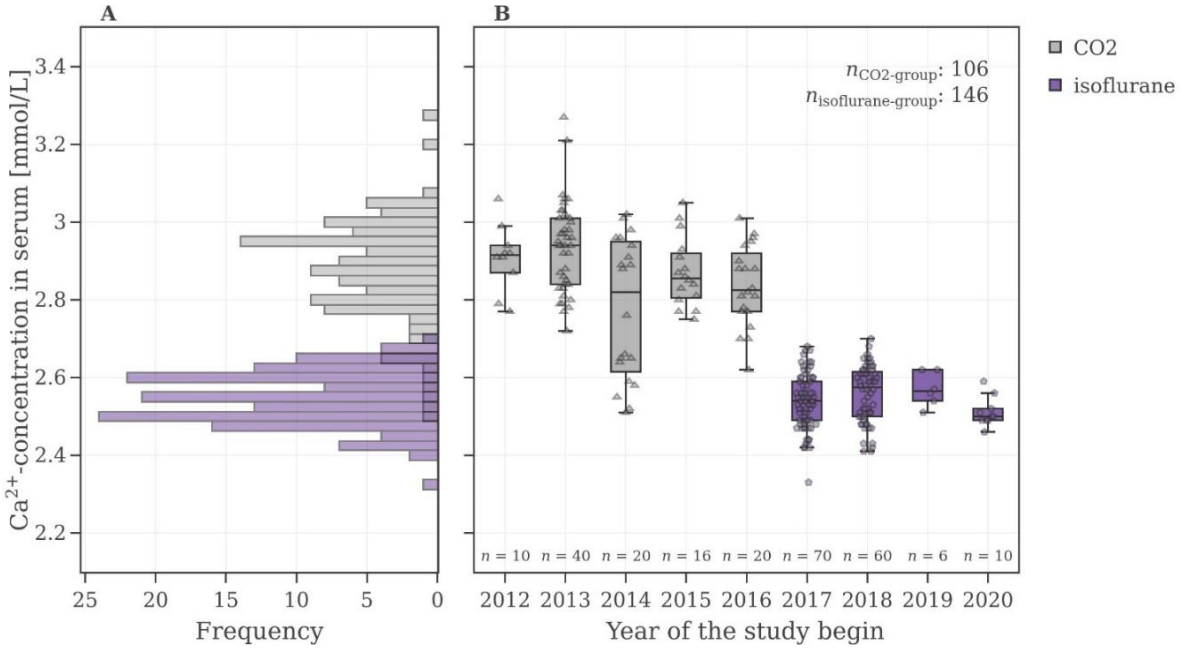

Figure S12: (A) Calcium value distributions of male Wistar-rats (B) Box plots of these calcium levels with respect to the study year. (C) Calcium values as box plots with respect to the anesthetic. The CO_2-group is colored grey, and the isoflurane-group is colored violet.

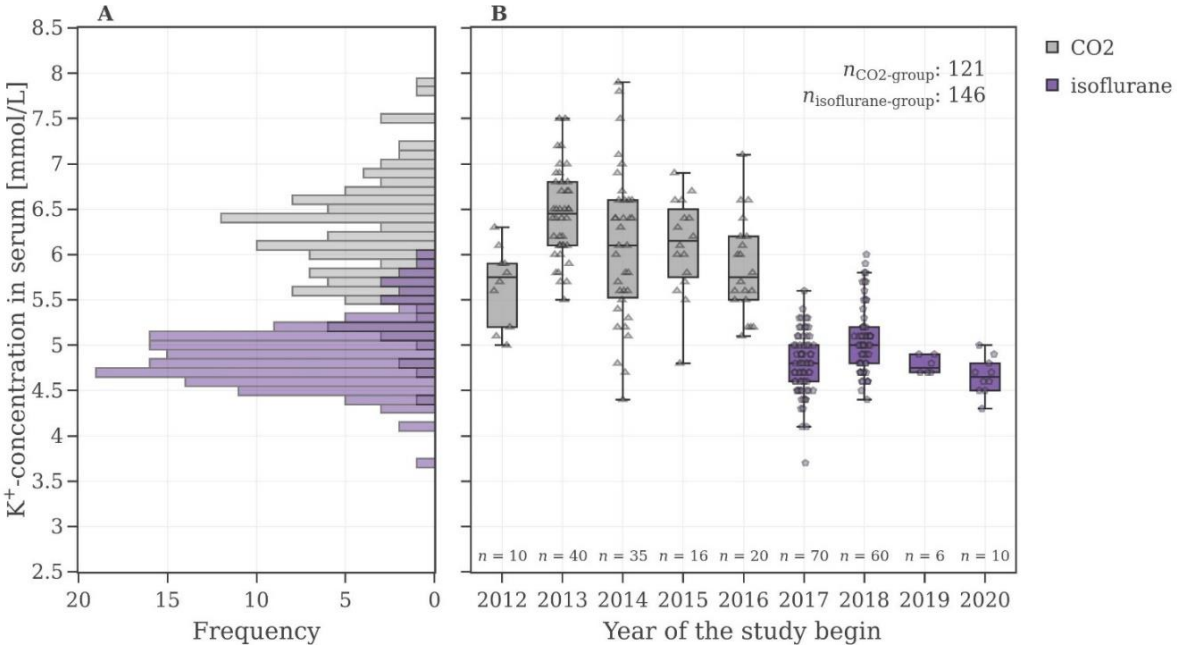

Figure S13: (A) Potassium value distributions of male Wistar-rats (B) Box plots of these potassium levels with respect to the study year. (C) Potassium values as box plots with respect to the anesthetic. The CO_2-group is colored grey, and the isoflurane-group is colored violet.

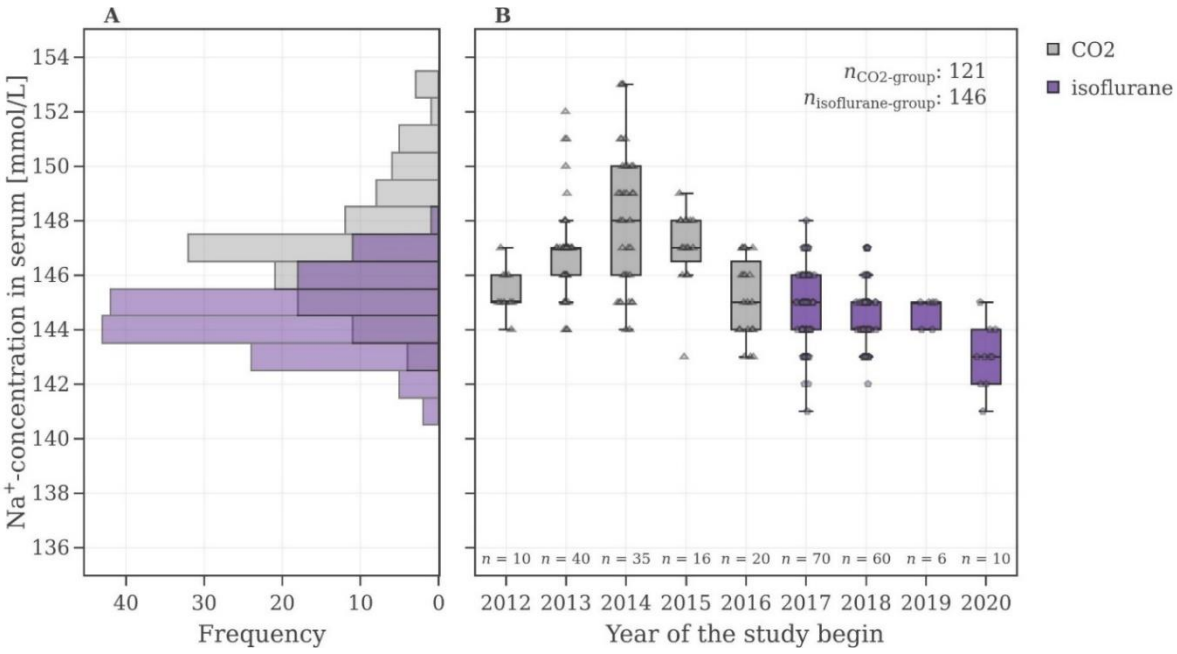

Figure S14: (A) Sodium value distributions of male Wistar-rats (B) Box plots of these sodium levels with respect to the study year. (C) Sodium values as box plots with respect to the anesthetic. The CO_2-group is colored grey, and the isoflurane-group is colored violet.

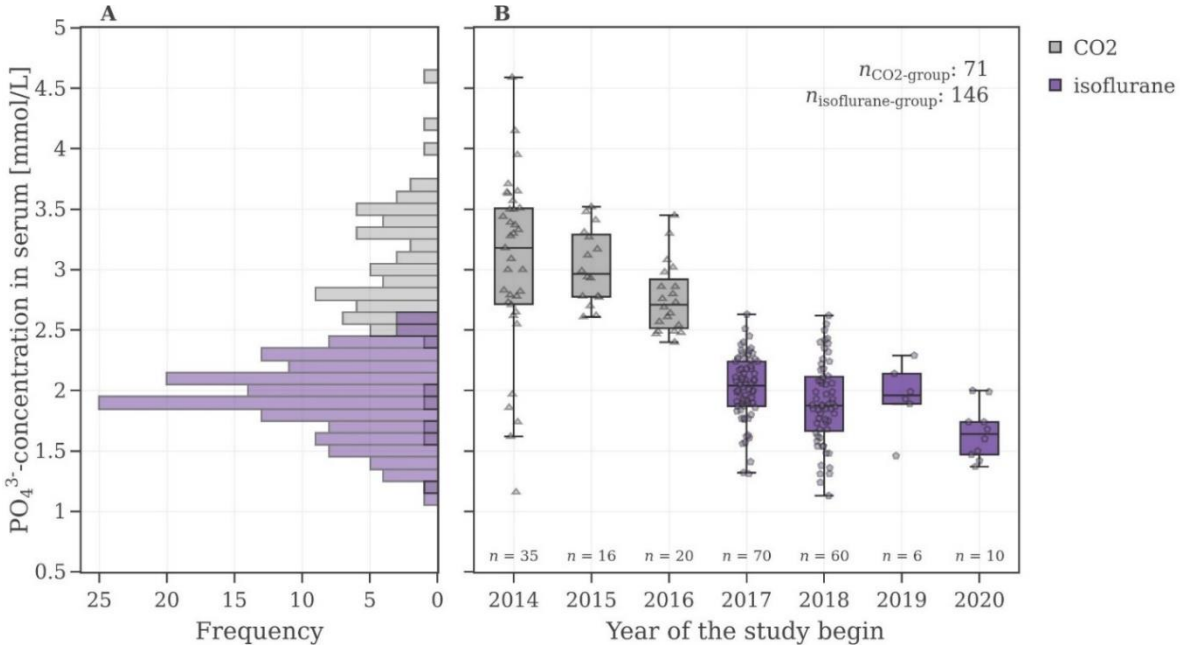

Figure S15: (A) Phosphate value distributions of male Wistar-rats (B) Box plots of these phosphate levels with respect to the study year. (C) Phosphate values as box plots with respect to the anesthetic. The CO_2-group is colored grey, and the isoflurane-group is colored violet.

2.3 Resampling results on phosphate and body weight

In the supplementary Table S1 we show the VCG performance for the parameter inorganic phosphate, a parameter strongly correlated with calcium which is the main endpoint observed in this article. The legacy study which is used here as the reference reported a significantly increased phosphate value in Dose group 3 (i.e., the high dose group) while Dose group 1 and 2 did not show any significant differences. The reproducibility % shows the same behavior in this endpoint as observed in calcium (see Table 2 of the main body of the article). Namely, the performance of the “agnostic scenario” is very poor due to the presence of a confounder. Upon removing all data points affected by the confounder, the performance increased considerably. The presence of sentinel animals improved the performance even further. Note that the VCG was filtered to match the sentinel animals in the parameter calcium. Phosphate—being strongly correlated to calcium—profited from this additional filtering step as well.

Apart from phosphate, another parameter was observed where it wasn’t expected that the anesthetic procedure (being a confounder for electrolyte values measured in blood serum) would affect the values: the body weight measured on day 28 of the study (supplementary Table S2). This parameter indeed didn’t show any improvements in performance after removing all data from animals affected by the confounder. However, keeping five sentinel animals in the set of the concurrent control group improved the performance slightly.

Table S1: Resampling results of the legacy study on the parameter phosphate after replacing the concurrent control group with virtual control groups (VCG) sampled from the respective subgroups. The sampling was performed 500 times and the percentage of consistent statistical results are given for each sex and each dose group (DG).

Mean value of the CCG [mmol/L]	Scenario	Mean value of the VCG sample population [mmol/L]	Sub-scenario	DG 1 consistency	DG 2 consistency	DG 3 consistency
1.52 ± 0 28	1: Confounder is unknown	2.30 ± 0.65	1a: Replace all CCG animals	**11 % consistently non-significant** 89 % inconsistently significant	**18% consistently non-significant** 82 % inconsistently significant	**4 % consistently significant** 93 % inconsistently non-significant 3 % inverted significant
			1b: Keep 2 sentinel animals	**95 % consistently non-significant** 5 % inconsistently significant	**99 % consistently non-significant** 1 % inconsistently significant	**66 % consistently significant** 34 % inconsistently non-significant
			1c: Replace half of the CCG animals	**100 % consistently non-significant**	**100 % consistently non-significant**	**82 % consistently significant** 18 % inconsistently non-significant
	2: Confounder is known	1.94 ± 0.31	2a: Replace all CCG animals	**62 % consistently non-significant** 38 % inconsistently significant	**80 % consistently non-significant** 20 % inconsistently significant	**83 % consistently significant** 17 % inconsistently non-significant
			2b: Keep 2 sentinel animals	**99 % consistently non-significant** 1 % inconsistently significant	**100 % consistently non-significant**	**97 % consistently significant** 3 % inconsistently non-significant
			2c: Replace half of the CCG animals	**100 % consistently non-significant**	**99 % consistently non-significant** 1 % inconsistently significant	**100 % consistently significant**

Table S2: Resampling results of the legacy study on the parameter body weight (on day 28) after replacing the concurrent control group with virtual control groups (VCG) sampled from the respective subgroups. The sampling was performed 500 times and the percentage of consistent statistical results are given for each sex and each dose group (DG).

Mean value of the CCG [g]	Scenario	Mean value of the VCG sample population [g]	Sub-scenario	DG 1 consistency	DG 2 consistency	DG 3 consistency
319 ± 24	1: Confounder is unknown	303 ± 23	1a: Replace all CCG animals	**94 % consistently non-significant** 6 % inconsistently significant	**100 % consistently non-significant**	**100 % consistently non-significant**
			1b: Keep 2 sentinel animals	**95 % consistently non-significant** 5 % inconsistently significant	**100 % consistently non-significant**	**100 % consistently non-significant**
			1c: Replace half of the CCG animals	**100 % consistently non-significant**	**100 % consistently non-significant**	**100 % consistently non-significant**
	2: Confounder is known	295 ± 21	2a: Replace all CCG animals	**76 % consistently non-significant** 24 % inconsistently significant	**95 % consistently non-significant** 5 % inconsistently significant	**100 % consistently non-significant**
			2b: Keep 2 sentinel animals	**96 % consistently non-significant** 4 % inconsistently significant	**99 % consistently non-significant** 1 % inconsistently significant	**100 % consistently non-significant**
			2c: Replace half of the CCG animals	**100 % consistently non-significant**	**100 % consistently non-significant**	**100 % consistently non-significant**

3 List of proposed study design parameters to filter for VCG selection.

In the supplementary Table S3 we present a set of parameters we propose to control in order to derive meaningful virtual controls.

Table S3: List of study design parameters proposed to use as filters for VCG selection.

Parameter (SEND-name)	**Selected value**	**Rationale**
Essential		
Species (SPECIES)	RAT	Animals differ strongly in their parameters.
Strain (STRAIN)	WISTAR HAN	Different strains, differ in various parameters.(de Kort et al., 2020)
Route of administration (ROUTE)	ORAL GAVAGE	Different routes of administration may alter the outcome of parameters (*e.g.,* skin findings after intravenous injections)(Gad, 1994).
Dosing duration (DOSDUR)	at least 28 d	Studies shorter than 4 weeks often are performed for exploratory purposes in a non-standardized setting where not all parameters of a GLP 4-week study are monitored.
Study day (BWDY/LBDY)	1 - 35	It was decided to observe only the study days 1 - 28, however 1 week was added in case that some time passed between finishing the dosing period and measuring respective parameters.
Aspect to be considered		
Study start-year (STDSTDTC)	2018 – 2022	Genetic drift may introduce variation of parameters over time, which should be prevented. In addition, analytical methods

		may change over time (e.g., flame photometry vs. Ion-selective electrodes for determination of cations)
Initial age (AGE)	6 weeks – 9 weeks	Aging affects many parameters rats. However, the age entries are rather vague. In addition, animals are usually ordered according to weight and not age. Thus, the body weight should be used as a surrogate for age.(Wolford et al., 1987; de Kort et al., 2020)
Body weight (BWORRES) (as a surrogate for age)	100 g – 250 g	Controlling body weight prevents entering abnormally young/old animals in the set which might influence other parameters (e.g., ALT, ASP, GGT). It is proposed to select a body weight distribution similar to the one reported for the study under investigation.
Supplier/Breeder (SPLRNAM)	CHARLES RIVER	There is weak evidence for a change in supplier which may influencing parameters (potassium, calcium). Further investigations needed.
Test facility location (TSTFLOC)	E.g., Bayer Pharmaceuticals, Wuppertal	The laboratory has a high impact on different measured endpoints and it is recommended by the guidelines to use historical data from only one laboratory.(OECD, 2008; Keenan et al., 2009) While taking data from only one laboratory is preferable, certain endpoints might still be enriched by consulting data from different laboratories as well as long as GLP criteria are met.

Useful		
Treatment vehicle (TRTV)	KolliphorHS15	Different vehicles may have a different impact on the metabolism of the animals.(Stokes et al., 2013; de Kort et al., 2020)
Time from beginning of the substance treatment until sacrifice of animals (TRMSAC).	1 - 35	Unlike the dosing duration, this parameter takes the “recovery period” into account. During this time, control animals are not exposed to the treatment vehicle which can have an outcome on certain values.

4 References

de Kort, M.,Weber, K.,Wimmer, B.,Wilutzky, K., Neuenhahn, P., Allingham, P., et al. (2020). Historical control data for hematology parameters obtained from toxicity studies performed on different wistar rat strains: Acceptable value ranges, definition of severity degrees, and vehicle effects. Toxicol. Res. Appl. 4, 239784732093148. https://doi.org/10.1177/2397847320931484.

Dowle, M. and Srinivasan, A. (2023). Data.Table: Extension of `data.Frame`. R package version 1.14.8. Https://cran.R-project.Org/package=data.Table.

Gad, S. C. (1994). Routes in toxicology: An overview. J. Am. Coll. Toxicol. 13 (1), 34–39. https://doi.org/10.3109/10915819409140653.

Keenan, C., Elmore, S., Francke-Carroll, S., Kemp, R., Kerlin, R., Peddada, S., et al. (2009). Best practices for use of historical control data of proliferative rodent lesions. Toxicologic pathology 37(5), 679-693. https://doi.org/10.1177/0192623309336154.

OECD (2008). Test no. 407: Repeated dose 28-day oral toxicity study in rodents. https://doi.org/10.1787/9789264070684-en.

Sievert, C. (2018). plotly for R. https://plotly-r.com.

Signorell, A., and al., e.m. (2019). DescTools: Tools for descriptive statistics. R package version 0.99.28. https://cran.r-project.org/package=DescTools.

Stokes, A. H., Kemp, D. C., Faiola, B., Jordan, H. L., Merrill, C. L., Hailey, J. R., et al. (2013). Effects of Solutol (Kolliphor) and cremophor in polyethylene glycol 400 vehicle formulations in Sprague-Dawley rats and beagle dogs. Int. J. Toxicol. 32 (3), 189–197. https://doi.org/10.1177/1091581813485452.

Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Winston, C. (2019). webshot: Take Screenshots of Web Pages. R package version 0.5.2. https://CRAN.R-project.org/package=webshot.

Wolford, S., Schroer, R., Gallo, P., Gohs, F., Brodeck, M., Falk, H., et al. (1987). Age related changes in serum chemistry and hematology values in normal Sprague-Dawley rats. Fundam. Appl. Toxicol. 8 (1), 80–88. https://doi.org/10.1016/0272-0590(87)90102-3.

3.2 The Road to Virtual Control Groups and the Importance of proper Body-Weight Selection

Beyond the control of study design parameters, as presented in the previous chapter, the VCG performance greatly relies on matching HCD with the animals from the legacy study. An essential value for HCD matching is the age of the test subjects since numerous animal values are influenced by age. This short communication presents the use of initial body weight for HCD matching as a surrogate for age. Using the reproducibility of statistical outcomes for animal body weights as an example, this study demonstrates that aligning HCD to the initial body weights of legacy-study animals improves the performance of VCGs. Furthermore, this study presents alternative sampling approaches to improve the VCG performance and discusses the requirements HCD needs to meet for VCG creation along with the suitability of the presented sampling methods.

Authors: A. Gurjanov, L. A. I. Vaas, T. Steger-Hartmann

CRediT author statement: *Conceptualization*: AG, LV, TSH; *Methodology*: AG, LV, TSH; Software: AG; *Validation*: AG, LV, TSH; *Formal analysis*: AG; *Investigation*: AG; *Resources*: AG; *Data curation*: AG; *Writing – original draft*: AG; *Writing – review and editing*: AG, LV, TSH; *Visualization*: AG; *Supervision*: LV, TSH; *Project administration*: TSH; *Funding acquisition*: TSH

This article was submitted to the journal *ALTEX – Alternatives to animal experimentation* on March 14, 2024.

Citation: Gurjanov, A., Vaas, L. A. I. and Steger-Hartmann, T. (2024) “The road to virtual control groups and the importance of proper body weight selection”, ALTEX - Alternatives to animal experimentation, 41(4), pp. 660–665. doi: https://doi.org/10.14573/altex.2403141.

Short Communication

The Road to Virtual Control Groups and the Importance of proper Body-Weight Selection

Alexander Gurjanov[1], Lea Vaas[2], Thomas Steger-Hartmann[1]

[1]Bayer Research & Development, Pharmaceuticals, Investigative Toxicology, Berlin, Germany;

[2]Bayer Research & Development, Pharmaceuticals, Research & Pre-Clinical Statistics Group, Berlin, Germany

Abstract

Virtual control groups (VCGs) created from historical control data (HCD) can reduce the number of concurrent control group animals needed in regulatory toxicity studies by up to 25%. This study investigates the performance of VCGs on statistical outcomes of body weight development between treatment and control groups in legacy studies. The objective is to reproduce the statistical outcomes of 28-day sub-chronic studies (legacy studies) after replacing the concurrent control group with virtual ones. In rodent toxicity studies initial body weight is used as surrogate for the age of animals. For the assessment of VCG-sampling methods three different approaches are explored: (i) sampling VCGs from the entire HCD ignoring initial body weight information of the legacy study, (ii) sampling from HCD matching the legacy study's initial body weights, and (iii) sampling from HCD with assigned statistical weights derived from legacy study initial body weight information. It is shown that the ability to reproduce statistical outcomes by virtual controls is mainly determined by the congruence between the legacy study and the HCD weight distribution: regardless of the chosen approach, the ability to reproduce statistical outcomes was well for VCGs when the legacy study's initial-body-weight distribution was similar to the HCD's. When the initial body weight range of the legacy study was at the extreme ends of the HCD's distribution, the weighted-sampling approach was superior. This article highlights the importance of proper HCD-matching by the legacy study's initial body weight and discusses required conditions to accurately reproduce body weight development.

Introduction

The concept of virtual control groups (VCGs), i.e., using historical control data (HCD) to replace concurrent control group animals (CCGs) with virtual ones (Steger-Hartmann et al., 2020) suggests that VCGs may be able to reduce the number of animals in toxicity studies by up to 25% and thus contributing to the 3R concept (Russel and Burch, 1959).

HCDs are generally used for comparative purposes in regulatory toxicity studies (Kluxen et al., 2021) provided that HCD are “originating from the same laboratory, species, strain, and collected under similar conditions” (OECD, 2018). The same requirements apply for the establishment of reliable, well-performing VCGs (Steger-Hartmann et al. 2020; Gurjanov et al., 2023).28-day repeat-dose toxicity studies in rodents and non-rodents usually determine the safe starting dose for the First-in-Man trials (ICH, 2009; OECD, 2008). HCD was collected from these studies (Kluxen et al., 2021), where the largest amount of HCD has accumulated for rodents, since the rat is the most frequently used species (Namdari et al., 2021). Due to the abundance of HCD, first insights on the use of VCGs have been made with this species (Gurjanov et al., 2023; Wright et al., 2023, Gurjanov et al., 2024).

The ongoing maturation of the concept of VCGs (Golden et al., 2023; Steger-Hartmann and Clark, 2023) involves the evaluation of the VCG performance on legacy studies using them as a benchmark: a virtual control group can be considered “well performing” if the legacy study results can be reproduced after the CCG was replaced by VCGs (Gurjanov et al., 2023).

In conventional 28-day repeat-dose toxicity studies, roughly 100 different parameters are measured in each animal comprising hematological parameters, clinical chemistry, urine measurements, clinical observations, and histopathological findings. At the beginning of the experiment, animals are randomly allocated to a control group and several dose groups, where the latter receive the test item. The parameter for randomization in rodent studies is usually the initial body weight (Hoffmann et al, 2002). If an effect in animals is observed, one can compare it to the control to assess whether this effect is caused by the test item or by other influences (OECD, 2008; ICH, 2009). Differences in quantitative parameters, such as body weight or clinical chemistry parameters are hereby assessed using appropriate statistical tests (Hamada, 2018) such as the Dunnett’s test (Dunnett, 1955; Hothorn, 2016).

A key parameter in toxicological assessment is animal body weight. It is the only quantitative parameter measured daily during an ongoing dosing period (i.e., in-life phase) making body weight the most densely populated datapoint. Furthermore, body weight serves as a stratum for randomizing animals into the different experimental groups before the start of the treatment (Du Sert et al., 2020). Together with clinical observations, body weight is used to monitor *ad hoc* toxic effects and animal wellbeing (Jourdan, 2013); a critical decrease in body weight is

even a criterion to terminate a study entirely (Talbot et al., 2020). Furthermore, body weight is an important parameter in animal purchase, where it may be used as a surrogate for age or date of birth.

Given the fact that many physiological parameters correlate with age and body weight of animals, any mismatch in the initial body weight (i.e., body weight taken at the beginning of the experiment prior to first test-item application) might impair the outcome of a study using VCGs (Wolford et al., 1987; McCutcheon and Marinelli, 2009, Jacob Filho et al., 2018).

This article presents the performance of VCGs regarding body weight values of 14 legacy studies. The statistical outcomes of the legacy studies ("significantly different to the control" vs. "non-significantly different to the control") serve as a benchmark and the objective is to reproduce them after replacing CCGs with VCGs. This study examines the impact of different sampling methodologies on the VCG performance, discusses potential benefits and pitfalls of the approaches, and explores the necessity for careful selection and validation of HCD prior to VCG creation.

Methods

Software

For data gathering, statistical calculations, model development, evaluation, and visualization the statistical programming software R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria) was used. The details of R packages used are listed in Table S1 of the supplementary material. The code used and the control data are stored in Bayer's GitHub repository (Gurjanov, 2024).

Graphical representation

All data was visualized as histograms to check for normality of value distribution. The growth of the animals is shown as a line plot where mean body weight values are displayed with respect to the study day; the figures for males and females respectively can be found in Figures S4-S31 in the supplementary material. The initial body weight, *i.e.,* the body weight measured on day 1 prior to the first application of the test substance is shown as box plots with the body weight values for each group (control and dose groups); the respective plots are shown in the Figures S32-S59 in the supplementary material.

Selection of historical control data

Data was selected as described in Gurjanov et al. (2023) and extracted from Bayer's internal repository. From a total of 1609 rat studies, 14 studies with similar criteria regarding species, strain, route of administration, treatment vehicle, number of treatment groups, animal supplier, food and water supply, caging conditions, and study length were used for this report. To avoid potential impact of genetic drift, studies were selected from within a 5-year timeframe. All selected studies were performed at Bayer's test facility in Wuppertal, Germany between 2017 and 2022 according to European and national animal protection regulations. The details for selection and data filtering can be found in chapter 1.2 of the supplementary material.

Studies were numbered according to ascending initial body weight of male animals (Study-01 to Study-14), i.e., Study-01 having animals with the lowest mean initial body weight and Study-14 having animals with the highest. The studies' initial body weights and mean weight gains are shown in Table 1.

Table 1: Virtual control group (VCG) performance of body weight reproducibility of legacy studies for male and female animals. VCGs were generated from historical control data (HCD) followed by statistical reanalysis of body weight values between VCG and dose groups across all time points (day 1-28). The sampling was repeated 1000 times and the subsequent performance is the percentage of cases where VCGs reproduced the original statistical outcomes (significant vs. non-significant) of the legacy study with the concurrent control. VCGs were created from the HCD in three approaches, (i) sampled from HCD not matched to legacy study dose groups' weight distribution, (ii) sampled from HCD matched to legacy study dose groups' initial body weight, (iii) weighted sampling with weights assigned to HCD to match the probability density of legacy study dose groups.

	VCG performance in males [%]			**VCG performance in females [%]**		
Legacy study number	Sampling on un-filtered HCD	Sampling on HCD matched to legacy study's dose-groups' initial body weight	Weighted sampling	Sampling on un-filtered HCD	Sampling on HCD matched to legacy study's dose-groups' initial body weight	Weighted sampling
Study-01	15	0	45	14	75	88
Study-02	27	77	79	31	72	71
Study-03	66	69	76	72	78	79
Study-04	83	77	87	48	97	94
Study-05	90	89	91	36	69	89
Study-06	63	40	64	78	90	89
Study-07	99	99	99	93	95	97
Study-08	83	88	91	39	77	79
Study-09	66	87	83	8	60	87
Study-10	96	99	98	95	92	92
Study-11	71	83	81	81	82	77
Study-12	74	91	91	94	97	97
Study-13	87	94	93	93	97	93
Study-14	36	55	56	94	94	95
Mean	**68**	**75**	**81**	**63**	**84**	**88**

Each study consists of a concurrent control group and three treatment groups (later on called dose groups: low dose (LD), mid dose (MD), and high dose (HD)). Body weight was measured prior to application of the test item daily throughout the study from day 1 to day 28. To assess the differences between control and treatment groups, a Dunnett's exact homogenous test (Dunnett, 1955) was computed for each day, and the resulting *p*-values serve as classifiers into "significantly different to the control" ($p \leq 0.05$) and "non-significantly different to the control" ($p > 0.05$). In the subsequent sections of this article, these terms will be abbreviated as "significant" and "non-significant".

Creation and evaluation of virtual control groups

From the selected 14 studies, the HCD consisted of 165 male and 157 female control group animals. VCGs representing a drawn sample of the HCD pool were created using a resampling approach introduced by Steger-Hartmann et al. (2020) and further explored in recent publications (Gurjanov et al., 2023; Gurjanov et al., 2024). The underlying idea is that forming VCGs by drawing animals from the HCD pool ensures that all body weight measurements over the 28-day period are assigned to the respective subject. VCGs substituted then the CCG in the legacy study whereas the number of animals drawn is equal to the number CCG animals originally used in the respective legacy study. The objective of this study is to assess how well the replacement of the CCG with virtual controls reproduces the original statistical outcomes (significant and non-significant). The entire approach was performed iteratively for all studies with each single study serving as a legacy study once, while the control data of the remaining studies were used as HCD for generating corresponding VCGs. VCGs were created from HCD in three different approaches:

Approach 1: VCGs were created by randomly sampling animals without replacement from the respective HCD animal pool. The respective legacy study's initial body weight information was not considered for the sampling. A representative sampling is given in Figure 1A. The sampled animals and their respective body weight values replaced the CCGs in the legacy study throughout the entire time-course of the study.

Approach 2: the initial body weight value ranges of the legacy study's dose groups are extracted and used to match the HCD, i.e., HCD animals not falling within the range of dose groups' initial body weights were discarded. A representative example is shown in Figure 1B. The black dashed lines illustrate the weight ranges used for matching the HCD. Animals were then randomly sampled without replacement from this initial-body-weight matched set of HCD.

Approach 3: aiming for a selection of animals in a way that virtual controls reproduced the initial body weight distribution in the legacy study, a weighted sampling method was implemented assigning statistical weights to the HCD animal pool. This increased the

probability of selecting animals that matched the dose groups' distribution. To form this distribution, a kernel density estimation (KDE), i.e. a non-parametric method for estimating a density curve representing the probability distribution, was computed for the initial body weights of legacy study's treatment-naïve dose groups. As an output, each initial body weight value of HCD was assigned with a numerical value indicating frequency of occurrence within the distribution estimated by the KDE. This information was then used to assign statistical weights to each initial body weight value of the HCD corresponding to the probability density on the dose groups' KDE curve. Afterwards, a weighted sampling was conducted on the HCD where animals within the dose groups' initial body weight distribution were more likely to be drawn than animals outside. This approach allowed for sampling with replacement, consequently, one animal may be represented several times in a VCG. This approach is shown exemplary for one iteration in Figure 1C where the black line illustrates the original probability distribution of the HCD, and the black line shows the newly assigned weighted probability distribution used for weighted sampling.

After sampling the animals, Dunnett's tests were re-calculated, and the test results were compared to the original result. Each sampling and subsequent statistical evaluation was repeated 1000 times independently. The resulting performance is expressed as the percentage of reproduced original statistical outcomes, summarized across all time points, and dose groups over all iterations.

Results

Mean body weight development

Rat (males and females, respectively) mean body weight development for each study are illustrated in Figures S4-S31 of the supplementary material. The mean initial body weights and weight gains per sex are shown in Table S2 of the supplementary material. The studies are sorted by ascending mean initial body weight of male rats. The mean body weight gain correlates negatively with the mean initial body weight, i.e., a low initial body weight leads to a high weight gain within the observed timeframe of 28 days (males: r = -.085, p = 1.08E-4; females: r = -.082, p = 2.54E-4)). Two studies are present (Study-01 and Study-02) where the initial body weight of males is markedly lower than in the remaining studies and, subsequently, their weight gain is respectively higher.

Finally, a noteworthy observation is the marked decrease in mean body weight which may occur in the last week of dosing (day 22-26), followed by strong increases on the following day. Four studies with such prominent transient decreases are highlighted in Figure S2 in the supplementary material. This phenomenon is caused by urine measurement procedures:

animals were placed in metabolic cages and fasted to ensure that feces do not contaminate the collected urine.

Resampling process

After replacing the CCG of a legacy study with VCGs and subsequent statistical reanalysis, the percentage of all reproduced statistical outcomes (significant vs. non-significant) for body weight out of 1000 iterations was calculated. Table 1 summarizes the performance-evaluation of VCGs generated by the three tested approaches.

The first approach generates VCGs by randomly sampling animals from the entire HCD without taking initial body weight into consideration and resulted in a rather low performance. Only about 65% (68% males, 63% females) of all statistical outcomes of body weights from day 1 to 28 were reproduced by VCGs. In only 6 out of 14 studies a consistent reproducibility of 80% or higher could be achieved. In Study-01, only 15% of the results for males and in Study-09, only 8% of all results for females were reproducible. The reason for the poor reproducibility for Study-01 is illustrated in supplementary Figure S4: the body weight values of the CCG are far outside of the 2 x standard-deviation range of the HCD which in turn is used to create VCGs. Subsequentially, a high difference between VCG and dose-group was classified as significant while there was no statistically significant difference between CCG and dose groups. Hence, the original results were not reproduced. Only in the last days of the study the CCG was close to the HCD so that the original results could be reproduced consistently. Figure 1A shows the VCG generation process on another study (Study-05 in females): the initial body weight of the CCG and dose groups (white and red boxes) are high in comparison to the weight distribution of the female HCD (density plot on the left). A virtual control (grey box in the box plot), consisting of 10 randomly sampled animals from this HCD distribution is therefore more likely to be lower than the CCG and a statistical reanalysis will lead to results inconsistent to the original one.

Approach 2 improved the VCG performance by filtering the HCD so that it matches the initial body weight values of the respective legacy-study dose-group animals. Thus, the reproducibility of statistical outcomes increased compared to the first approach with 79% (75% for males and 84% for females) of all statistical outcomes being reproducible. Further, the reproducibility was above 80% for 8 out of 14 studies. However, this approach had shortcomings: a prerequisite for sampling animals without replacement is that the HCD pool has at least the same number of animals as the legacy study. This could not be achieved for the male group in Study-01 where the initial body weight was outside of the HCD distribution. Further, as shown in Figure 1B on the example of females in Study-05, the initial body weight values of the dose groups have been used as filter ranges to remove body weight values of

HCD outside of these borders (dashed lines). This, however, resulted in HCD with body weight distribution heavily skewed to the left side. Thus, derived VCGs are more likely to be at the lower end within the body weight boundaries, which still results in many inconsistencies to the original results.

Approach 3: the above limitations could be addressed by conceptualizing a weighted-sampling approach (shown in Figure 1C) where the kernel density of the dose-group body weights (red line) was mapped to the density distribution of the HCD. This increases the probability of generating VCGs with a similar weight distribution as the respective legacy-study's dose group animals. The weighted-sampling approach improved the reproducibility of the original statistical outcomes to 85% in all cases (81% in males, 88% in females). Out of 14 studies, 9 studies with males and 10 with females showed a consistency above 80% regarding their statistical outcomes. In Study-01 in males—the study with the animals having very low initial body weights—45% consistency was reached.

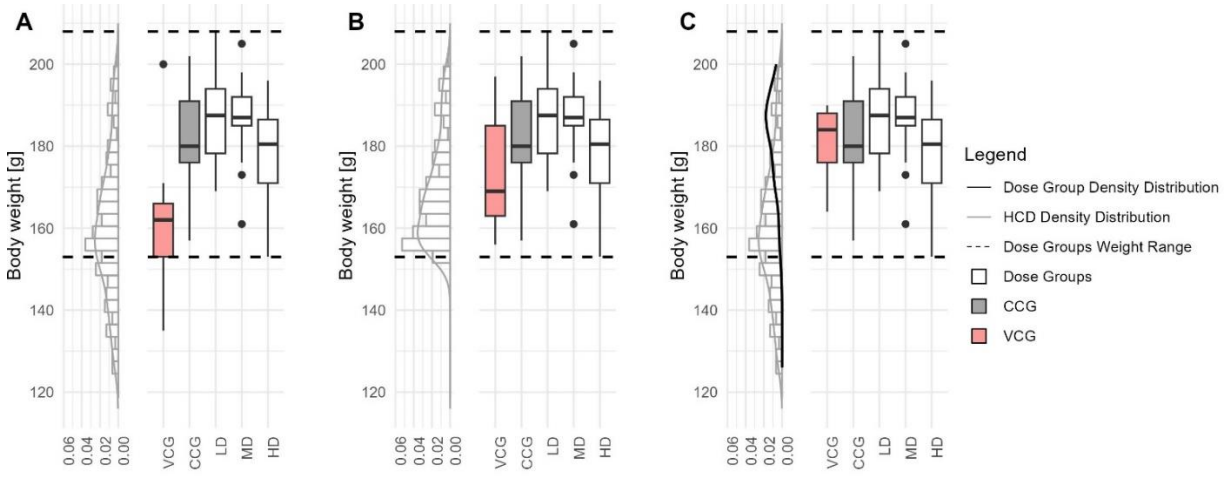

Figure 1: Exemplary virtual control groups for female animals of Study-05. Box plots show the VCG (red box) and the legacy study's concurrent control group (CCG) (grey box) and dose groups (low dose (LD), mid dose (MD), and high dose (HD)) (white boxes). The black dashed lines represent the ranges (min and max body weight) of the dose groups. Adjacent on the left, the histogram shows the probability density distribution of historical control data (HCD) pool used to generate virtual control groups (VCGs): **1A)** VCG was sampled from HCD not matched to legacy study weight distribution. **1B)** VCG was sampled from HCD matched to legacy-study-dose groups' initial body weight range. **1C)** VCG generated by weighted sampling with weights assigned to HCD to match the probability density of legacy study dose groups (black solid line on the histogram).

Discussion

This article presents the performance of virtual control groups assessed on their ability to reproduce statistical outcomes of a legacy study's body weight values. Body weight is of particular interest in the context of VCGs since it serves as a surrogate for age for HCD selection particularly for rodent species (McCutcheon and Marinelli, 2009). Although matching by age would be preferrable as age is the most influencing factor for all animal physiology, animal birth dates in rodents may be missing or documented imprecisely (cross-company study directors, personal communication, December 2023). This study shows that statistical outcomes comparing body weight parameters can be reproduced with virtual control groups at an average rate of 66%–85%, depending on the sampling approach. Establishing reliable and well performing VCGs requires that certain conditions are fulfilled:

1) Historical control data should be selected based on study-design parameters of a planned study.

It was discussed in previous works (Gurjanov et al., 2023; Golden et al., 2023) that HCD should consist of animals from studies performed under conditions closely comparable to the concurrent study. Animal data from different strains, or study-design parameters such as sex, route of administration or study length should not be mixed as this would increase the experimental variability of endpoints decreasing sensitivity towards a test-substance (Howard, 2002).

2) The resulting HCD pool should have a substantial number of animals for proper resampling.

If the number of available HCD animals is smaller than the CCG size, the resulting VCGs become biased leading to limitations in the reproducibility of the original study design. While technically a statistical evaluation with a Dunnett's test is possible with only two animals per group, a small group size limited to certain values leads to biased variance estimators and lower statistical power and thus, less robust results.

3) The distribution of the legacy-study's initial body weights should ideally be placed well within the HCD distribution.

Even when study-design parameters are similar, VCGs should only be generated from HCD which was matched to the initial body weight distribution of the study. Our results show that the VCG performance was below 60% if the CCG body weight mean value was outside the 1 x standard deviation range of the HCD, while sampling from an initial-body-weight matched HCD led to superior performance. For legacy studies, where initial body weights were are at

the far ends of the HCD distribution (Study-01, 02, and 14 in males, and Study-01, 05, 08, and 09 in females), VCGs performed best when generated from HCD in a weighted-sampling approach.

However, the weighted-sampling approach should be taken with caution: assigning weights to a resampling process may lead to bias as (i) the VCGs may become unrepresentative of the true population, and (ii) individual animals may be over-represented resulting in overfitting phenomena. For instance, one iteration in males of Study-01 resulted in an animal represented 5 times. On the other hand, a weighted-sampling approach may be suitable for cases where the initial-body-weight distributions of the HCD and the studies are largely different, since this approach creates VCGs similar to the legacy study's initial body weights' distribution (see Figure 1C). This is not achievable through initial-body-weight matching (Figure 1B). Future research should seek to establish clear decision criteria that can suggest the most appropriate sampling methodology for specific situations.

It should be noted that the reproducibility of statistical outcomes here is made exclusively on body weight values. While our study shows that VCGs can be generally used to reproduce the original body-weight statistical outcomes, reproducibility of the entire studies comprising all quantitative and qualitative parameters needs to be investigated as recently done by Gurjanov et al. (2024). However, given the fact that many physiological parameters are correlated with body weight (Jacob Filho et al., 2018), a thorough understanding of body weight matching is of key importance for the further assessment of the VCG concept.

Acknowledgements

The authors are deeply grateful for the work of Annika Kreuchwig, Adam Zalewski, and Carlos Vieira-Vieira from Bayer's department of computational toxicology for setup and maintenance of the database from which the data used in this publication has been sourced.

Conflict of interest

Part of the described research has been performed under the Innovative Medicine Initiative (IMI) Enhancing TRANslational SAFEty Assessment through Integrative Knowledge Management, (eTRANSAFE) project. eTRANSAFE has received support from IMI2 Joint Undertaking under Grant Agreement No. 777365. This Joint Undertaking received support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations (EFPIA).

Data availability statement

The historical control data and code be found on GitHub (https://github.com/Bayer-Group/VCG-INITBW). Due to confidentiality and propriety reasons, individual dose-group data cannot be shared.

References

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association 50*, 1096-1121. https://doi.org/10.1080/01621459.1955.10501294.

Du Sert, N. P., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., ... & Würbel, H. (2020). Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLoS biology, 18(7), e3000411. https://doi.org/10.1371/journal.pbio.3000411.

Golden, E., Allen, D., Amberg, A. et al. (2023). Toward implementing virtual control groups in nonclinical safety studies: Workshop report and roadmap to implementation. *ALTEX-Alternatives to animal experimentation.* https://doi.org/10.14573/altex.2310041.

Gurjanov, A., Kreuchwig, A., Steger-Hartmann, T. et al. (2023). Hurdles and signposts on the road to virtual control groups—a case study illustrating the influence of anesthesia protocols on electrolyte levels in rats. *Frontiers in Pharmacology 14*, 1142534. https://doi.org/10.3389/fphar.2023.1142534.

Gurjanov, A. (2024) VCG Initial Body Weight [Software]. GitHub. https://github.com/Bayer-Group/VCG-INITBW.

Gurjanov, A., Vieira-Vieira, C., Vienenkoetter, J. et al. (2024). Replacing concurrent controls with virtual control groups in rat toxicity studies. Regulatory Toxicology and Pharmacology, 105592. https://doi.org/10.1016/j.yrtph.2024.105592.

Hamada, C. (2018). Statistical analysis for toxicity studies. *Journal of toxicologic pathology 31*, 15-22. https://doi.org/10.1293/tox.2017-0050.

Hoffman, W. P., Ness, D. K., van Lier, R. B. L.. (2002). Analysis of Rodent Growth Data in Toxicology Studies, *Toxicological Sciences, 66, 2*, 313–319, https://doi.org/10.1093/toxsci/66.2.313.

Hothorn, L. A. (2016). The two-step approach—a significant anova f-test before dunnett's comparisons against a control—is not recommended. *Communications in Statistics-Theory and Methods 45*, 3332-3343. https://doi.org/10.1080/03610926.2014.902225.

ICH (2009). Guidance on nonclinical safety studies for the conduct of human clinical trials and marketing authorization for pharmaceuticals m3 (r2). International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-m3r2-non-clinical-safety-studies-conduct-human-clinical-trials-and-marketing-authorisation-pharmaceuticals-step-5_en.pdf.

Howard, B. R. (2002). Control of variability. ILAR journal, 43(4), 194-201. https://doi.org/10.1093/ilar.43.4.194.

Jacob Filho, W., Lima, C. C., Paunksnis, M. R. R. et al. (2018). Reference database of hematological parameters for growing and aging rats. *The Aging Male 21*, 145-148. https://doi.org/10.1080/13685538.2017.1350156.

Kluxen, F. M., Weber, K., Strupp, C. et al. (2021). Using historical control data in bioassays for regulatory toxicology. *Regulatory Toxicology and Pharmacology 125*, 105024. https://doi.org/10.1016/j.yrtph.2021.105024.

McCutcheon, J. E. and Marinelli, M. (2009). Age matters. *European Journal of Neuroscience 29*, 997-1014. https://doi.org/10.1111/j.1460-9568.2009.06648.x.

Namdari, R., Jones, K., Chuang, S. S. et al. (2021). Species selection for nonclinical safety assessment of drug candidates: Examples of current industry practice. *Regulatory Toxicology and Pharmacology 126*, 105029. https://doi.org/10.1016/j.yrtph.2021.105029.

OECD (2008). *Test no. 407: Repeated dose 28-day oral toxicity study in rodents*. Vol. https://www.oecd-ilibrary.org/content/publication/9789264070684-en doi:doi: https://doi.org/10.1787/9789264070684-en.

OECD (2018). Test No. 408: Repeated Dose 90-Day Oral Toxicity Study in Rodents. https://doi.org/10.1787/20745788.

Russel, W. M. S. and Burch, R. L. . The principles of humane experimental technique. Methuen, 1959. http://117.239.25.194:7000/jspui/bitstream/123456789/1342/1/PRILIMINERY%20%20AND%20%20CONTENTS.pdf.

Steger-Hartmann, T., Kreuchwig, A., Vaas, L. et al. (2020). Introducing the concept of virtual control groups into preclinical toxicology testing. *ALTEX-Alternatives to animal experimentation 37*, 343-349. https://doi.org/10.14573/altex.2001311.

Steger-Hartmann, T. and Clark, M. (2023). Can historical control group data be used to replace concurrent controls in animal studies? *Toxicologic Pathology* 01926233231208987. https://doi.org/10.1177/01926233231208987.

Talbot, S. R., Biernot, S., Bleich, A. et al. (2020). Defining body-weight reduction as a humane endpoint: A critical appraisal. *Laboratory animals 54*, 99-110. https://doi.org/10.1177/0023677219883319.

Wolford, S., Schroer, R., Gallo, P. et al. (1987). Age-related changes in serum chemistry and hematology values in normal sprague-dawley rats. *Fundamental and Applied Toxicology 8*, 80-88. https://doi.org/10.1016/0272-0590(87)90102-3.

Wright, P. S., Smith, G. F., Briggs, K. A. et al. (2023). Retrospective analysis of the potential use of virtual control groups in preclinical toxicity assessment using the etox database. *Regulatory Toxicology and Pharmacology 138*, 105309. https://doi.org/10.1016/j.yrtph.2022.105309.

3.2 Supplementary Material

Table of Contents

1 Methods

1.1 Software

Data gathering, statistical calculations, model development, evaluation, and visualization was performed using the statistical programming software R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria). The details of R packages used are listed in Table below. The R code can be accessed by Bayer's open-source GitHub repository (Gurjanov, 2024b)

Table S1: R packages used in the VCG qualification procedure.

Package name	**Usage**
data.table (Dowle and Srinivasan, 2023)	Fast loading of large tables
parallel (R Core Team, 2021)	Divide calculations on several cores for faster computing
tidyverse_2.0.0 (Wickham, 2017)	Efficient processing and visualizing of data
openxslx_4.2.5.2 (Schauberger and Walker, 2023)	Reading and writing of Excel (XLSX) sheets
PMCRplus_1.9.6 (Pohlert, 2022)	Pipe friendly computing of significance tests
cowplot_1.1.1 (Wilke, 2020)	Combining several plots together

1.2 Selection of legacy studies

All animal studies were conducted internally and recorded using Pristima's laboratory information management system (LIMS) (Xybion, 2024). Harmonization of the study data was performed according to the standard for exchange of nonclinical data (SEND) (CDISC, 2022) controlled terminology for enabling data analysis using Amazon Warehouse System (AWS) S3 solution. The selection criteria for the three legacy studies used for the qualification procedure were based on the following study design parameters:

- Dosing duration of 4 weeks.
- Studies were initiated between 2017 or 2022.
- A control group was used along with three dose groups: low dose (LD), mid dose (MD), and high dose (HD).
- Studies were conducted in the laboratory of Bayer Pharmaceuticals AG, Wuppertal, Germany.
- Study was performed with Wistar HAN rats.

- Route of administration was oral by gavage.
- The studies had been performed at the same test facility.
- Treatment vehicle contained either Kolliphor® HS 15 or polyethylene glycol 400.
- Animals were supplied by the same breeder.
- Food and water supply during the study was *ad libitum*.
- Animals were housed in group cages with 2-3 animals per cage.
- Body weight was measured on day 1-28.

1.2.1 Study attrition after filtering

From a total of 1602 rat studies, 14 studies remained with a maximum similar study design. The attrition of studies is illustrated in supplementary Figure S1 below.

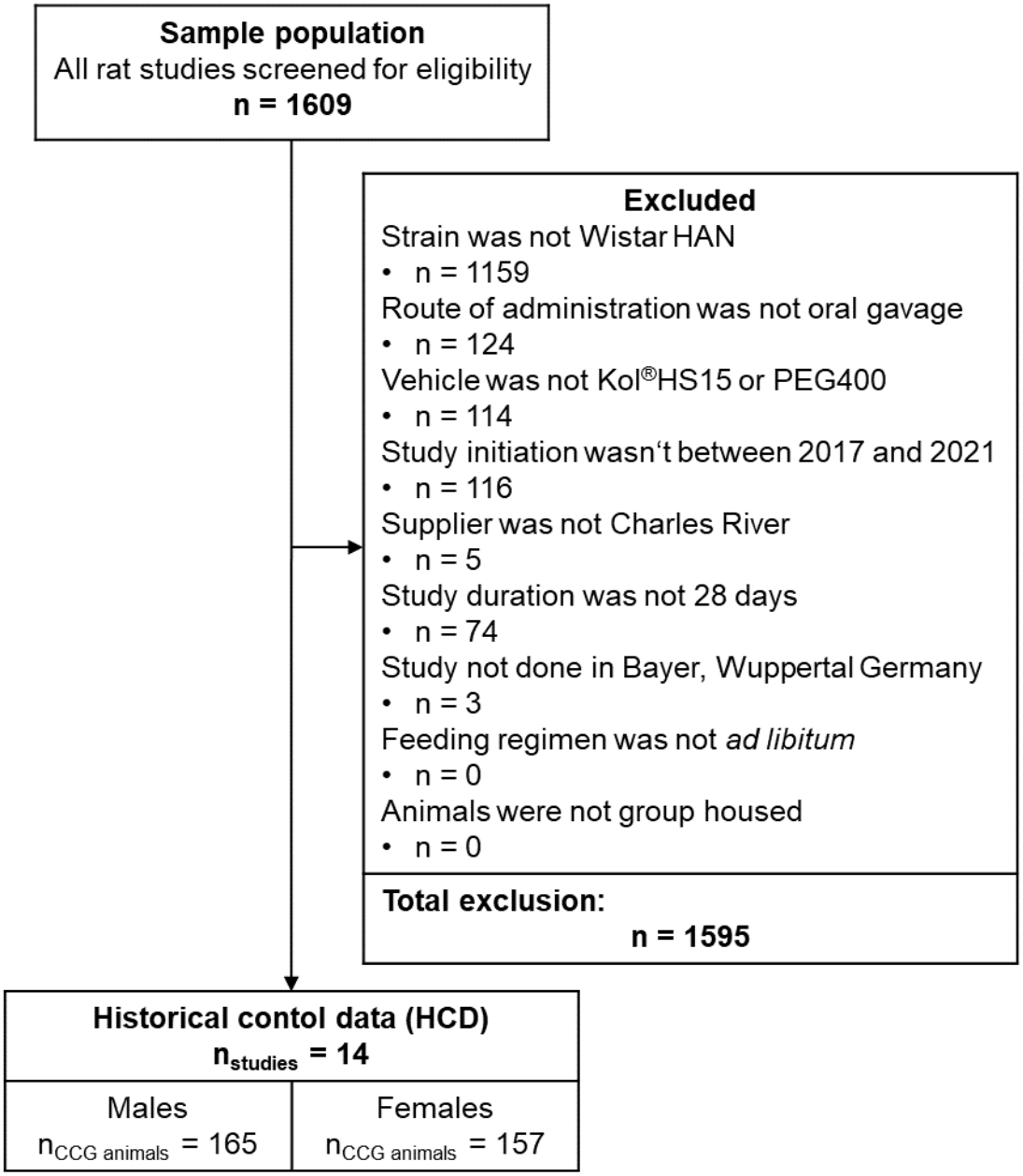

Figure S1: Rat-study selection and attrition.

2 Results

2.1 Mean growth of selected studies

The following Table S2 shows the age at study initiation, initial body weight, and mean weight gain of males and females for Study-01 to 14. The selected studies are sorted by ascending initial body weight of male control group animals. For calculation of the mean weight gain across all studies body weights from days 22-26 were removed to avoid bias due to urine measurements frequently taken in that time. Of note is the difference in mean initial body weights which varies considerably between males and females. In Study 1, there's a difference of 16 g while in Study 14, males and females differ by 72 g in Study 14 in the mean initial body weights. The subsequent figures show the mean body weight per study with respect to time (Figure S2 for males and Figure S3 for females). Four studies are highlighted where urine measurements were taken for males resulting in strong transient body weight declines. Historical control data should be examined for such transient declines and data should be flagged accordingly.

Table S2: Age, mean initial body weight ± SD, and mean weight gain ± SD with relative weight gain [%] ± SD of male and female concurrent control group (CCG) animals for Study-01 to 14. Mean weight gain was cleaned of days where urine measurements were usually taken, i.e., day 22 to 26 were not taken for calculating weight gains.

Study number	Age of animals [weeks]	Initial mean body weight of CCG [g] ± SD		Mean weight gain [g/day] ± SD (relative weight gain [%] ± SD) cleaned of urine-measurement-days	
		Males	Females	Males	Females
Study-01	4-7	157 ± 8	141 ± 9	5.94 ± 2.83 (3.78 ± 1.81)	2.41 ± 5.49 (1.71 ± 3.90)
Study-02	6-7	180 ± 7	137 ± 7	5.21 ± 3.06 (2.89 ± 1.70)	2.73 ± 4.70 (1.99 ± 3.43)
Study-03	7	199 ± 10	165 ± 8	4.21 ± 2.43 (2.12 ± 1.23)	2.03 ± 5.01 (1.23 ± 3.04)
Study-04	7-11	206 ± 6	182 ± 8	4.52 ± 2.44 (2.19 ± 1.19)	1.69 ± 6.27 (0.93 ± 3.45)
Study-05	7-9	206 ± 12	181 ± 11	3.65 ± 2.43 (1.77 ± 1.18)	1.10 ± 4.40 (0.61 ± 2.43)
Study-06	7-8	210 ± 10	155 ± 8	3.45 ± 2.42 (1.64 ± 1.16)	2.41 ± 5.44 (1.55 ± 3.51)
Study-07	7-8	212 ± 9	157 ± 10	4.14 ± 2.70 (1.95 ± 1.28)	1.67 ± 4.03 (1.06 ± 2.57)
Study-08	6-8	213 ± 9	184 ± 10	4.05 ± 2.51 (1.90 ± 1.18)	1.57 ± 6.10 (0.85 ± 3.32)
Study-09	7	214 ± 7	183 ± 11	4.20 ± 2.83 (1.96 ± 1.32)	1.75 ± 6.21 (0.96 ± 3.39)
Study-10	7	216 ± 8	162 ± 7	3.77 ± 2.68 (1.75 ± 1.24)	2.30 ± 4.68 (1.42 ± 2.89)
Study-11	7-8	219 ± 8	160 ± 8	3.92 ± 2.31 (1.79 ± 1.06)	2.35 ± 6.82 (1.47 ± 4.26)
Study-12	7-8	222 ± 10	155 ± 12	4.13 ± 2.45 (1.86 ± 1.11)	2.03 ± 5.88 (1.31 ± 3.79)
Study-13	6	225 ± 7	160 ± 7	4.28 ± 2.76 (1.90 ± 1.23)	2.02 ± 4.44 (1.26 ± 2.78)
Study-14	8	229 ± 12	157 ± 10	3.49 ± 2.39 (1.52 ± 1.05)	2.07 ± 5.98 (1.32 ± 3.81)

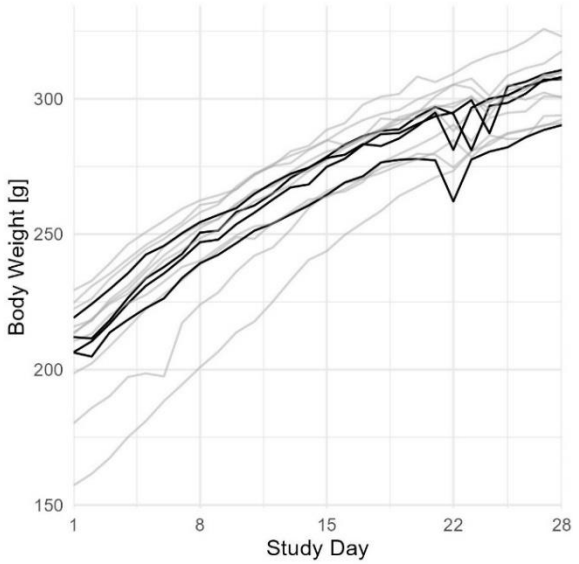

Figure S2: Mean body weight growth of control-group animals for all 14 studies (males). Four studies are highlighted where a transient decrease in body weight was observed. These transient decreases are caused by urine measurements.

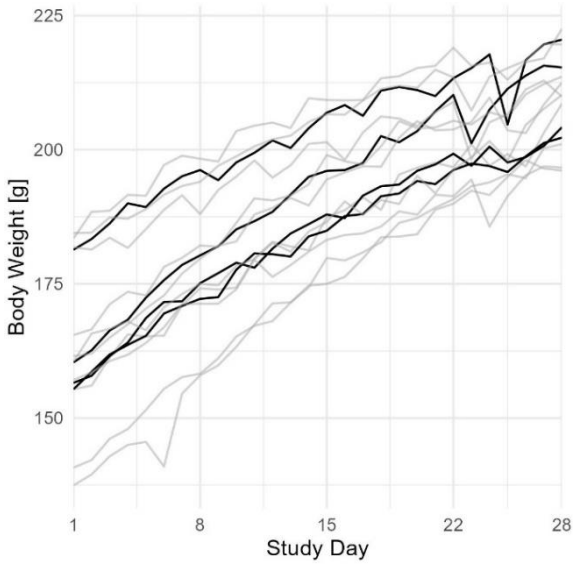

Figure S3: Mean body weight growth of control-group animals for all 14 studies (females). The same studies are highlighted as for the males in Figure S1, where a transient decrease in body weight was observed. Only two studies show the same transient decrease in body weight.

2.2 Mean body weight gain of legacy study control groups compared to mean body weight gain of historical control data

The following 28 figures (Figure S4-S31) show the mean body weight gain of the legacy studies Study-01 to Study-14 of concurrent control group animals, first for males (Figure S4-S17), then fore females (Figure S18-S31). The solid lines represent the mean body weight with respect to the study day and the dashed lines show the 2•standard deviation area. Of note is the mean growth of male animals in Study-01 in males (Figure S4). The mean initial body weight (on day 1) is very low compared to the historical control data while the growth of the animals is high compared to the growth of the historical control data animals. Generating virtual control groups and replacing the concurrent controls resulted therefore in poor reproducibility of subsequent statistical outcomes. Only the body weight values of the last week (day 22 and onwards) were reproduced well by historical control data. As a result, statistical outcomes were reproduced only in 15% of all VCG iterations.

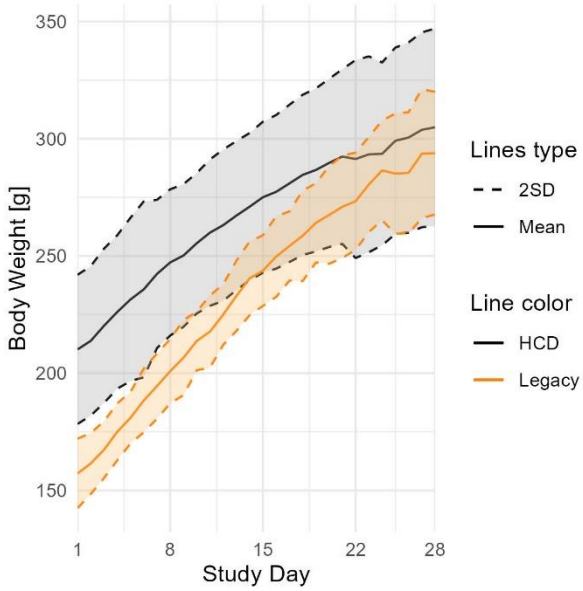

Figure S4: Mean body weight growth of male control group animals of Study-01 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

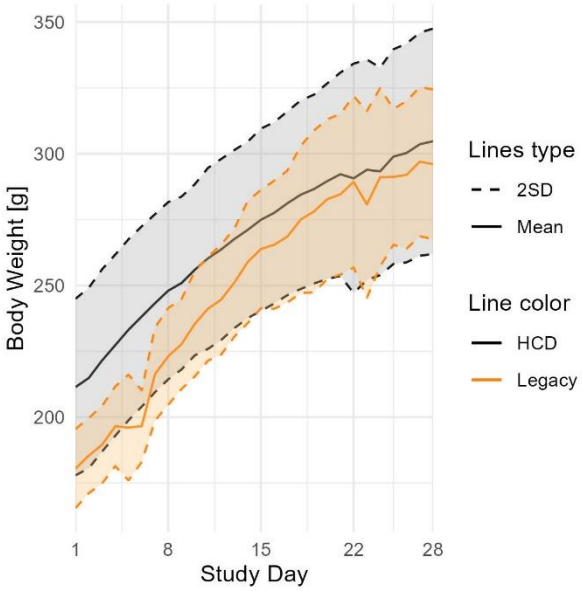

Figure S5: Mean body weight growth of male control group animals of Study-02 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

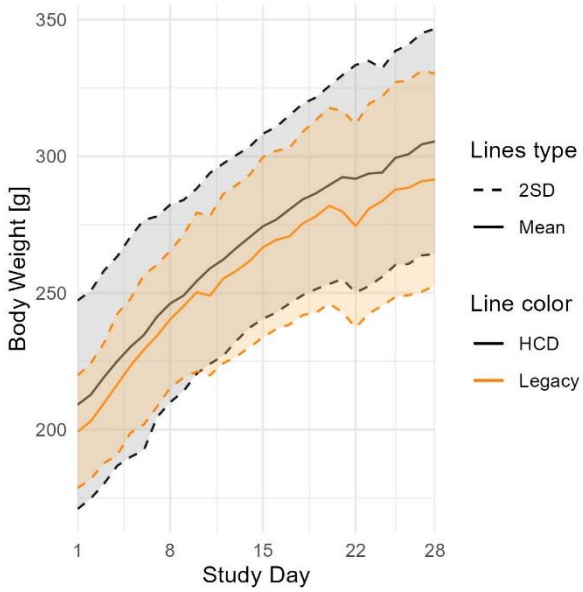

Figure S6: Mean body weight growth of male control group animals of Study-03 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

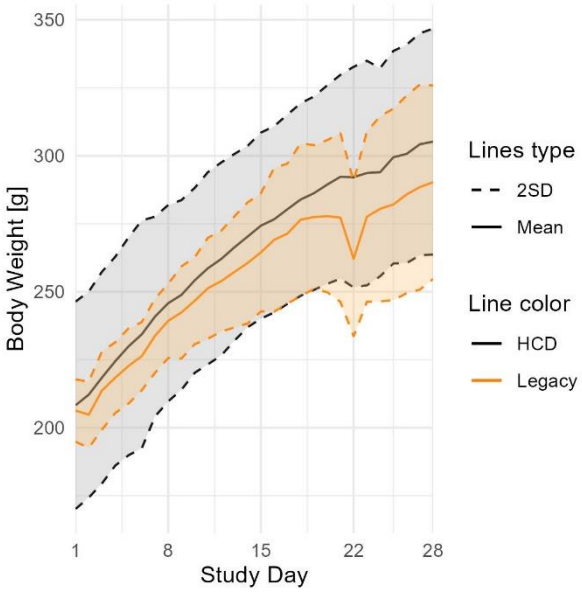

Figure S7: Mean body weight growth of male control group animals of Study-04 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

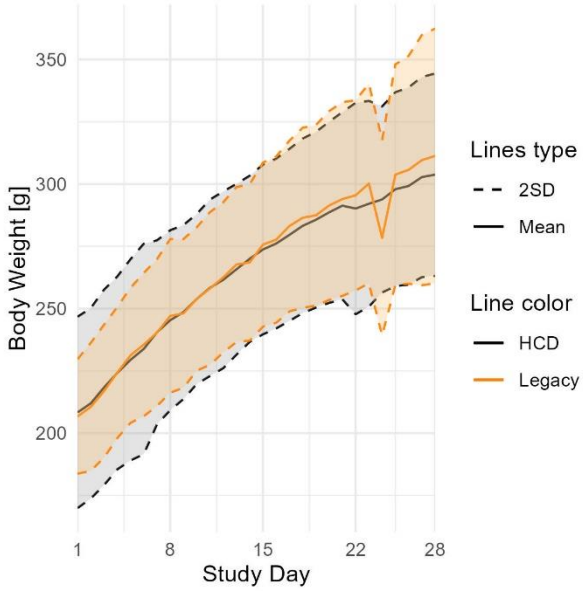

Figure S8: Mean body weight growth of male control group animals of Study-05 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

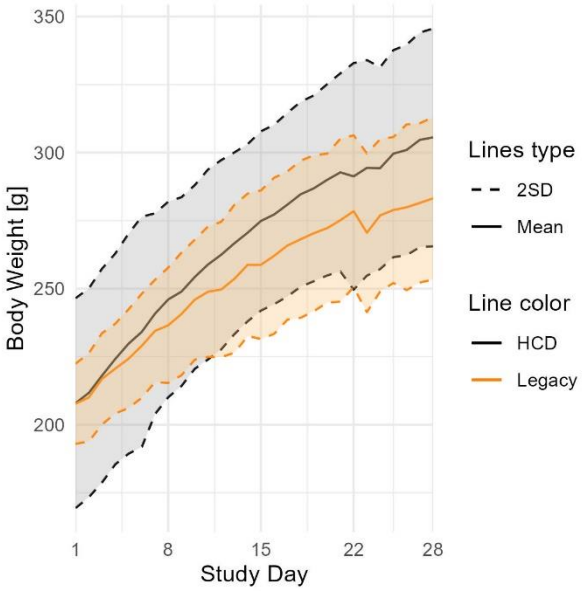

Figure S9: Mean body weight growth of male control group animals of Study-06 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

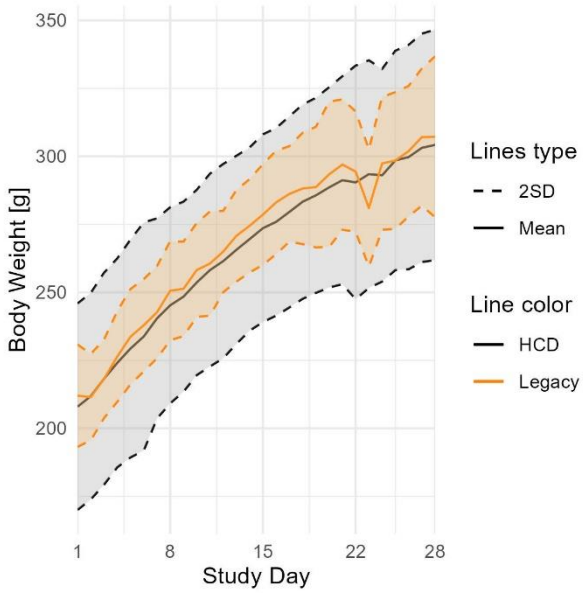

Figure S10: Mean body weight growth of male control group animals of Study-07 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

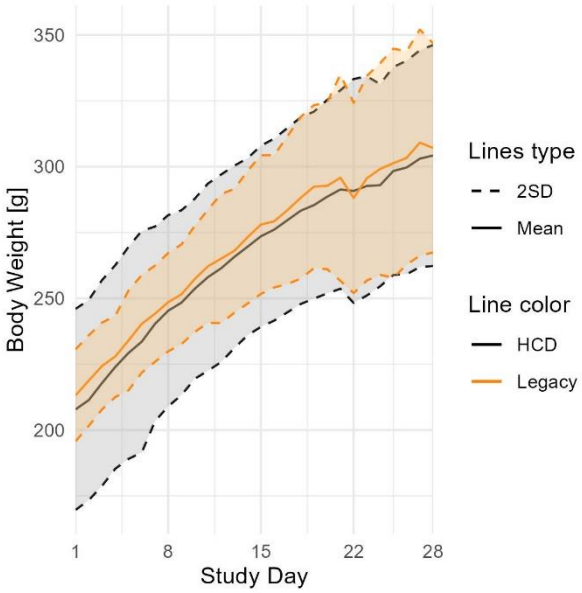

Figure S11: Mean body weight growth of male control group animals of Study-08 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

the 2•sd range of the legacy study and historical control data respectively.

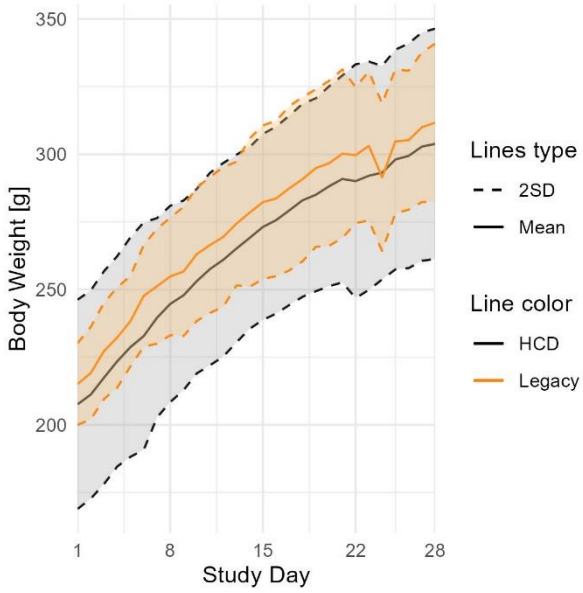

Figure S12: Mean body weight growth of male control group animals of Study-09 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

the 2•sd range of the legacy study and historical control data respectively.

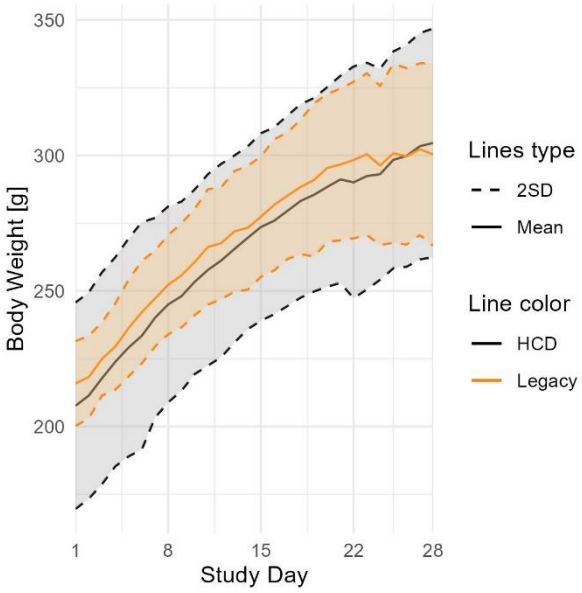

Figure S13: Mean body weight growth of male control group animals of Study-10 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

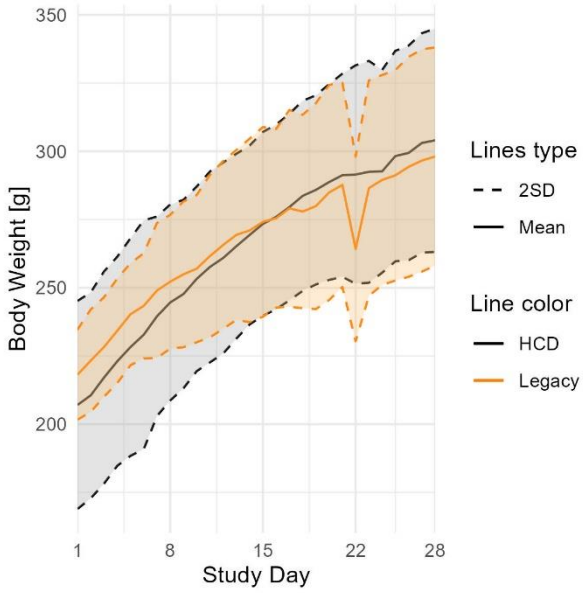

Figure S14: Mean body weight growth of male control group animals of Study-11 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

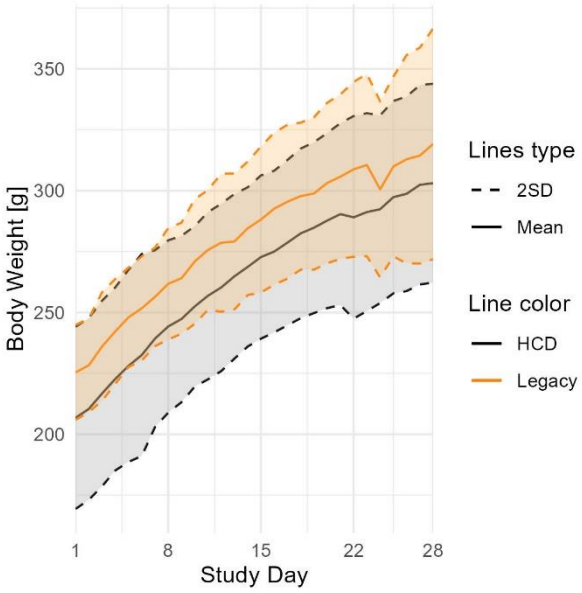

Figure S15: Mean body weight growth of male control group animals of Study-12 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

the 2•sd range of the legacy study and historical control data respectively.

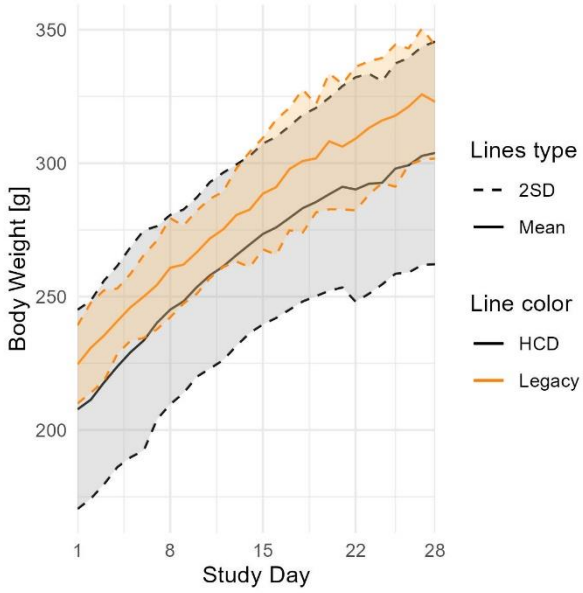

Figure S16: Mean body weight growth of male control group animals of Study-13 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

the 2•sd range of the legacy study and historical control data respectively.

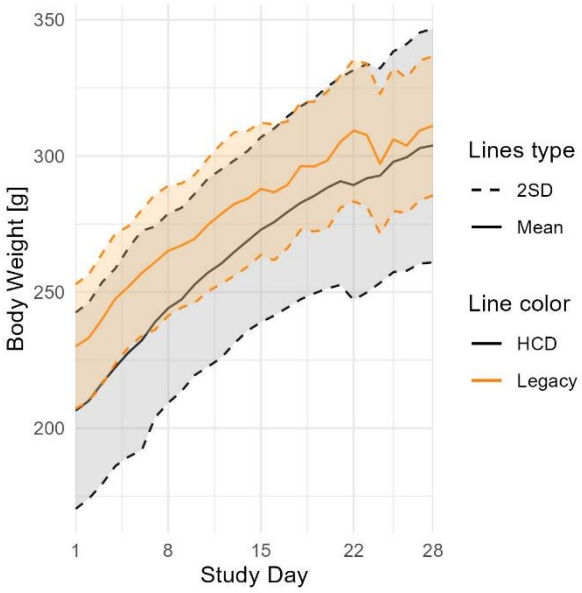

Figure S17: Mean body weight growth of male control group animals of Study-14 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

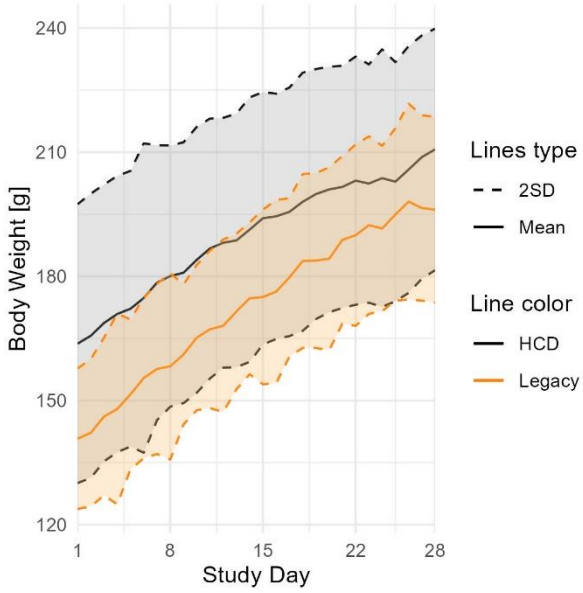

Figure S18: Mean body weight growth of female control group animals of Study-01 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

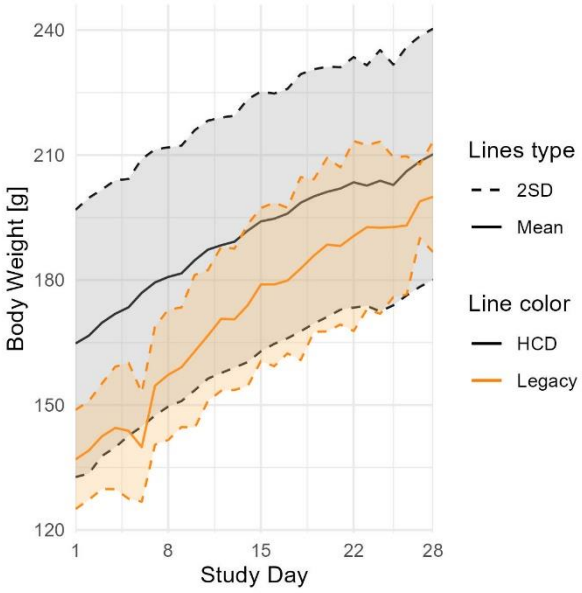

Figure S19: Mean body weight growth of female control group animals of Study-02 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

the 2•sd range of the legacy study and historical control data respectively.

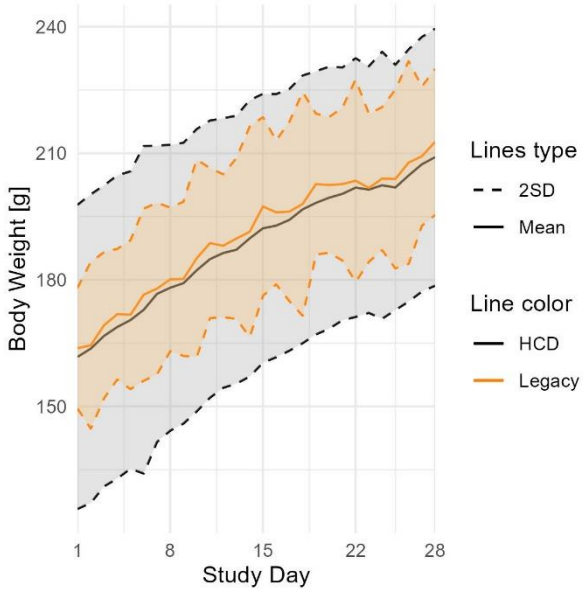

Figure S20: Mean body weight growth of female control group animals of Study-03 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

the 2•sd range of the legacy study and historical control data respectively.

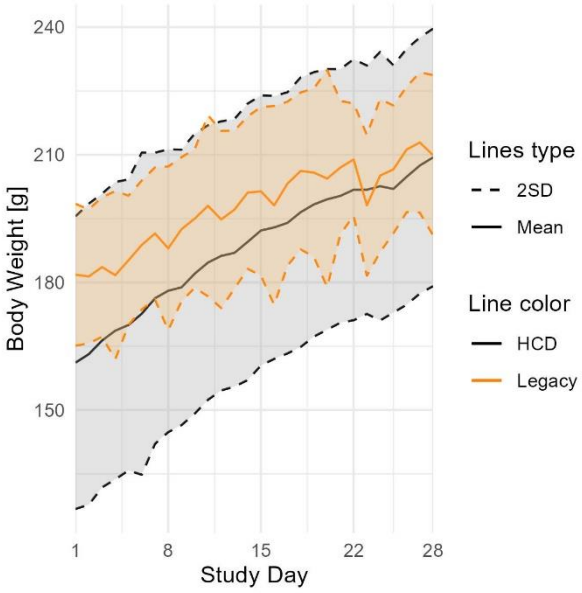

Figure S21: Mean body weight growth of female control group animals of Study-04 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

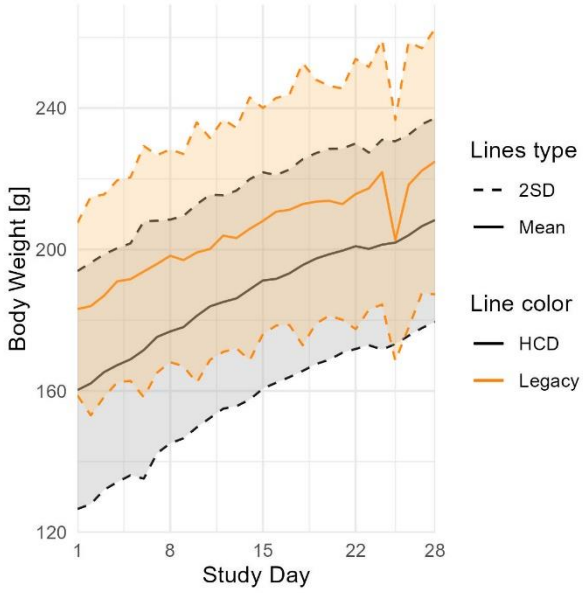

Figure S22: Mean body weight growth of female control group animals of Study-05 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

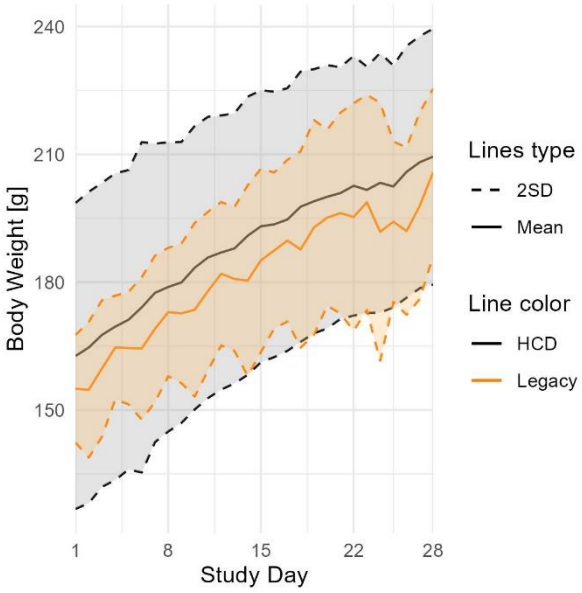

Figure S23: Mean body weight growth of female control group animals of Study-06 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

the 2•sd range of the legacy study and historical control data respectively.

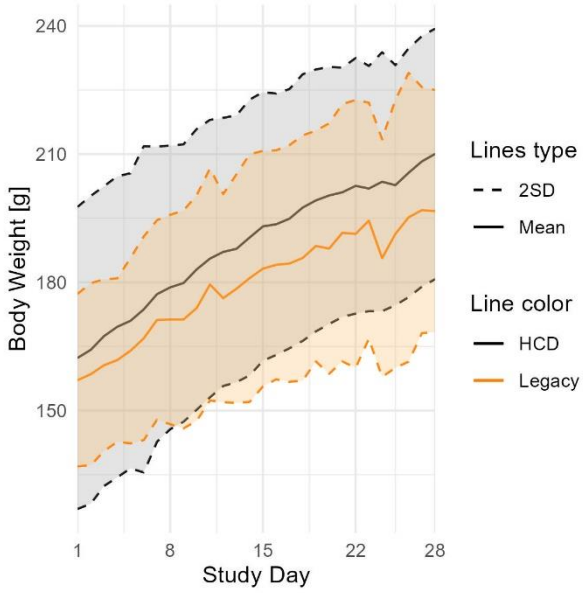

Figure S24: Mean body weight growth of female control group animals of Study-07 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

the 2•sd range of the legacy study and historical control data respectively.

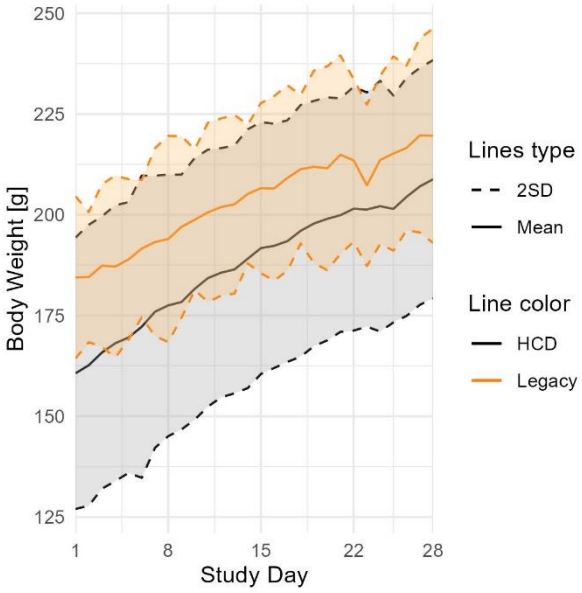

Figure S25: Mean body weight growth of female control group animals of Study-08 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

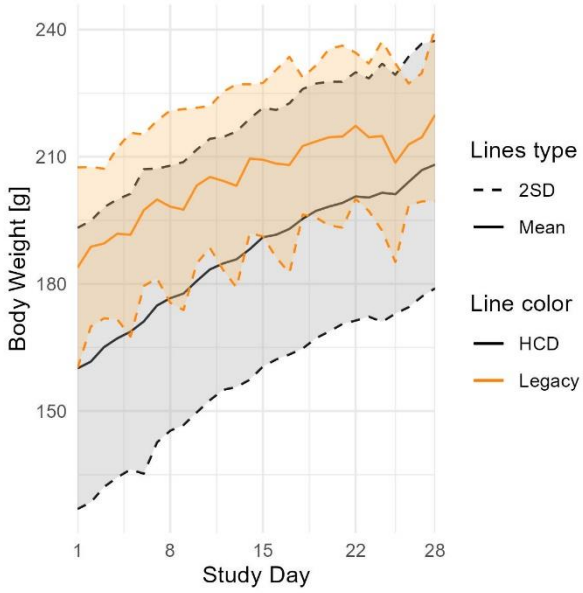

Figure S26: Mean body weight growth of female control group animals of Study-09 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

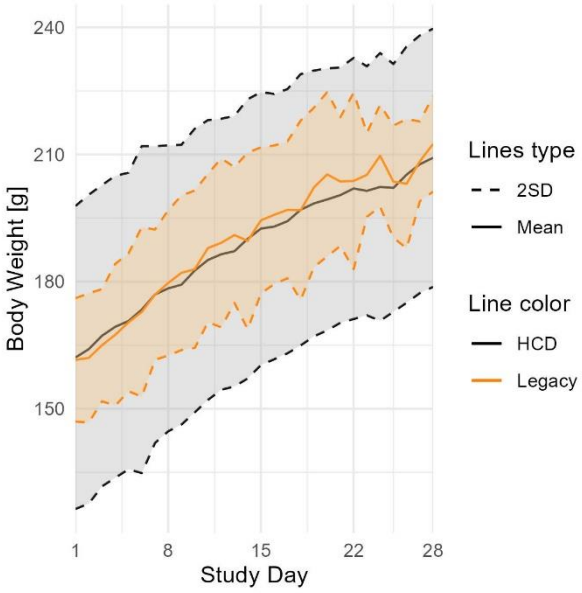

Figure S27: Mean body weight growth of female control group animals of Study-10 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent

the 2•sd range of the legacy study and historical control data respectively.

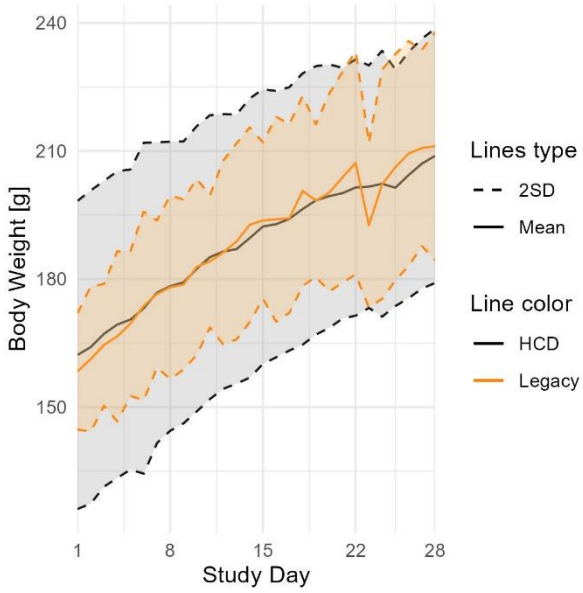

Figure S28: Mean body weight growth of female control group animals of Study-11 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

the 2•sd range of the legacy study and historical control data respectively.

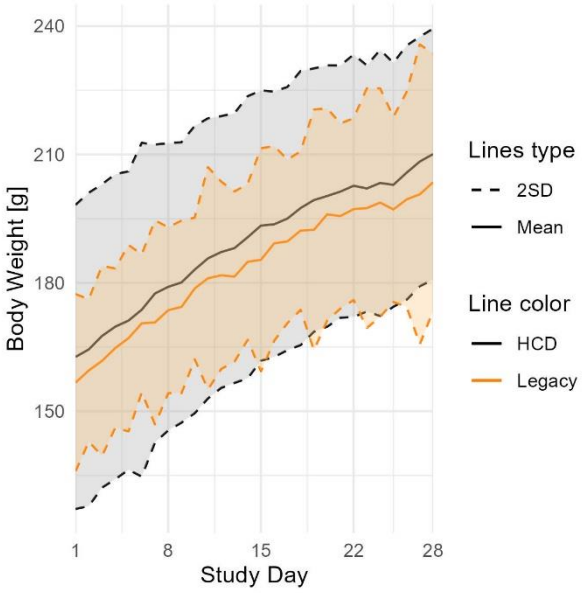

Figure S29: Mean body weight growth of female control group animals of Study-12 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

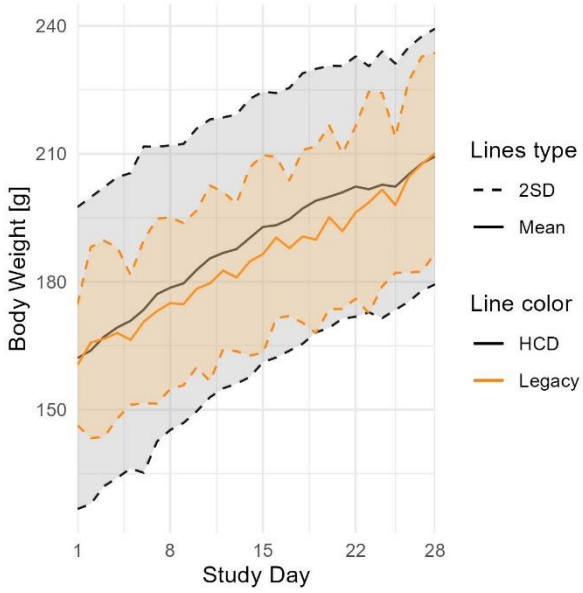

Figure S30: Mean body weight growth of female control group animals of Study-13 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

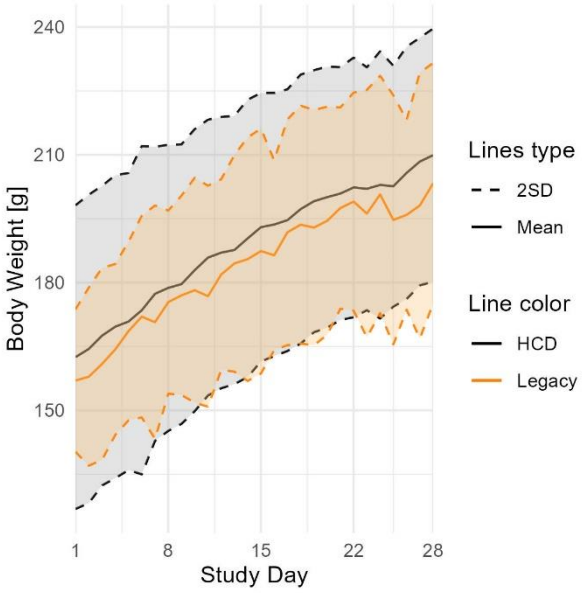

Figure S31: Mean body weight growth of female control group animals of Study-14 (orange line) compared to the mean growth of historical data control groups (grey line) from study day 1-28. The dashed lines with the painted area represent the 2•sd range of the legacy study and historical control data respectively.

2.3 HCD histograms, exemplary VCG selection and initial body weight box plots of legacy studies

In the following 28 figures, initial body weights (body weights measured on day 1 prior to the first substance application) of historical control data (HCD) are displayed as histograms in comparison to the initial body weight of the legacy study. Objective is to replace the concurrent control with a virtual control. Depending on the location parameters of the legacy study's groups, the VCG is either close or far away from the concurrent control. The latter would result in poor statistical reproducibility of the body weight differences between control and dose groups. The studies (Study-01 to Study-14) are sorted by increasing initial body weight of male animals with Study-01 having the lightest animals (compared to the HCD) and Study-14 having the heaviest (see Figures S32-S45). Note, that the same study nomenclature has been used for females as well which however is not sorted by increasing initial body weight of females (Figures S46-S59). This means for instance, that while Study-14 has the heaviest males, it does not necessary mean that females are the heaviest as well.

The figures share the same format: On the left, the histogram shows the probability density distribution of historical control data (HCD) used to generate virtual control groups (VCGs). Adjacent to the histogram, box plots show the one example iteration of a VCG in red, the legacy study's concurrent control group (CCG) in grey, and dose groups (low dose (LD), mid dose (MD), and high dose (HD)) in white. In Panel A, VCG was sampled from HCD not matched to legacy study weight distribution. Panel B shows VCG sampled from HCD matched to legacy-study-dose groups' initial body weight (black dashed line). Panel C shows VCG generated by weighted sampling with weights assigned to HCD to match the probability density of legacy study dose groups (black solid line on the histogram).

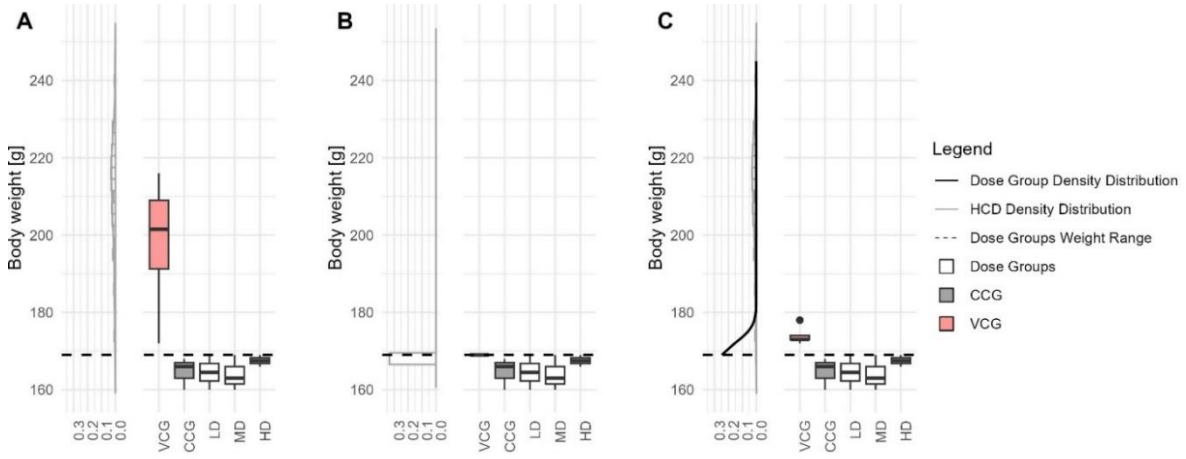

Figure S32: Virtual control groups for male animals Study-01, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S32A)** Non-matched HCD. **S32B)** HCD matched to initial body weights. **S32C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

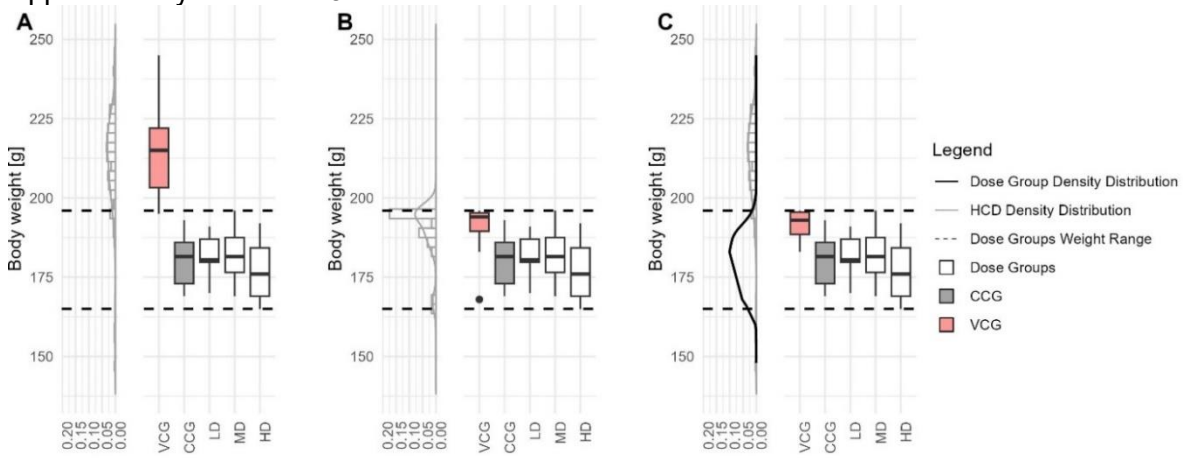

Figure S33: Virtual control groups for male animals Study-02, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S33A)** Non-matched HCD. **S33B)** HCD matched to initial body weights. **S33C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

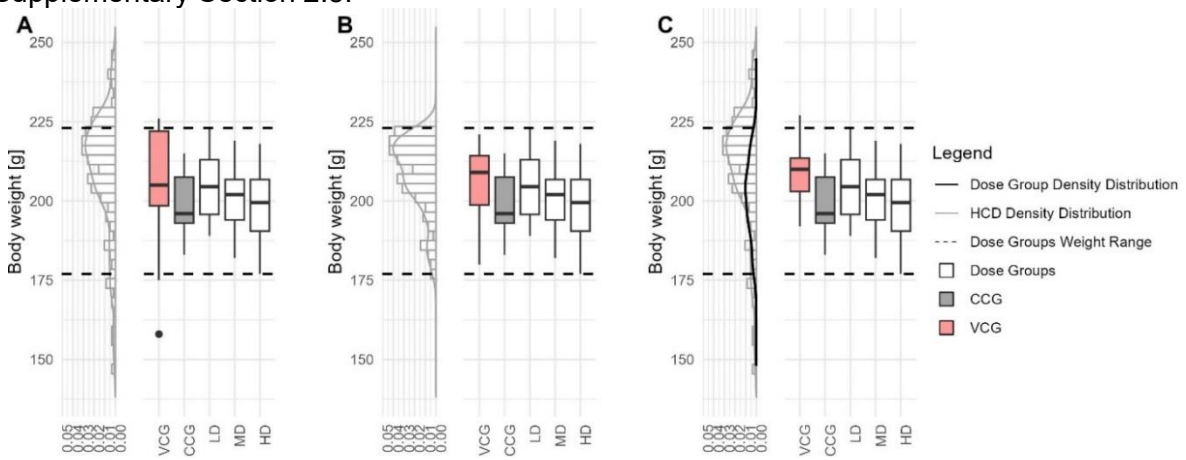

Figure S34: Virtual control groups for male animals Study-03, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S34A)** Non-matched HCD. **S34B)** HCD matched to initial body weights. **S34C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

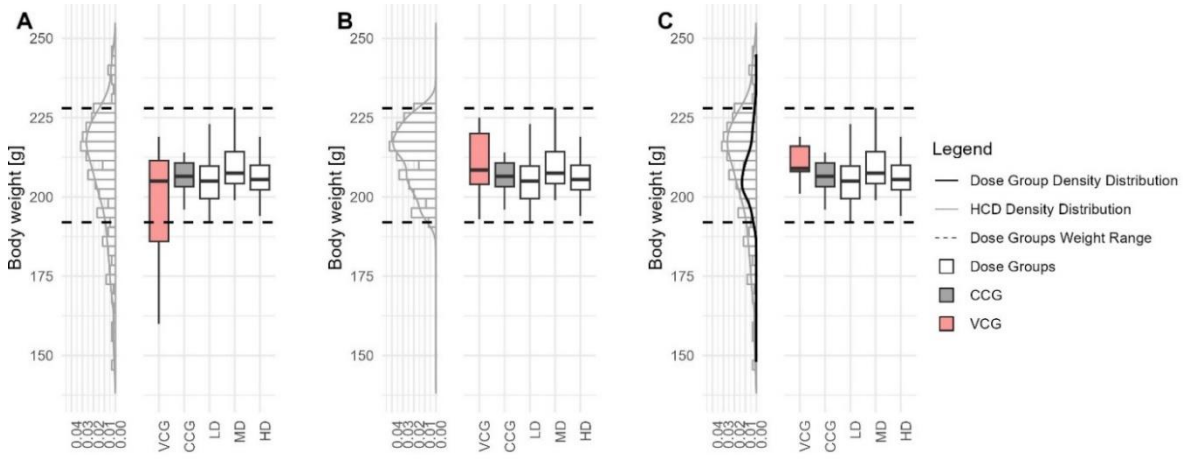

Figure S35: Virtual control groups for male animals Study-04, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S35A)** Non-matched HCD. **S35B)** HCD matched to initial body weights. **S35C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

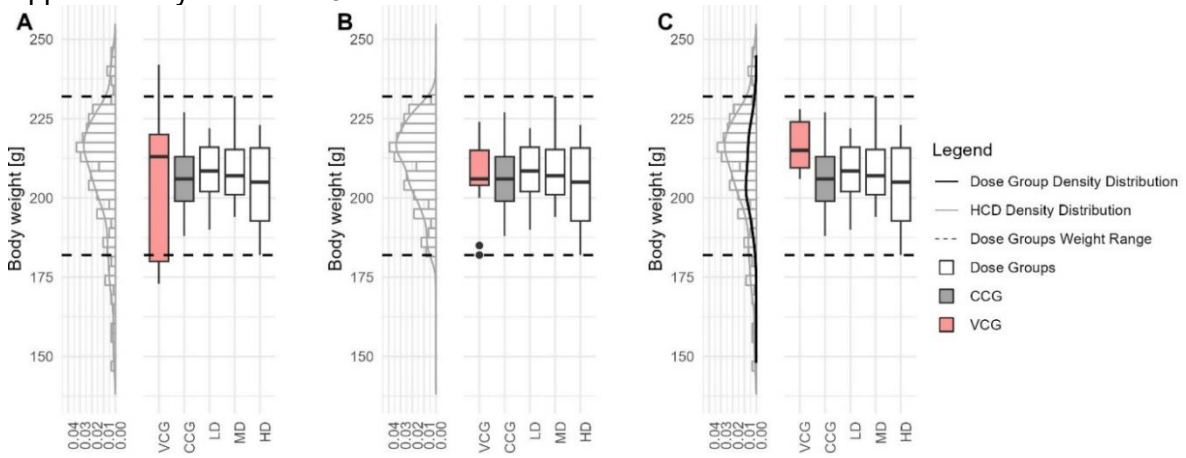

Figure S36: Virtual control groups for male animals Study-05, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S36A)** Non-matched HCD. **S36B)** HCD matched to initial body weights. **S36C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

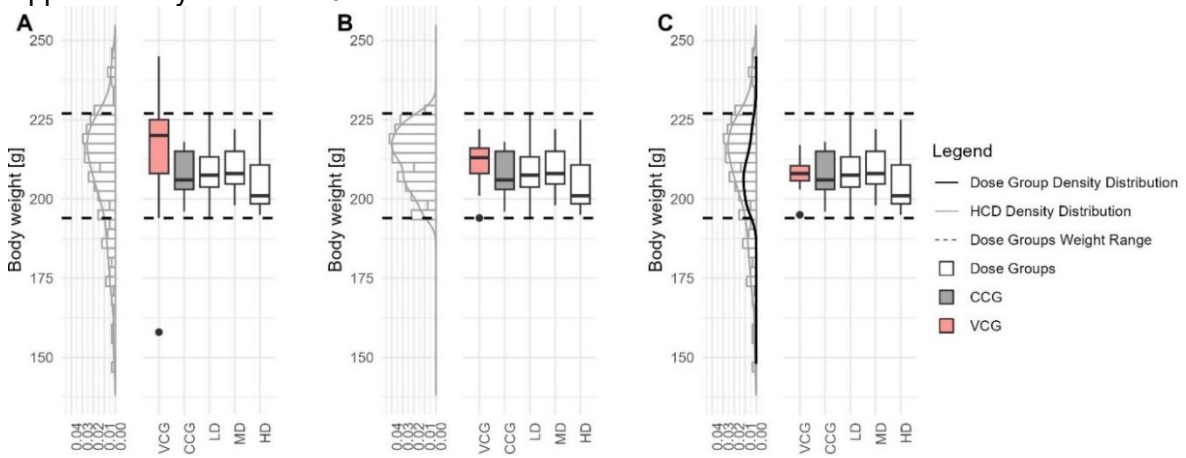

Figure S37: Virtual control groups for male animals Study-06, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S37A)** Non-matched HCD. **S37B)** HCD matched to initial body weights. **S37C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

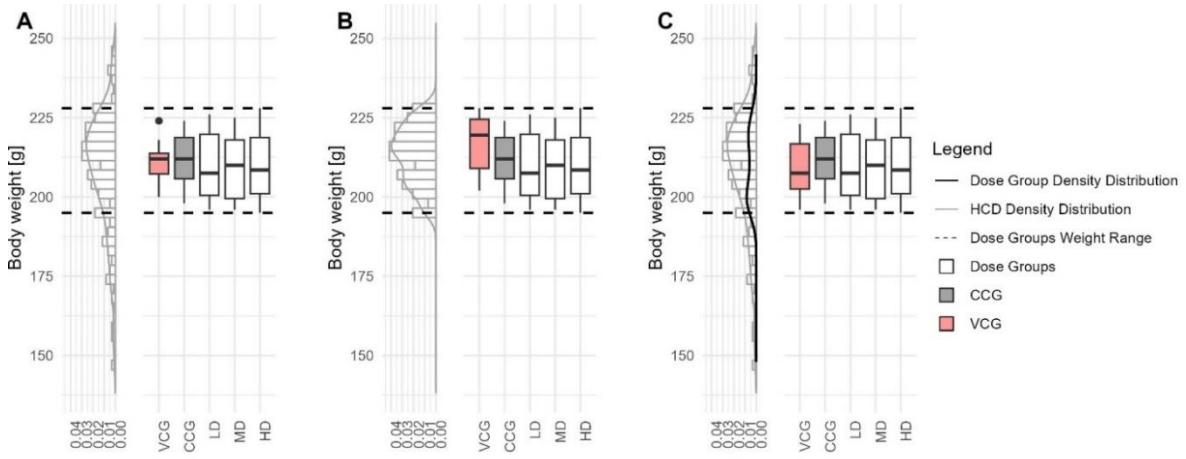

Figure S38: Virtual control groups for male animals Study-07, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S38A)** Non-matched HCD. **S38B)** HCD matched to initial body weights. **S38C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

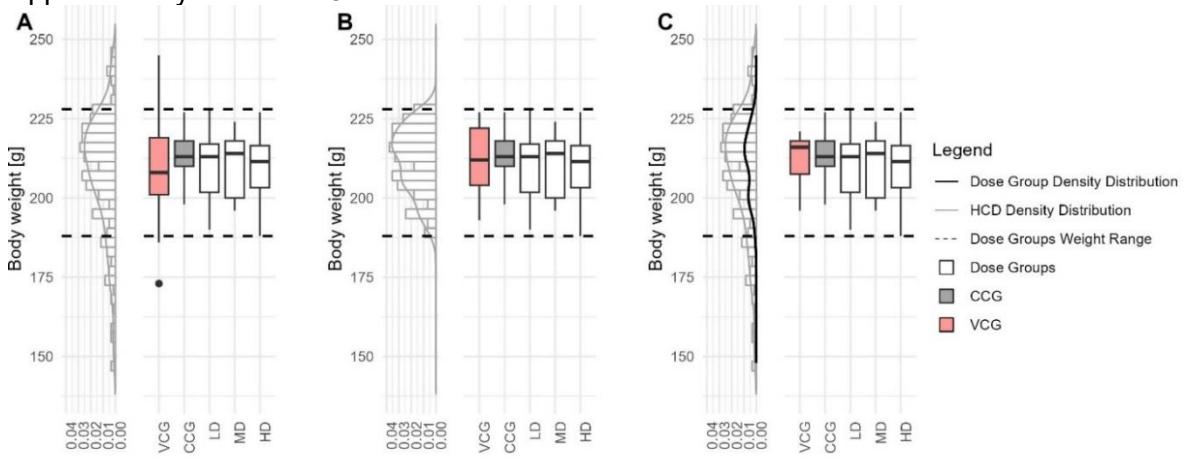

Figure S39: Virtual control groups for male animals Study-08, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S39A)** Non-matched HCD. **S39B)** HCD matched to initial body weights. **S39C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

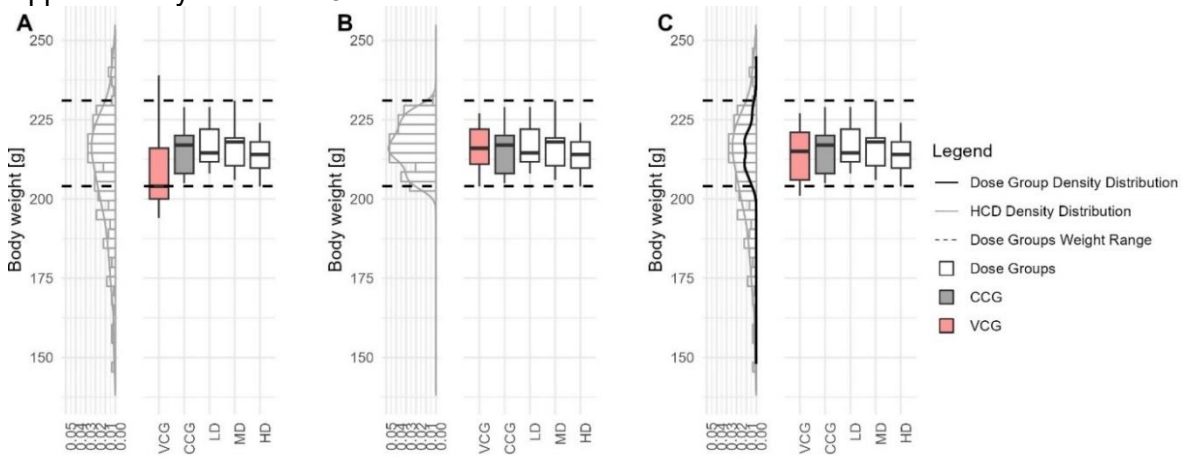

Figure S40: Virtual control groups for male animals Study-09, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S40A)** Non-matched HCD. **S40B)** HCD matched to initial body weights. **S40C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

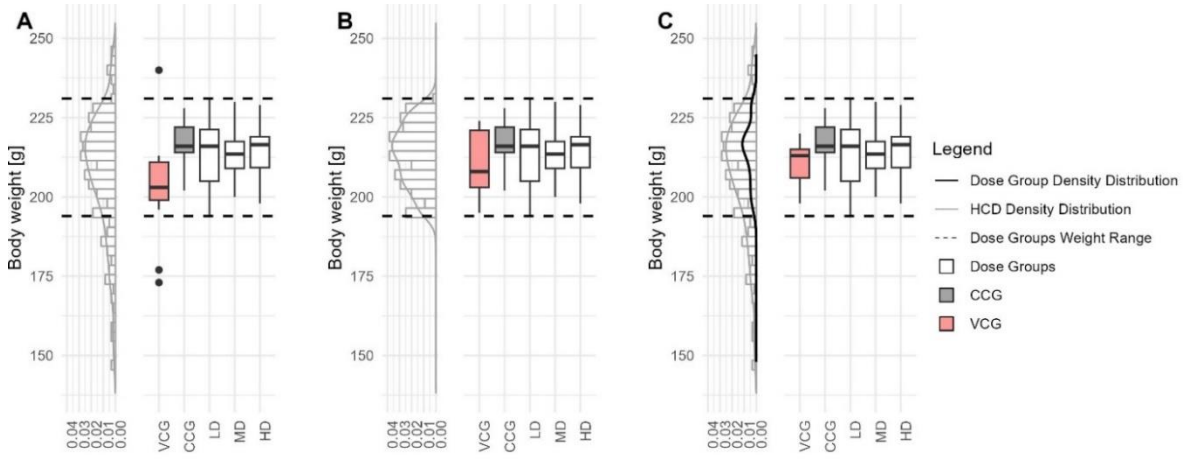

Figure S41: Virtual control groups for male animals Study-10, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S41A)** Non-matched HCD. **S41B)** HCD matched to initial body weights. **S41C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

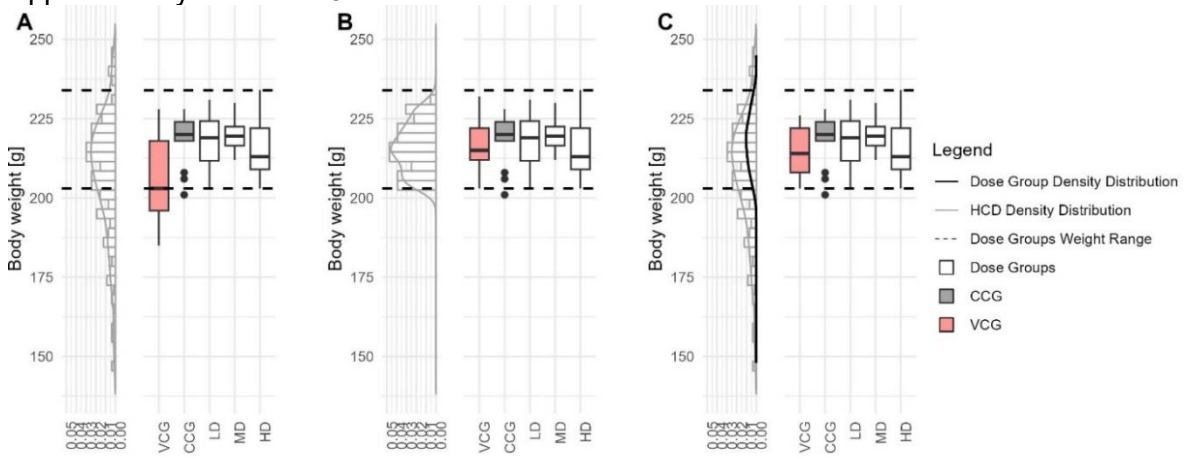

Figure S42: Virtual control groups for male animals Study-11, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S42A)** Non-matched HCD. **S42B)** HCD matched to initial body weights. **S42C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

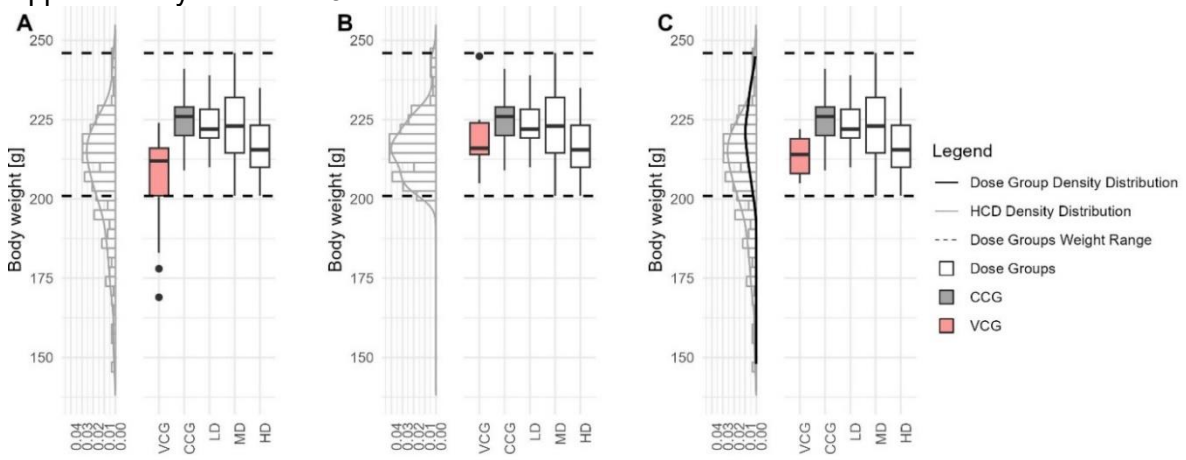

Figure S43: Virtual control groups for male animals Study-12, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S43A)** Non-matched HCD. **S43B)** HCD matched to initial body weights. **S43C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

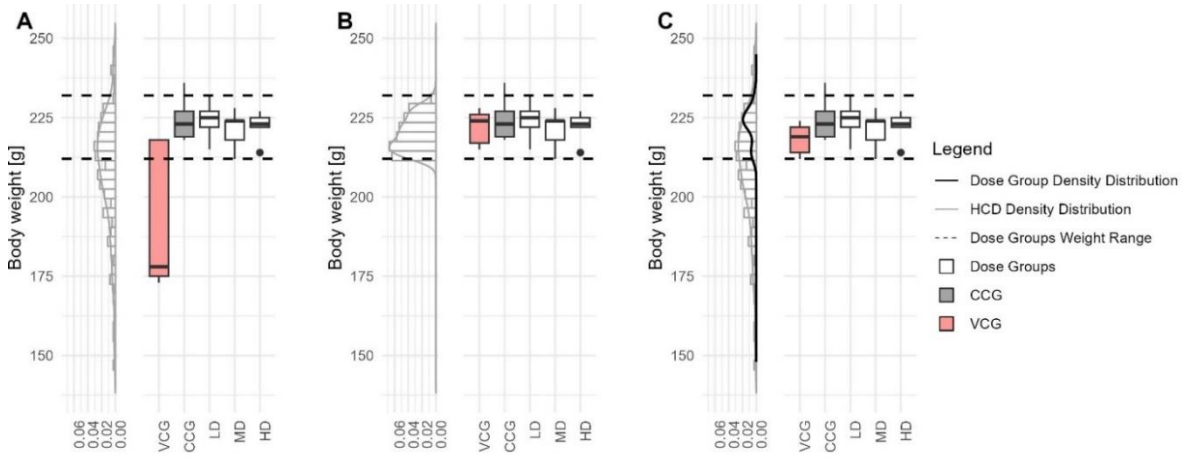

Figure S44: Virtual control groups for male animals Study-13, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S44A)** Non-matched HCD. **S44B)** HCD matched to initial body weights. **S44C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

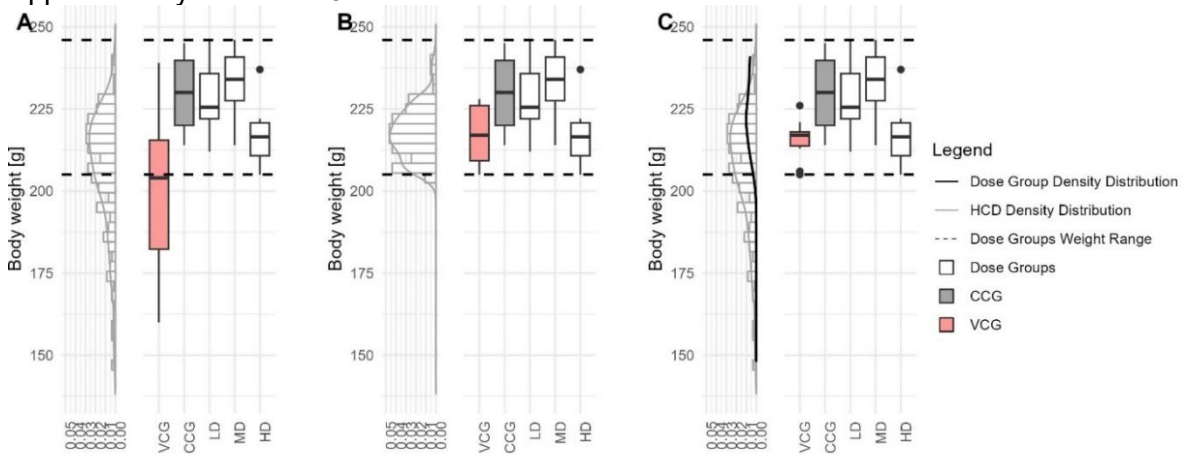

Figure S45: Virtual control groups for male animals Study-14, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S45A)** Non-matched HCD. **S45B)** HCD matched to initial body weights. **S45C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

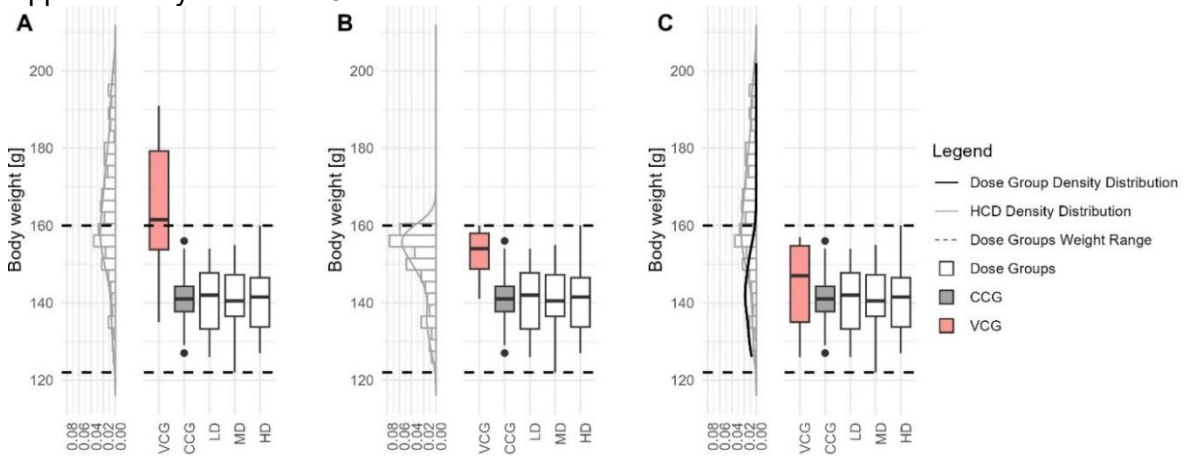

Figure S46: Virtual control groups for female animals Study-01, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S46A)** Non-matched HCD. **S46B)** HCD matched to initial body weights. **S46C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

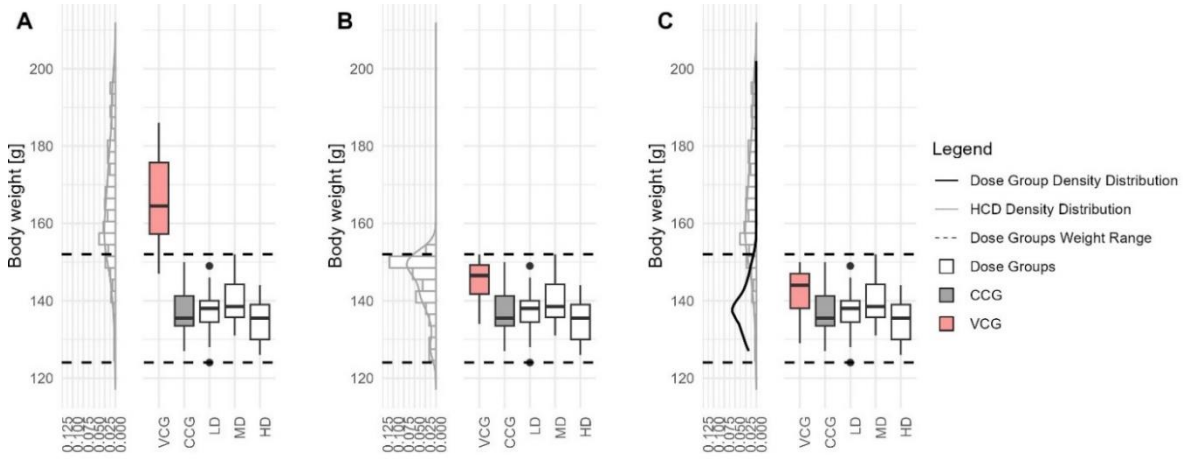

Figure S47: Virtual control groups for female animals Study-02, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S47A)** Non-matched HCD. **S47B)** HCD matched to initial body weights. **S47C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

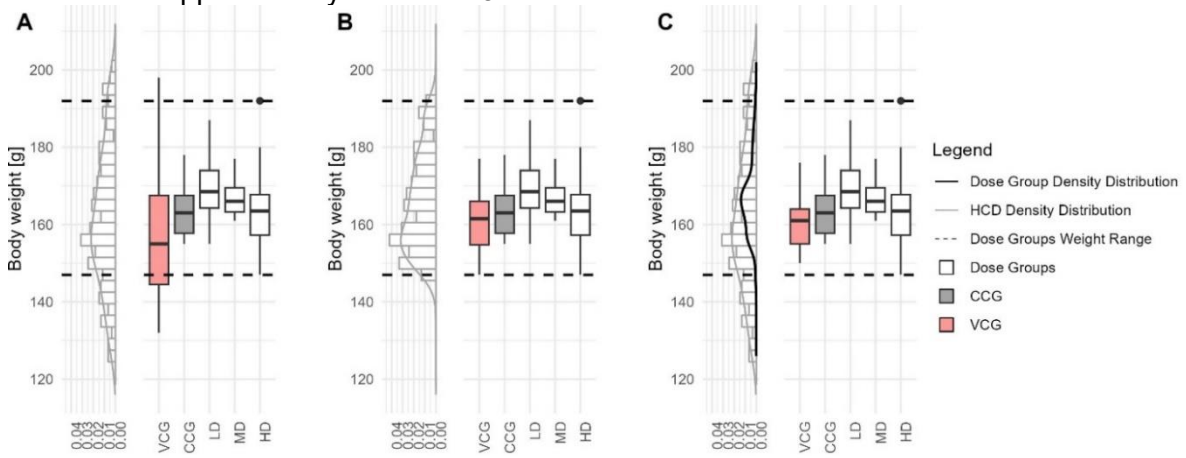

Figure S48: Virtual control groups for female animals Study-03, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S48A)** Non-matched HCD. **S48B)** HCD matched to initial body weights. **S48C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

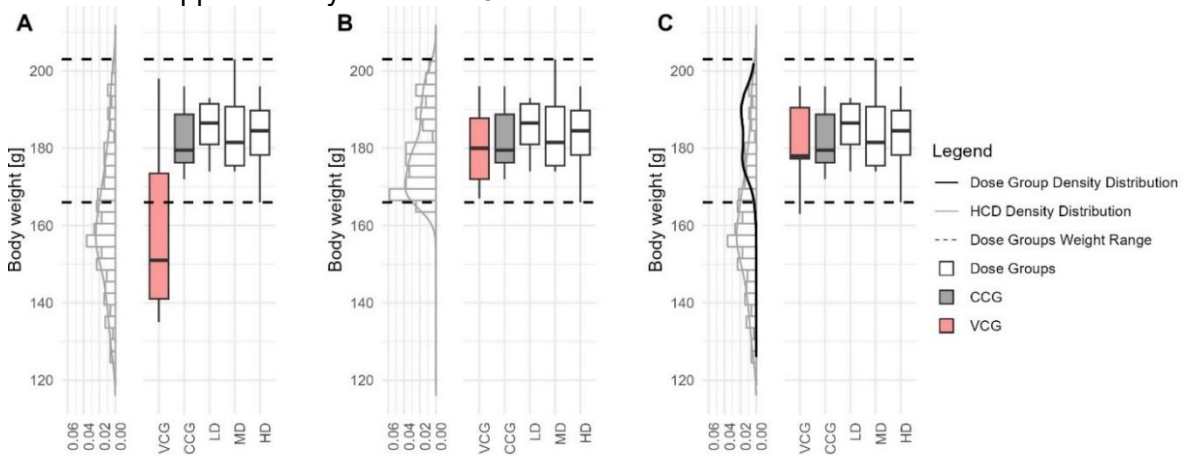

Figure S49: Virtual control groups for female animals Study-04, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S49A)** Non-matched HCD. **S49B)** HCD matched to initial body weights. **S49C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

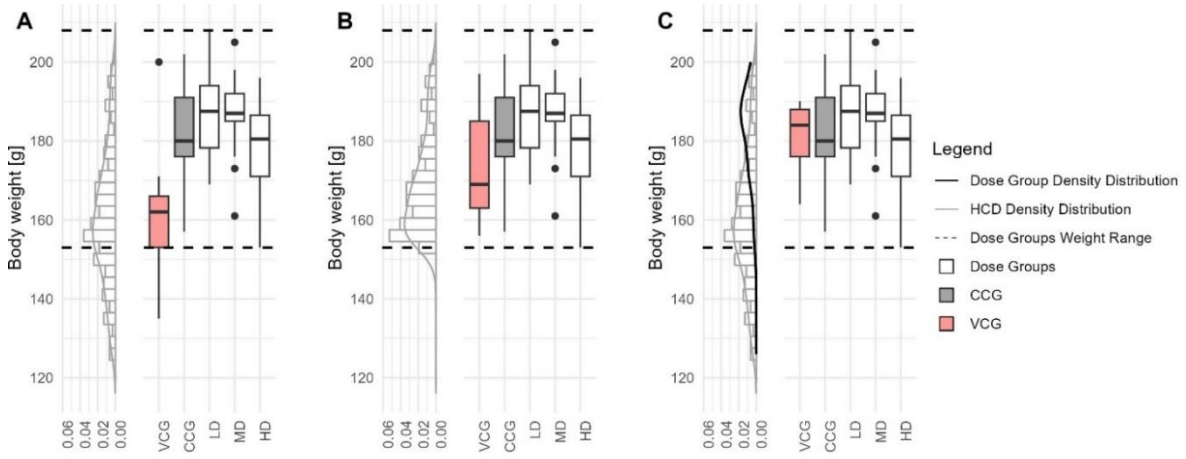

Figure S50: Virtual control groups for female animals Study-05, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S50A)** Non-matched HCD. **S50B)** HCD matched to initial body weights. **S50C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

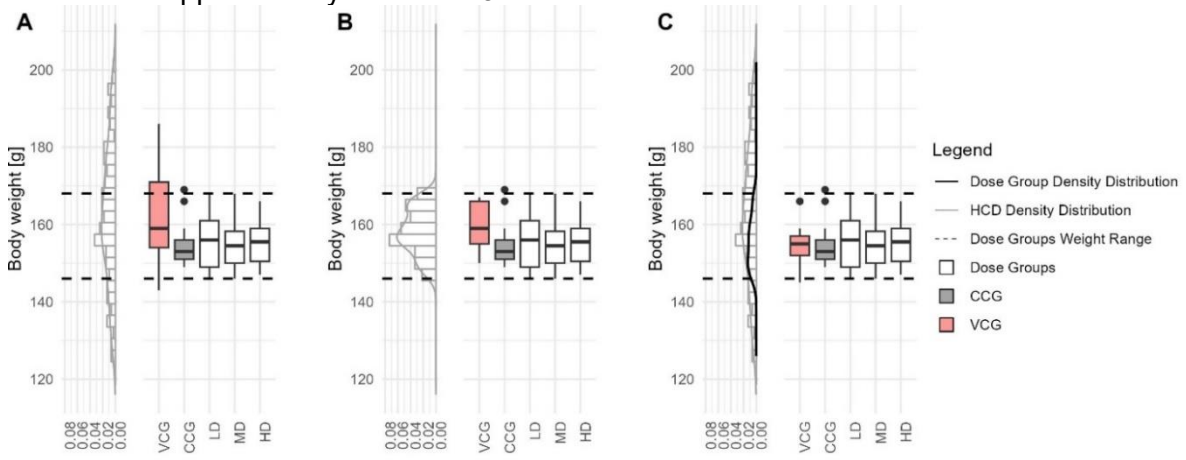

Figure S51: Virtual control groups for female animals Study-06, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S51A)** Non-matched HCD. **S51B)** HCD matched to initial body weights. **S51C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

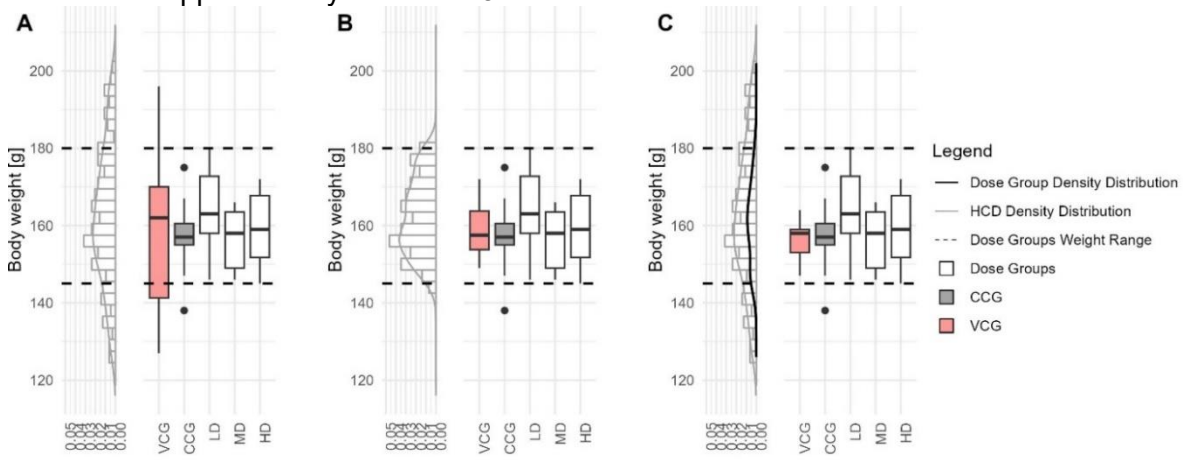

Figure S52: Virtual control groups for female animals Study-07, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S52A)** Non-matched HCD. **S52B)** HCD matched to initial body weights. **S52C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

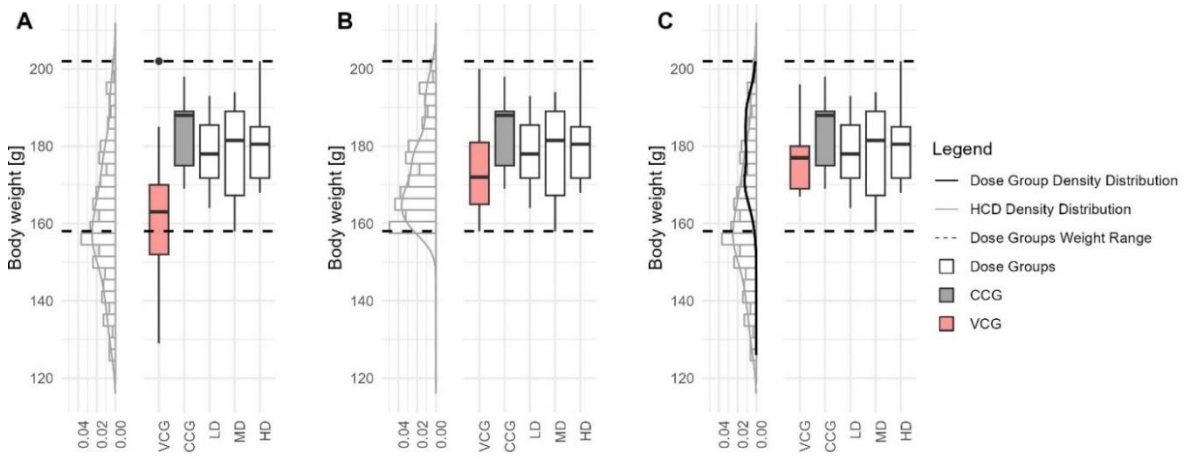

Figure S53: Virtual control groups for female animals Study-08, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S53A)** Non-matched HCD. **S53B)** HCD matched to initial body weights. **S53C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

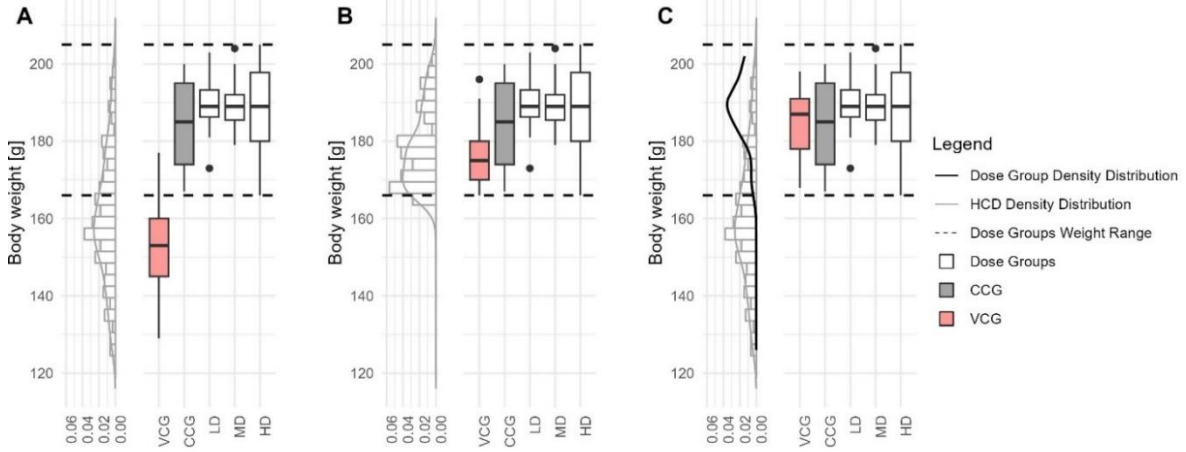

Figure S54: Virtual control groups for female animals Study-09, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S54A)** Non-matched HCD. **S54B)** HCD matched to initial body weights. **S54C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

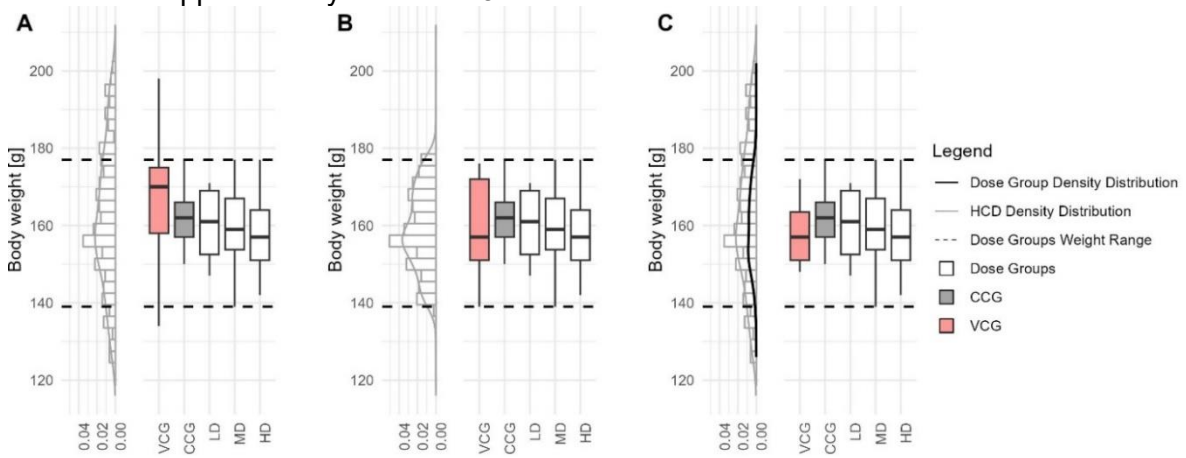

Figure S55: Virtual control groups for female animals Study-10, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S55A)** Non-matched HCD. **S55B)** HCD matched to initial body weights. **S55C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

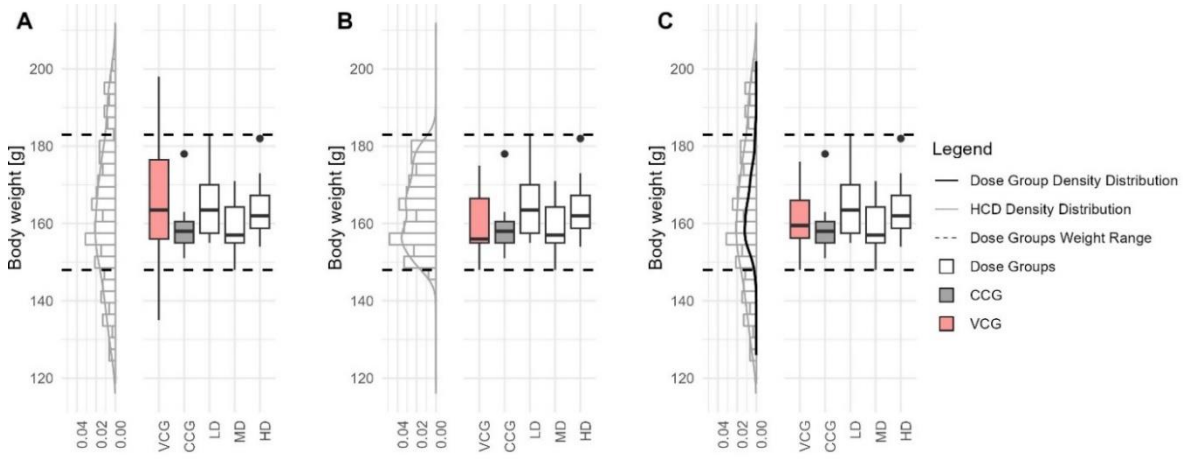

Figure S56: Figure S56: Virtual control groups for female animals Study-11, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S56A)** Non-matched HCD. **S56B)** HCD matched to initial body weights. **S56C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

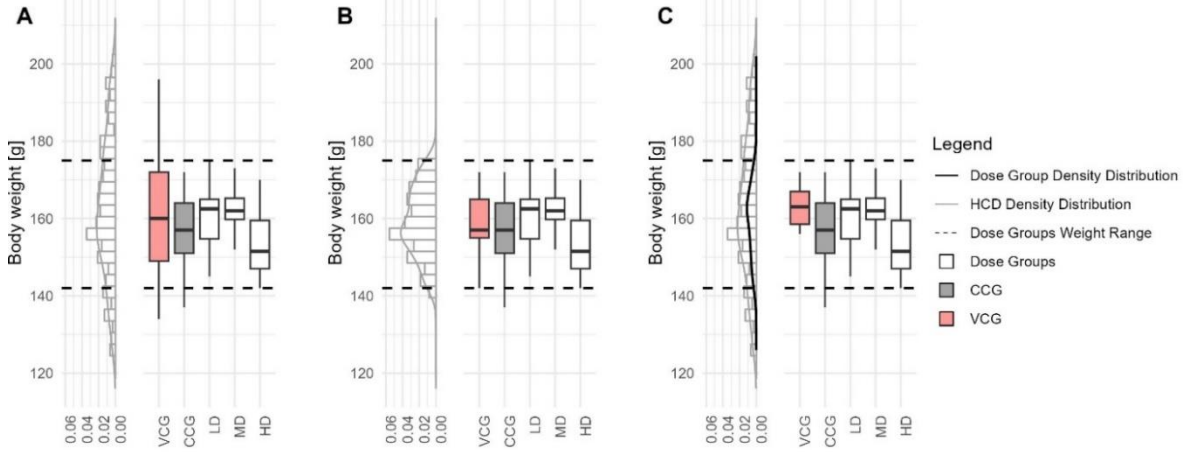

Figure S57: Virtual control groups for female animals Study-12, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S57A)** Non-matched HCD. **S57B)** HCD matched to initial body weights. **S57C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

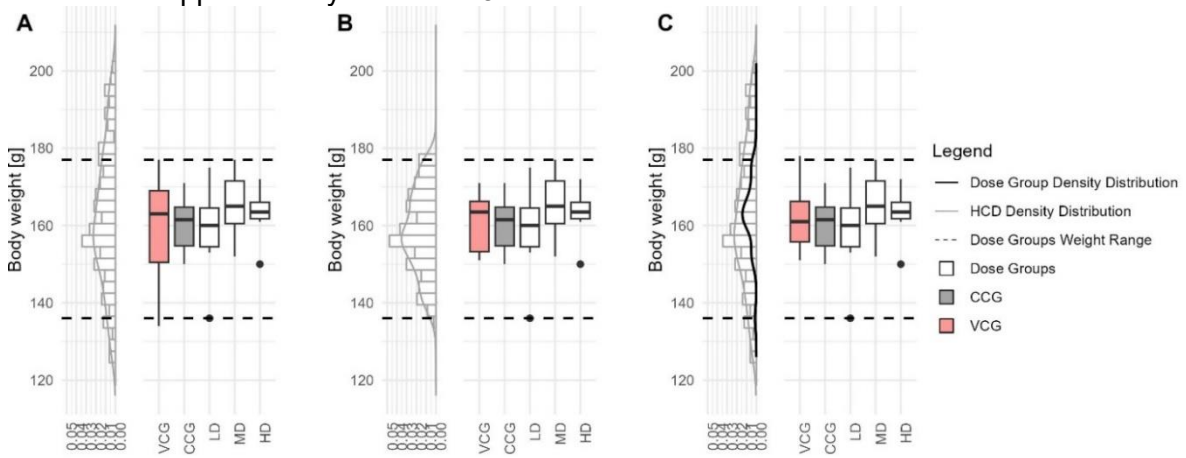

Figure S58: Virtual control groups for female animals Study-13, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S58A)** Non-matched HCD. **S58B)** HCD matched to initial body weights. **S58C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

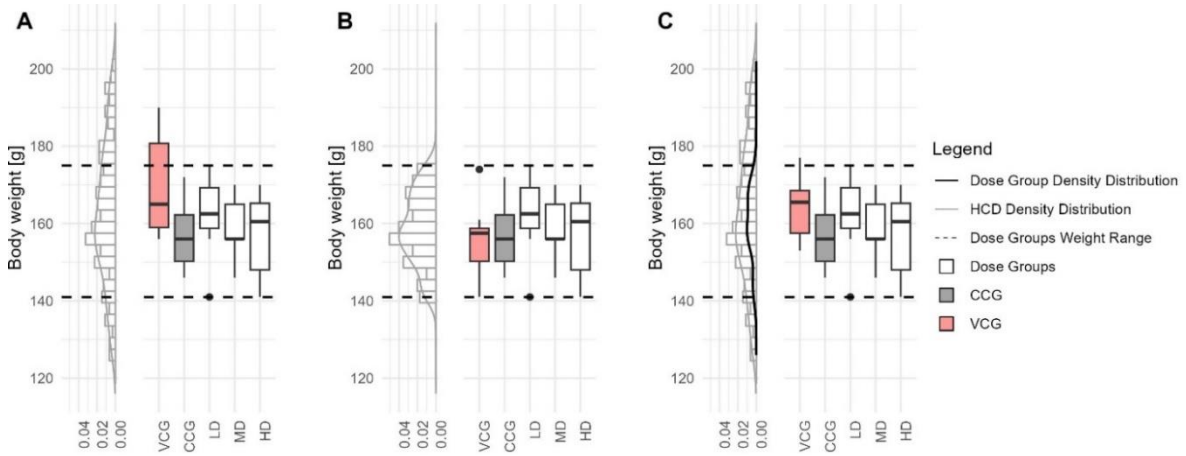

Figure S59: Virtual control groups for female animals Study-14, showing VCG (red), CCG (grey), and dose groups (LD, MD, HD in white). **S59A)** Non-matched HCD. **S59B)** HCD matched to initial body weights. **S59C)** VCG via weighted sampling. Detailed descriptions are provided in Supplementary Section 2.3.

3 References

Dowle, M. and Srinivasan, A. (2023). Data.Table: Extension of `data.Frame`. R package version 1.14.8. Https://cran.R-project.Org/package=data.Table.

Gurjanov (2024b) https://github.com/bayer-group/VCG-INITBW.

Pohlert, T. (2022). Pmcmrplus: Calculate pairwise multiple comparisons of mean rank sums extended. R package version 1.9.6. Https://cran.R-project.Org/package=pmcmrplus.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schauberger, P. and Walker, A. (2023). Openxlsx: Read, write and edit xlsx files. R package version 4.2.5.2. Https://cran.R-project.Org/package=openxlsx.

Wilke, C. O. (2020). Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.1.1. https://cran.R-project.Org/package=cowplot.

3.3 Replacing concurrent controls with virtual control groups in rat toxicity studies

The previous chapters showed an approach to assess the VCG performance on the ability to reproduce statistical outcomes of legacy studies and presented methods to improve this performance. The article in this chapter goes beyond this assessment approach by presenting a method to evaluate VCG performance across all endpoints measured in rat toxicity studies, underscoring the notion that statistical significance does not automatically result in biological relevance. In this study, all parameters—both quantitative and qualitative—are taken into account, ensuring that the study evaluation process is faithfully reproduced in the manner typically employed in assessments of regulatory toxicology studies. This article presents a workflow to semi-systematically evaluate the performance of VCGs and lays the foundation for future validations in the context of nonclinical regulatory toxicity studies.

Authors: A. Gurjanov, C. Vieira-Vieira, J. Vienenkoetter, L. A. I. Vaas, T. Steger-Hartmann

CRediT author statement: *Conceptualization*: AG, LV, TSH; *Methodology*: AG, CVV, LV, TSH Software: AG; *Validation*: AG, CVV, JV, LV, TSH; *Formal analysis*: AG; *Investigation*: AG; *Resources*: CVV; *Data curation*: AG, CVV; *Writing – original draft*: AG; *Writing – review and editing*: AG, CVV, JV, LV, TSH; *Visualization*: AG; *Supervision*: LV, TSH; *Project administration*: TSH; *Funding acquisition*: TSH

Citation: Alexander Gurjanov, Carlos Vieira-Vieira, Julia Vienenkoetter, Lea A.I. Vaas, Thomas Steger-Hartmann, Replacing concurrent controls with virtual control groups in rat toxicity studies, Regulatory Toxicology and Pharmacology, Volume 148, 2024, 105592, ISSN 0273-2300, https://doi.org/10.1016/j.yrtph.2024.105592. (https://www.sciencedirect.com/science/article/pii/S0273230024000333)

Regulatory Toxicology and Pharmacology 148 (2024) 105592

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Replacing concurrent controls with virtual control groups in rat toxicity studies

Alexander Gurjanov [a,*], Carlos Vieira-Vieira [a], Julia Vienenkoetter [b], Lea A.I. Vaas [c], Thomas Steger-Hartmann [a]

[a] Bayer Research & Development, Pharmaceuticals, Investigative Toxicology, Berlin, Germany
[b] Bayer Research & Development, Pharmaceuticals, Pathology and Clinical Pathology, Wuppertal, Germany
[c] Bayer Research & Development, Pharmaceuticals, Research & Pre-Clinical Statistics Group, Berlin, Germany

ARTICLE INFO

Handling Editor: Lesa Aylward

Keywords:
3R
Historical control data
Virtual control groups
Biological relevance
Statistical modeling

ABSTRACT

Virtual control groups (VCGs) in nonclinical toxicity represent the concept of using appropriate historical control data for replacing concurrent control group animals. Historical control data collected from standardized studies can serve as base for constructing VCGs and legacy study reports can be used as a benchmark to evaluate the VCG performance. Replacing concurrent controls of legacy studies with VCGs should ideally reproduce the results of these studies. Based on three four-week rat oral toxicity legacy studies with varying degrees of toxicity findings we developed a concept to evaluate VCG performance on different levels: the ability of VCGs to (i) reproduce statistically significant deviations from the concurrent control, (ii) reproduce test substance-related effects, and (iii) reproduce the conclusion of the toxicity study in terms of threshold dose, target organs, toxicological biomarkers (clinical pathology) and reversibility. Although VCGs have shown a low to moderate ability to reproduce statistical results, the general study conclusions remained unchanged. Our results provide a first indication that carefully selected historical control data can be used to replace concurrent control without impairing the general study conclusion. Additionally, the developed procedures and workflows lay the foundation for the future validation of virtual controls for a use in regulatory toxicology.

1. Introduction

Preclinical toxicity studies are conducted for safety assessment of drug candidates following strict regulatory frames regarding employed animals, environmental conditions, and analytical procedures (ICH, 2009). They should characterize potential test substance-induced toxicity and its exposure-relationship ideally in pharmacologically relevant animal species, determining a safe starting dose for later clinical trials. In particular, the no observed adverse effect level (NOAEL) which describes the highest tested exposure of the test substance that did not cause any adverse effects in the animals, is a key parameter for estimating a safe starting dose in humans.

Study standardization efforts aim to limit experimental variability and to increase sensitivity towards effects caused by a test substance (Howard, 2002; Carroll, 2016). Standardization also allows pooling data from control group animals performed under the same experimental conditions. Pooled animal control data is then used for calculating background incidences of histological lesions (Haseman et al., 1984), as reference values for organ weights or clinical pathology parameters, and as a general quality control (i.e., estimating expected value range in bioassays) (Kluxen et al., 2021). A milestone for standardization was the introduction of the Standard for Exchange of Nonclinical Studies (SEND) (Wood, 2008; CDISC, 2022) facilitating the collection and evaluation of nonclinical safety studies. Being aware of the potential of animal historical control data (HCD) in research, the European Innovative Medicine Initiative consortium *eTOX* (Sanz et al., 2017) and the follow-up consortium *eTRANSAFE* (Sanz et al., 2023) developed a database for animal toxicity studies, also collecting large sets of control animal data.

Previously, we proposed to explore the extension of the use of historical control data to construct virtual control groups (VCGs) (Steger-Hartmann et al., 2020). VCGs would allow to replace concurrent control group (CCG) animals and potentially reduce the number of animals used in animal toxicity studies for safety assessment by up to 25%. VCGs shall be built from selected control animals from the HCD pool matching animals in the treatment groups by age or initial body weights, route of administration, vehicle, study duration and further study design parameters (Steger-Hartmann et al., 2020). A previous study has

* Corresponding author. Bayer Aktiengesellschaft, Genetic Tox & Computational Tox, S116, 5.549, Müllerstraße 178, 13353, Berlin, Germany.
E-mail address: alexander.gurjanov@bayer.com (A. Gurjanov).

https://doi.org/10.1016/j.yrtph.2024.105592
Received 12 October 2023; Received in revised form 15 February 2024; Accepted 21 February 2024
Available online 22 February 2024

Abbreviations

VCG	Virtual Control Group
HCD	Historical Control Data
CCG	Concurrent Control Group
LD	Low Dose
MD	Mid Dose
HD	High Dose
NO(A)EL	No Observed (Adverse) Effect Level
STD	Severely Toxic Dose
GGT	gamma-Glutamyl Transferase
SEND	Standard for Exchange of Non-clinical Data
LON	Limits Of Normal
SME	Subject Matter Expert

explored the use of control data from the shared *eTOX* data base to analyze the impact of replacing concurrent control groups with VCGs on treatment-related findings (Wright et al., 2023). This study identified study design parameters (referred there as covariates) significantly influencing toxicological study outcomes when CCGs are replaced by VCGs and demonstrated that the best VCG performance was achieved if these covariates were kept at a high level of similarity. This study was, however, limited by the use of only aggregated animal data (i.e., mean and standard deviation).

In our recent study (Gurjanov et al., 2023), the collection of historical control data (HCD) at the level of individual animals served as basis for investigating the influence of animal handling procedures, particularly anesthesia protocols, on blood electrolyte measurement. Although we showed that generating VCGs from the wealth of accumulated HCD is generally possible, if influencing factors are strictly controlled, the impact of replacing CCGs on the overall toxicity study outcome has not yet been investigated.

After initial statistical analyses of the numerous quantitative parameters (e.g., body weights, organ weights, clinical pathology parameters), the entirety of quantitative and qualitative (e.g., in-life observations, gross and microscopic pathology) endpoints of a study are holistically assessed by the study director. There is currently no automated or generalized workflow for the directors to interpret in-life phase data (Baird et al., 2019; Baldrick et al., 2020). Consequentially, it is not sufficient to compare the results of original studies using a CCG to the re-analyzed results using VCGs merely on a statistical basis. Comparison of matching statistical results (Wright et al., 2023) does not necessarily provide full insight in the usability of VCGs, as changes in statistical significance are also observed when studies are repeated under maximum similar conditions (Poland et al., 2014). Experimental variability is expected in toxicity studies, but the conventional study design *per se* is another factor causing this lack of reproducibility of statistical results. Preclinical animal studies are usually statistically underpowered when assessing the null hypothesis for the multitude of parameters measured in such studies (Poland et al., 2014; Sena et al., 2014; Kluxen et al., 2021). While higher animal numbers would be desirable from a statistical point of view, the group sizes are set forth in international guidelines taking into account economical and ethical considerations. In this context, a difference in the statistical outcome between CCG and VCG might be of minor relevance or even irrelevant, if the affected parameter is not biologically related to the observed adversity (Kluxen et al., 2021). Therefore, a more comprehensive process for assessing the impact that a replacement of CCGs with VCGs might have on the toxicity study outcome is necessary. Such a process needs to recapitulate the way a toxicological study is assessed by the study director (Golden et al., 2023; Steger-Hartmann and Clark 2023). The statistical analysis of the collected data forms the basis for a further assessment performed by the study director or a subject matter expert (e.g., a clinical pathologist), who identifies the dose-response relationship, evaluates biological relevance or plausibility of findings while taking reference values into account and also considering prior knowledge on the test substance in order to differentiate between desired pharmacological effects or off-target findings.

In this article, we present a pilot process for a systematic validation for the feasibility of VCGs. This is done by re-assessing key study findings after substituting CCGs by VCGs in three legacy studies. The VCG performance, that is, the impact of replacing CCGs for VCGs in the outcome of a 4-week repeat dose systemic toxicity study in rats, was assessed in a three-step procedure as recently recommended (Golden et al., 2023). First, differences between VCG and dose groups were re-analyzed statistically for quantitative parameters. Study directors then assessed whether observed findings in the re-analyzed studies are to be attributed to the test substance treatment. Finally, based on the test-substance related findings study directors derived the no observed (adverse) effect level (NO(A)EL) or severely toxic dose affecting 10% of the animals (STD_{10}), identified target organs and relevant clinical pathology biomarkers (monitorability) and assessed the reversibility of adverse effects.

2. Materials and methods

This section provides an explanation of the creation process of VCGs and how we assessed their performance. An overview of the process is given in Fig. 1. For a better understanding, a definition of the used terminology is given below.

- *Legacy study*: toxicity study performed in the past. Contains data of concurrent control groups (CCGs) and dose groups. The written results of the study are used as a benchmark. Our goal is to reproduce the results with VCGs.
- *VCGs* (virtual control groups): a set of control group animal data derived from HCD applying appropriate filters for matching the animals of the treatment groups.
- *HCD* (historical control data): control-group data of toxicity studies performed in the past. HCD is used for creating VCGs and for toxicological decision making.
- *Initial body weight*: Body weight of dose group animals on day 1 of the study, prior to the first application of the test substance. Since the animals are assigned to the different control or treatment groups through randomization based on the initial body weight, the distribution of the initial body weight can be assumed to be identical between CCGs and treatment groups.

2.1. Selection of legacy studies

The selection criteria for the three legacy studies used for the qualification procedure were based on the following study design parameters:

- Dosing duration of 4 weeks.
- Studies were initiated in 2021 or 2022.
- A control group was used along with three dose groups: low dose (LD), mid dose (MD), and high dose (HD).
- Study was performed with Wistar HAN rats.
- Route of administration was oral by gavage.
- The studies had been performed at the same test facility.
- Treatment vehicle contained either Kolliphor® HS 15 or polyethylene glycol 400.
- Animals were supplied by the same breeder.
- Food and water supply during the study was *ad libitum*.
- Animals were housed in group cages with 2–3 animals per cage.
- Clinical pathology and histopathology parameters were measured on day 28 ± 7.

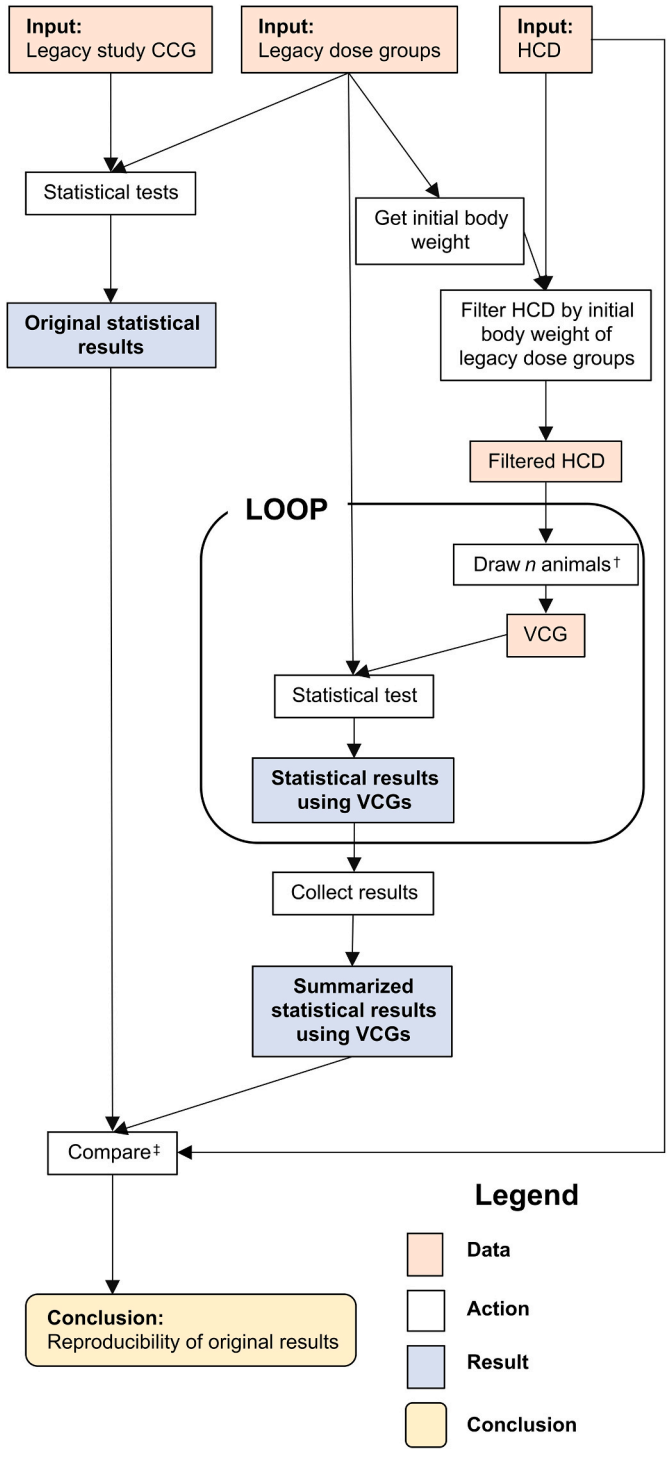

Fig. 1. Overview of the VCG creation and validation process.

○ In recovery groups, parameters were measured on day 42 ± 7.

Based on these criteria, three legacy studies were selected performed in 2021 and 2022, labeled in this article as legacy study A, B, and C, respectively. The severity of toxicity findings increased from A to C with legacy study A showing almost no effects (see Table 4 and chapter 3.4.1.1), legacy study B showing some effects but none which were found to be adverse (Table 5 and chapter 3.4.2.1), and legacy study C showing clear toxicological effects (supplementary material E Table E1 and chapter 3.4.3.1).

2.2. Data sources

Control animal data was gathered from the HCD repository for preclinical safety studies of Bayer AG, Germany. This repository contains data from animal studies completed over more than 40 years. The animals used in these studies were kept and treated in accordance with the German Animal Welfare Act and approved by the competent state authorities. All data from animal studies were recorded in SEND format (Standard for Exchange of Non-Clinical Data) (CDISC, 2022).

The data obtained from the legacy studies consisted of measurements for quantitative and qualitative parameters, plus information on the study design. The category of quantitative parameters included body weight and body weight gain, food and water consumption, hematology parameters (e.g., erythrocyte count), clinical chemistry parameters (e.g., liver transaminases, electrolytes), quantitative urine parameters, and organ weights which were measured after necropsy of the animals. Mortality of animals is technically a quantitative parameter, but since mortality in control groups is rare, it was considered as zero. Qualitative parameters consisted of clinical observations during the in-life phase of the study such as piloerection, increased salivation, and others. Qualitative histopathological findings were obtained after necropsy of the animals, including microscopic findings and macroscopic findings. An overview on the workflow how HCD was collected is provided in Fig. 2. Table 1 summarizes data domains and how HCD was used to assess possible changes in the toxicity studies.

2.3. Software

Data gathering, statistical calculations, model development, evaluation, and visualization was performed using the statistical programming software R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria). The details of R packages used, visualization methods for quality control, and imputation methods are listed in supplementary material A. The code used along with the control data is stored in Bayer's GitHub repository (Gurjanov and Vieira e Vieira, 2023). Classification of results into test-substance related or not (see below for further explanation) was done using a German version of Microsoft Excel, version 2302.

2.4. Statistical analysis of quantitative parameters

For the evaluation of the legacy study, validated statistical significance tests were used for each quantitative parameter. These significance tests are specified in the original report for every parameter and were exactly replicated for the VCG study evaluation in R. In brief, Dunnett's test was used for parameters in which a normal distribution is empirically known. For some parameters logarithmic transformation was required to achieve normal distribution. For parameters assumed to be normally distributed but with heterogenous variances, Dunnett's T3 test was used. For data assumed to be non-normally distributed, the Wilcox test with Bonferroni correction was used. A detailed description of all parameters tested within legacy studies A, B and C, and the respective statistical test used is presented in Table B1 of supplementary material B. A threshold of p-value ≤ 0.05 was used to interpret a result as "statistically significant".

To confirm that our scripts correctly recapitulate the statistical tests as in the original legacy study, we re-analyzed all parameters using the CCG and applying statistical tests with our script as described above. Out of the approximately 420 statistical tests—each sex and each dose group is compared to the respective control for an average of 70 blood and urine parameters—the results of only 3 parameters were not reproduced in legacy study A, 6 parameters in study B, and 7 parameters in study C.

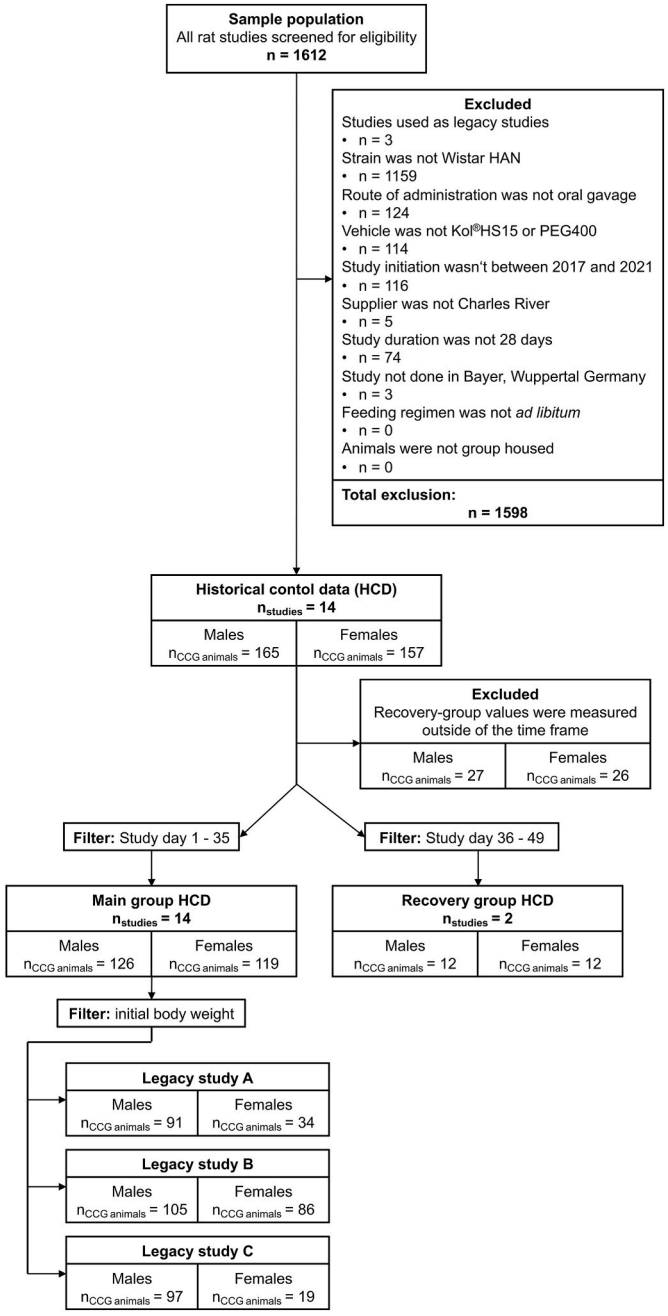

Fig. 2. Historical control data (HCD) selection flow diagram. Kol®HS15: Kolliphor® HS 15, PEG400: polyethylene glycol 400.

As all the non-reproduced statistical analyses were borderline scenarios where the resulting *p*-value of the statistical test in our reanalysis was close to 0.05, we ignored these small deviations which in most cases could be assigned to rounding issues.

2.5. VCG creation

Prior to generating VCGs from HCD, the HCD was pre-filtered to match the range of initial body weight values of the dose group animals in each study individually (see last step in Fig. 2). Out of the selected HCD animal data set, *n* animals were randomly drawn without replacement, where *n* corresponds to the number of CCG animals. This random drawing of animals from the selected initial-body-weight matched HCD animal pool was repeated 100 times independently. For each drawn animal, all measured parameters were obtained and used in each iteration of VCG. Statistical analysis was carried out independently using each VCG iteration in the same fashion as with the CCGs. The following results for each VCG repetition were obtained:

- *p*-value (for each dose group and each sex) resulting from the respective significance test.
- Post-hoc statistical power of the significance test to correctly reject at least one of the three null hypotheses, that the control group is not significantly different from the low dose, mid dose, or high dose group, respectively.
- Descriptive statistics (i.e., mean, and standard deviation of each parameter).

Table 1
Data domains and how HCD was used to compare the study outcomes after replacing the CCG with VCGs.

Data domain	Quantitative or qualitative	How the HCD was used
Mortality	Quantitative	Assumed to be zero for control animals. No reference to HCD was made.
Body weight and body weight gain	Quantitative	VCGs generated for each body weight per day and statistical outcome was compared to original result. Plots of BW and BG with respect to study day were created to assess changes over time.
Clinical observations	Qualitative	HCD was used to calculate background incidences for findings reported for treatment groups
Clinical pathology: laboratory parameters	Quantitative	VCGs generated for each parameter, significance tests calculated, and statistical outcome compared to original result.
Histopathology: macroscopic findings and microscopic findings	Qualitative	HCD was used to calculate background incidences for findings reported for treatment groups.

2.6. Preparation of results for performance assessment by study directors

The results of each of the 100 VCG iterations were summarized and presented directly to the study directors. Differences in statistical analysis for each parameter per dose group and per sex were classified as significant based on majority vote on the re-analysis with the VCG iterations. The percentage of iterations voting for the dominating class indicated the level of uncertainty. If the majority of statistical results was classified as significant, the results were further classified into three sub-categories: "*" for $p \leq 0.05$, "**" for $p \leq 0.01$, and "***" for $p \leq 0.001$. Of all VCG results for which the iterations resulted in significant changes, the predominant sub-category was shown in the final table. The final summarized parameter values which were presented to the study director are the median over all mean values and the median over all standard deviations of all 100 iterations.

In regulatory toxicology studies, HCD is used to assist in the assessment of test-substance related findings (OECD, 2008; Kluxen et al., 2021). HCD indicate the limits of normal of a measured parameter, i.e., even if a result of a parameter indicates a statistical difference, the change might still be within the limits of normal and therefore be of minor biological relevance. To replicate this established procedure also for our VCG studies, HCD reference ranges were calculated. Limits of normal ranges were displayed and calculated as follows: the lower limit of normal represented the 5th percentile of the HCD range, while the upper limit of normal represented the 95th percentile of the HCD for each parameter. The number of animals outside of the limit-of-normal range in each dose was also indicated. This procedure was applied to

each numerical parameter. The resulting values support study directors deciding the biological relevance of the statistically significant result (test-substance-relatedness) (Kluxen et al., 2021).

2.7. Calculation of background incidences for qualitative parameters

VCGs were not created to simulate qualitative animal data such as histopathological findings and clinical observations since no statistical evaluation was done in the legacy studies. Instead, HCD was used to calculate background incidences. For every noteworthy qualitative finding (see Table 4, Table 5 and Table E1 of supplementary material E for legacy study A, B, and C respectively), we have calculated the background incidence based on the HCD.

2.8. VCG overall performance assessment

The overall performance of the VCGs was assessed by the three-step approach described in this section. A graphical abstract of the assessment is shown in Fig. 3.

2.8.1. Reproducibility of statistical results

After calculating the statistical significance of differences between dose groups and controls for each parameter and sex over all 100 iterations, the VCG results were compared to the results of the legacy study. The statistical results of a parameter consist of 6 significance-test results (low, mid, and high dose, males and females, respectively). The statistical results of a parameter were classified as "reproduced" (i.e., consistent to the legacy-study result) if all statistical significance classes were correctly reproduced. Statistical results were classified into the following categories.

- Consistently significant: at least one dose group was significantly different to the CCG and the reanalysis with VCGs resulted in significant differences for the same group(s) and the differences showed the same direction, i.e., a significant increase or a significant decrease, respectively.
- Consistently non-significant: all dose groups in both the original study report and the reanalysis with VCGs found no statistically significant differences towards the control groups.
- Inconsistently significant: the reanalysis with VCGs found statistically significant differences between at least one dose group and the control group while there were none in the original study report.
- Inconsistently non-significant: the original study report found statistically significant differences between control and at least one dose group, but the group(s) were not significantly different after reanalysis with VCGs.
- Inversely significant: both the original study report and the reanalysis with VCGs found statistically significant differences between control and the same dose group(s), but the direction of the difference was dissimilar. For instance, this was the case when a significant increase was seen between a dose group and the CCGs, but the majority of the VCG iterations showed a significant decrease.

2.8.2. Reproducibility of test-substance-relatedness

All statistical results were presented to study directors or subject matter experts (SME) to assess test-substance-relatedness of statistically significant differences (Kluxen et al., 2021). When study directors or SMEs classified statistically significant differences as not related to the

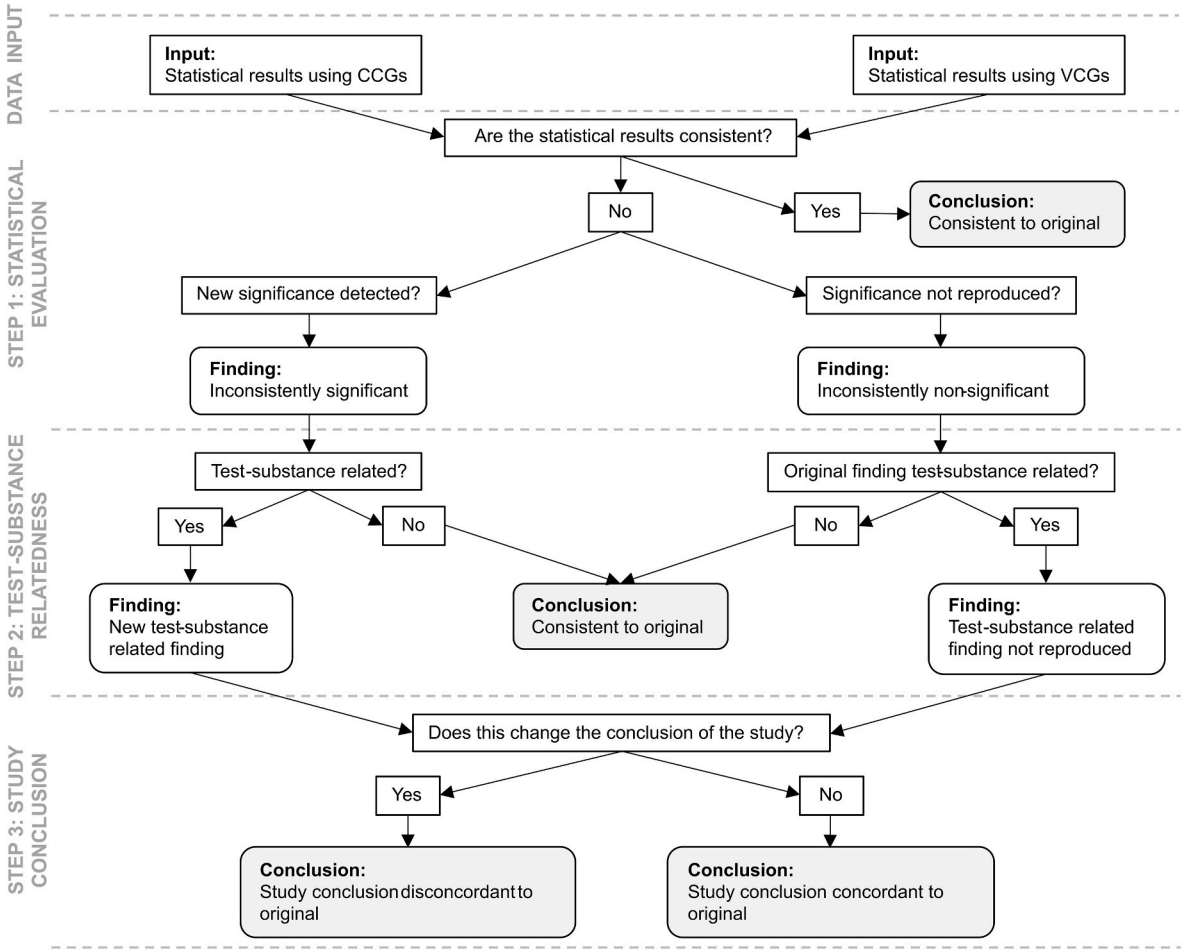

Fig. 3. VCG performance assessment flow diagram. CCG: concurrent control group; VCG: virtual control group.

test substance, we requested that at least one reason for this decision was provided. The most frequent reasons were collected from in-house toxicology reports and provided to study directors as a dropdown menu.

- Non-monotonic dose-response correlation: a value of a parameter of a lower dose group is significantly different to the control while the higher doses show no significant differences.
- Within the limits of normal (LON): while the effect in the treatment group was statistically significant, the values of the determined parameter of the treatment-group animals were still within the limits of normal of the HCD.
- No effect in correlating parameters: a parameter value was increased but since this parameter is usually highly correlated with another one, which was not affected, the change was considered implausible. For instance, a statistically significant decrease in food consumption may be disregarded if the body weight gain has not been reduced in parallel.
- Transient effect: a significant change was observed in a specific time point and disappeared over time. This is often seen in body weight or food intake.
- Only increase of parameter value is of toxicological relevance: some parameter values may be significantly decreased compared to control, but only the increase is known to be of toxicological concern. This may occur e.g., for liver damage biomarkers (FDA, 2009).
- Small effect size: the difference between control and dose may be significant but was too small to be of toxicological concern. This reason is often accompanied with other reasons, such as “within the LON” (in clinical chemistry parameters), “transient effect” (in body weight).
- Different deviations of the two sexes: a significant increase in males occurred while females showed a significant decrease or vice versa.
- Only one sex affected: effect was only visible in one sex. This reason is usually accompanied by the argument of a “small effect size”.
- Others: if none of the reasons above were suitable, the SME could provide other reasons as free text into another column of the Excel sheet.

2.8.3. Reproducibility of the study conclusion

Statistically significant differences classified as test-substance related were considered as noteworthy findings (see results section Table 4, Table 5, and Table E1 of supplementary material E). The goal of a 4-week repeated dose toxicity study in rodents is to set a dose limit for upcoming clinical trials (ICH, 2009). SMEs derived the overall study conclusion, in a non-blind setup, from noteworthy findings based on:

- NOAEL: no observed adverse effect level, i.e., the highest dose where no adversity is detected (Palazzi et al., 2016; Baird et al., 2019).
- Maximum tolerated dose: defined as the highest dose of a test substance that does not cause death or other unacceptable side effects or toxicities by causes other than carcinogenicity (Gad, 2023)
- Target organs: organs are adversely affected by the test substance. Identifying target organs is an integral part of a toxicity study (EMA, 2010).
- Monitorability: Identification of biomarkers for early detection of adverse effects available in the clinical setting (EMA, 2010).
- Reversibility: the ability of an organ, tissue, or a measured parameter to recover and return to the normal state after withdrawal of the test substance. Assessed using recovery groups in toxicity studies (EMA, 2010).

3. Results

The aim of this study was to investigate the impact of replacing CCG animals of 4-week repeated dose toxicity studies in rats (i.e., legacy studies) by VCGs. The performance of VCG was evaluated based on their capacity to replicate the findings and results of the legacy studies. Repeated dose studies were chosen as they are the most common type of study in toxicological safety assessment. This section presents the results, structured as follows:

First, we present the HCD resulting from our data-selection process. HCD were used to derive VCGs. Our objective was to faithfully reproduce the entire decision-making process of a toxicologist. We summarized this process into three main steps: statistical validation, interpretation of statistical results, and deriving a conclusion based on these results. Hence, the following sections comprise three parts: (i) we present the results of the statistical analyses comparing CCGs with VCGs; (ii) we present the findings classified as toxicologically relevant by subject matter experts, and compare them to the toxicologically relevant findings of the legacy study; and (iii) based on the test-substance related finding, we discuss whether the changes in the findings resulted in a change of the study conclusion from a toxicologist’s perspective.

Additional information can be found in the provided supplementary material. An overview of the material is provided below.

- Supplementary material A describes data curation and validation procedure (see methods section 2.3, 2.4, and 2.5).
- Supplementary material B lists all quantitative parameters with the corresponding statistical significance test for detection of significant differences (see methods section 2.8.1).
- Supplementary material C shows the missing data counts for each parameter (see results section 3.1).
- Supplementary material D shows the post-hoc statistical power results for each parameter (see results section 3.2.1).
- Supplementary material E shows the table of noteworthy findings for legacy study C (see results section 3.4.3).

3.1. Generating VCGs from HCD

To generate VCGs, we first built a matched HCD pool of animals by collecting animal data from legacy studies with the same filtering criteria used to select legacy studies A, B and C (see section 2.1). In brief, we selected animals in the HCD pool used for generating VCG animals who closely resemble the animals used in the respective legacy study. They share the same genetic background of the species, strain, and breeding facility, were treated for the same duration, using the same route of administration and vehicle, and were handled following the same procedures, in the same laboratory. To limit the effect of a potential genetic drift between animals in the CCG in the legacy studies A, B and C and the animals selected from the HCD for the VCGs (Steger-Hartmann et al., 2020), we also limited the study starting date in the HCD to 2017. In total, individual subject data from 126 male and 119 female animals from 14 legacy studies were extracted and form the HCD pool (Fig. 2). Additionally, 12 male and 12 female animals were used as VCG for analyzing the recovery group of legacy study B.

Body weight is the only quantitative parameter measured in the dose group animals prior to test substance application since body weights are usually used for animal purchase (as a surrogate for animal age), for randomizing animal allocation among experimental groups, and for calculating the amount of test substance to be administered per animal (OECD, 2008). Therefore, before building the VCG for a legacy study, we further filtered the HCD to match the initial body weight range of the dose group animals by animal sex. Therefore, no animal in the HCD pool was lighter or heavier than the lightest and heaviest animal in the dose groups of each legacy study, respectively. This reduced the animal pool to 125, 191, and 116 animals for legacy study A, B, and C respectively (see Fig. 2). From this pool of animals, the same number of animals of each sex as in the CCG in the original study were randomly sampled without replacement. All parameter measurements per animal were extracted into the VCG to ensure that possible correlations between different parameters were kept.

Not all parameters were measured in all animals leading to missing

data in VCG. Most parameters exhibited less than 20% of missing values. For the parameters which exceeded 20% of missing values, we were able to provide explanations. The number of missing data as well as the rationale for observed data gaps can be found in supplementary material C.

To increase the robustness of our analysis with VCGs, the animal sampling process was repeated 100 times and we computed the significance test *p*-value of the comparison between each dose group and VCG controls using the same statistical methods as in the original study report using CCG. Results from all 100 iterations were summarized in an ensemble-method fashion which has been described in method section 2.6 (Opitz and Maclin, 1999).

3.2. Reproducibility of statistical results

To test whether results and conclusions from the original study reports can be reproduced when legacy studies using a CCG are reanalyzed using VCGs, we implemented a three-step process that simulates the analysis process commonly performed in animal toxicity studies.

As a first step, we evaluated the performance of the reanalysis with VCGs in reproducing statistical significance and the direction of changes, i.e., decrease or increase observed in the original study report between dose and control groups on quantitative parameters. The reproducibility of statistical results is shown in Table 2. Each parameter provides 6 statistical results for low, mid, and high doses for both males and females. We have defined the statistical results of a parameter as correctly reproduced only if all statistical results in each dose group and sex were consistent between legacy study and VCG. In legacy study A, B and C, we were able to reproduce the results of 69%, 60%, and 51% of the quantitative parameters with VCGs. We have further separated this number in "consistently significant" and "consistently non-significant". If at least one dose group was significantly different to the CCG for a parameter, we classified this result as "consistently significant" only if all significant differences were reproduced exactly the same way as observed for the CCGs. We were able to achieve this only in 22% of all cases. This means that in 78% of all parameters, statistically significant differences either appeared in new dose groups—apart from the groups which were significantly already in the legacy study—or significant findings were non-significant after replacing CCGs with VCGs. A "consistently non-significant" result means that neither the original results nor our replicated ones with VCGs showed any statistically significant differences. In average this was achieved for 77% of the non-significant results. In the remaining parameters, at least one group was significantly different to the control after replacing the CCG with VCGs. In summary, the statistical results from the original analysis with CCG were not fully reproducible in the reanalysis with VCGs.

The performance in reproducing CCG statistical results was not affected by missing data in VCGs animals, as imputation of missing values did not improve reproduction of statistical results (results shown in supplementary material A).

Table 2
Summary of statistical results after replacing CCGs with VCGs. A parameter consists of up-to 6 statistical results (control vs low, mid, and high dose for males and females respectively). The percentage of inconsistent classes may overlap since parameters may contain both inconsistently non-significant as well as inversely significant classes. "Consistently significant" is defined as "at least one dose group is significantly different to the control in the legacy study and the VCG was able to reproduce the same significant classes while non-significant differences remained non-significant". "Consistently non-significant" is defined as "none of the dose groups in both legacy-study results and VCG results have any statistically significant differences to the control". "Inconsistently significant" is defined as "there is at least one statistically significant difference between VCG and dose groups while there was none in the legacy study". "Inconsistently non-significant" is defined as "there was at least one statistically significant class between CCG and dose group which was not reproduced after replacing the CCG with VCGs". "Inversely significant" is defined as "there was at least one statistically significant difference between CCG and dose group which was reproduced by VCGs, however the direction of the difference differs".

	Legacy study A	Legacy study B	Legacy study C[a]
n parameters	108	107	115
n significantly different parameters ($p \leq 0.05$)	24	33	51
n not-significantly different ($p > 0.05$)	84	74	64
Reproducibility by VCGs			
Consistently significant	3 out of 24 **(12%)**	10 out of 33 **(30%)**	13 out of 51 **(25%)**
Consistently non-significant	71 out of 84 **(85%)**	54 out of 74 **(73%)**	46 out of 64 **(72%)**
Inconsistently significant	13 out of 84 (15%)	20 out of 74 (27%)	18 out of 64 (28%)
Inconsistently non-significant	21 out of 24 (88%)	23 out of 33 (70%)	38 out of 51 (75%)
Inversely significant	0 out of 24 (0%)	0 out of 33 (0%)	1 out of 51 (2%)

[a] The female high-dose group of legacy study C was prematurely sacrificed. Therefore, blood values were not measured in the specified time window and body weight values are only partially available.

3.2.1. Post-hoc power calculations

To expand our statistical investigation on the performance of replacing CCG by VCGs as control groups, we performed post-hoc power calculations for the statistical tests of all quantitative parameters measured in the legacy studies. Calculating power for statistical analysis determines the sample size needed to correctly reject the null hypothesis at a given effect size. Low statistical power thus reduces the reliability of a bioassay (Ioannidis, 2005; Bonapersona et al., 2021). There are two types of statistical power in tests which account for multiple comparisons: the statistical power needed to correctly reject the null hypotheses of all groups, and the statistical power needed to correctly reject the null hypothesis of at least one group. We computed the latter since one statistically significant result suffices to raise awareness of the SME and triggers the next step of assessing the finding for test-substance-relatedness.

Our post-hoc statistical power analysis identified the statistical tests to be markedly underpowered (see Figure D1-D3 in supplementary material D). In at least one sex group more than 95 out of the 100 analyses with VCG iterations a statistical power of 0.8 or lower was found. While some statistical tests for parameters were sufficiently powered in one study, they were still underpowered in the other two studies. Consequently, no statistical test for any parameter consistently achieved a sufficient power of $\geq$0.8 across all studies. These results suggests that for non-significant results there is an over 20% probability that the null hypothesis was incorrectly accepted. Similar results were observed when statistical power calculations on the original analysis using CCG were done.

3.3. Reproducibility of test-substance-relatedness

Identifying statistically significant differences between dose and controls is not sufficient to determine biological relevance and expert knowledge is needed to correctly define test-substance-relatedness of findings. Therefore, we presented the statistical results from the reanalysis of the legacy studies after replacing CCG with VCGs to an SME, who assessed test-substance-relatedness of findings. The goal of the SME when preparing the assessment is to identify possible reasons to why a statistically significant difference between groups is not relevant and can therefore be disregarded. In average, 44% of all parameters where values found to be statistically significant were classified as not related to the test substance (66% for study A, 40% for study B, 29% for study C). The reasons for these decisions are summarized in Table 3.

In the following section, we highlight parameters where statistical

Table 3
Frequency of reasons provided to explain why a statistically significant finding was classified "not related to test substance". Note that the percentages do not add up to 100% since multiple reasons may apply.

Reasons why test-substance-relation was discarded	Frequency
Non-monotonic dose-response correlation	42%
Small effect size	31%
Within limits of normal (LON)	30%
Only one sex affected	19%
Transient effect	19%
Others	9%
Only increase of parameter value is of toxicological relevance	7%
No effect in correlating parameters	3%
Different direction of deviation in the two sexes	1%

results were not reproduced by the analysis with VCGs and their assessment by the SME. We especially focus on inconsistently significant parameters found to be test-substance related by the SME. We also focus on inconsistently non-significant parameters that were found to be test-substance related in the original study report.

3.3.1. Study A

Out of the 108 parameters examined, 34 parameters were inconsistent with the statistical results of the original data. 13 parameters had inconsistently significant statistical findings compared to the original result and 5 of them were found to be test-substance related (GGT, glucose, relative liver weight, absolute liver weight, food consumption on week 1). 15 parameters had inconsistently non-significant results compared to the original result and 4 of these parameters were considered as test-substance related in the original result (bilirubin and water consumption on week 2, 3, and 4). In summary, the table of noteworthy findings for legacy study A in our reanalysis with VCGs is significantly different than that from the original study report (Table 4).

3.3.2. Study B

Out of the 107 parameters examined, 43 discrepancies were found. From these 43 discrepant parameters, 20 were inconsistently significant in the statistical results and in 11 of these parameters test-substance-relatedness could not be excluded (alanine aminotransferase, basophils, bilirubin, chloride, GGT, glucose, urine protein/creatinine ratio, and body weight on several days). Notably, two noteworthy new findings were observed in the recovery group: glucose and protein/creatinine ratio. Similarly, 20 of the 43 discrepant parameters were inconsistently non-significant in their significance tests, out of which test-substance-relatedness for one parameter was not disregarded in the original study report (serum protein). Therefore, our table of noteworthy findings for legacy study B after our reanalysis with VCGs shows significant differences to the original study report where CCG was used for analysis (see Table 5).

3.3.3. Study C

In this study, 115 different quantitative parameters were monitored and 57 of them had inconsistent statistical results after replacing CCG with VCGs. From these inconsistent parameters, 18 were inconsistently significant in the statistical results, from which 9 were considered test-substance related (chloride, glucose, relative liver weight, serum protein, body weight on day 2 and 6, food consumption on week 1, 3, and 4). Also, 38 parameters showed inconsistently non-significant in their statistical results, from which test-substance-relatedness for 5 parameters were not disregarded in the original study report (absolute thymus weight, urine protein/creatinine ratio, serum protein, absolute liver weight). In summary, we found significant changes in the table of noteworthy findings after reanalyzing legacy study C with VCGs. Due to large size, this table is not shown in the main body of the article. Please refer to Table E1 in supplementary material E.

3.4. Study conclusions

The tables of noteworthy findings for each study reanalyzed with VCGs were again submitted to the study director or SMEs. The noteworthy findings were used by the study director to draw the conclusion for the study after reanalysis with VCGs. It is important to stress, that the subsequent comparison between the assessment based on CCGs and the subsequent based on VCGs was performed in a non-blinded fashion. Prior knowledge on the test substance, particularly the pharmacological mode of action, is essential to differentiate excessive pharmacological effects and off-target toxicity (Baird et al., 2019).

3.4.1. Legacy study A

3.4.1.1. Original study conclusion. In the 4-week toxicity study with daily oral gavage, no mortality, no clinical observations during the in-life phase, and no histopathological findings were noted that were regarded as test-substance related. Changes of four quantitative parameters (water intake, blood glucose, total bilirubin, and absolute plus relative liver weights) were found to be noteworthy (Table 4). While these findings were associated with the treatment, they were not regarded to be adverse. Therefore, the highest dose was determined as NOAEL.

3.4.1.2. Study conclusion after replacing the CCG with VCGs. While the increased water intake was not detected after replacing the CCG animals with VCGs, the findings in glucose were additionally visible in females starting from the lowest dose. Regarding relative liver weight, only mid dose group and high dose group showed an increase whereas with VCGs

Table 4
Legacy study A table of noteworthy findings (original and with VCG/HCD). CCG: concurrent control group, VCG: virtual control group, M: males, F: females, HCD: historical control data. LD: low dose, MD: mid dose, HD: high dose.

Mortality							
None							
Clinical findings							
None							
Quantitative parameters							
Original noteworthy findings with CCG				**Noteworthy findings after replacing CCG with VCGs**			
Parameter name	**Increase (+) decrease (−)**	**Sex (M/F)**	**Starting dose**	**Compared to CCG**	**Increase (+) decrease (−)**	**Sex (M/F)**	**Starting dose**
Water consumption	+	M + F	HD	not reproduced			
Glucose (whole blood)	–	M	HD	already at LD	–	M/F	HD/LD
Total bilirubin	+	F	LD	not reproduced			
gamma-Glutamyl transferase				new	+	M + F	LD
Absolute liver weight	+	F	MD	consistent	+	F	MD
Relative liver weight	+	F	MD	already at LD	+	F	LD
Noteworthy pathological findings							
None							

Table 5
Legacy study B table of noteworthy findings (original and with VCG/HCD). CCG: concurrent control group, VCG: virtual control group, M: males, F: females, HCD: historical control data. LD: low dose, MD: mid dose, HD: high dose.

Mortality							
None							
Clinical findings							
None							
Quantitative parameters							
Original noteworthy findings using the CCG				**Noteworthy findings after replacing CCG with VCGs**			
Parameter name	**Increase (+) decrease (−)**	**Sex (M/F)**	**Starting dose**	**Compared to CCG**	**Increase (+) decrease (−)**	**Sex (M/F)**	**Starting dose**
Body weight gain (1st week)	–	M + F	HD	consistent	–	M + F	HD
Food consumption (1st week)	–	M/F	HD/MD	consistent	–	M/F	HD/MD
Water consumption	+	M + F	LD	consistent	+	M/F	MD/LD
Total bilirubin	+	M + F	MD	now in LD (M)	+	M/F	LD/MD
Chloride	–	M	MD	now in LD	–	M	LD
Protein/Creatinine ratio	+	M + F	MD	now in LD	+	M + F	LD
Protein	–	M	HD	not reproduced			
Alanine aminotransferase				new	+	F	MD
Basophils				new	+	M + F	LD
gamma-Glutamyl transferase				new	+	F	LD
Glucose (whole blood)				new	+	M + F	LD

Noteworthy pathological findings

Organ/tissue	**Finding**	**Sex (M/F)**	**Starting dose**	**Background incidences in CCG**	**Background incidences in HCD**
Thyroid gland	Follicular cell hypertrophy	M + F	LD	M: 0 out of 10 (0 %) F: 0 out of 10 (0 %)	M: 11 out of 116 (9 %) F: 2 out of 109 (2 %)
Thymus	Tingible body macrophages, increased	M + F	HD	M: 0 out of 10 (0 %) F: 0 out of 10 (0 %)	M: 0 out of 116 (0 %) F: 0 out of 109 (0 %)
Harderian gland	Degeneration/regeneration	M + F	HD	M: 0 out of 10 (0 %) F: 0 out of 10 (0 %)	M: 1 out of 116 (1 %) F: 1 out of 109 (1 %)
Spleen	Extramedullary hematopoiesis	M + F	LD	M: 0 out of 10 (0 %) F: 1 out of 10 (10 %)	M: 22 out of 116 (19 %) F: 35 out of 109 (32 %)

the low dose animals were affected, too. Of note is that the significant change in bilirubin in female high dose was not reproduced with VCGs. However, a noteworthy new finding was observed: a significant increase in GGT was now present throughout all doses in all sexes. GGT is a sensitive—but not specific—marker for cholestatic drug induced liver injury (Trost, 2014; Robles-Diaz et al., 2015). The GGT values in all doses however were only slightly above the 1 x upper limit of normal. Given the small increases and the lack of concomitant histological findings, the increases in GGT were not considered as relevant. The conclusion of the legacy study A therefore remains unchanged: the NOAEL was still considered to be the highest dose of the study but some potential changes in liver parameters were seen.

3.4.2. Legacy study B

3.4.2.1. Original study conclusion. Study B investigated possible effects of a potential anti-cancer drug. Unlike the other two legacy studies, this study included a 2-week recovery phase to assess the reversibility of potential noteworthy findings. Cancer therapy is often accompanied with marked side effects resulting frequently in poor general condition of the animals. Therefore, instead of a NOAEL usually the severely toxic dose (STD_{10}) is the threshold of interest in rodents, i.e., the dose at which 10% of animals show marked clinical findings and/or body weight loss leading to premature sacrifice (Maziasz et al., 2010; ICH, 2010; FDA, 2020; Hukkanen et al., 2023).

In study B, no such effects on mortality or clinical findings were observed after repeated dosing. Some in-life parameters were affected, such as a decreased body weight gain during the first week of dosing accompanied by transiently reduced food consumption and an increased water intake. Several parameters of clinical chemistry (total bilirubin and protein), and urinalysis (protein/creatinine ratio) showed deviations from control means. Histopathological evaluation revealed findings of the spleen, thymus, thyroid and Harderian gland. Taking the test-substance related changes and findings into account, the STD_{10} was considered to exceed the highest dose (Table 5).

3.4.2.2. Study conclusion after replacing the CCG with VCGs. Replacing CCGs with VCGs altered the list of noteworthy findings only slightly. While the effects in body weight, food consumption and water consumption were consistent, the increase in total bilirubin, the decrease in chloride, and the increase in urine protein/creatinine ratio was now detectable at the lowest dose. Further, the decrease in protein was not reproduced by VCGs. New findings in alanine aminotransferase, basophils, GGT, and glucose were detected. However, neither of the findings changed the STD_{10} determined in this study as none of the findings were found severe. The conclusion of the study remained therefore unchanged: the STD_{10} was considered to be above the highest administered dose.

3.4.3. Legacy study C

3.4.3.1. Original study conclusion. In this study, high dose animals showed clinical findings, a decreased body weight development and food intake resulting in the need of premature sacrifice of several animals (see Table E1 in supplementary material E). The mid dose group also showed some clinical findings of less severity and lower incidence accompanied by changes of food and water intake, several changes of clinical pathology parameters, organ weights, and histopathological findings. There was one mid dose-animal which was prematurely necropsied. However, in the histopathological examinations of this animal, no finding was detected which would conclude a connection to the test substance. Low dose animals still revealed some clinical findings at low incidences as well as an increased water intake in females. Based on the clinical and histopathological findings, the NOAEL was set to the mid dose group in this study.

3.4.3.2. Study conclusion after replacing the CCG with VCGs. Replacing the CCG with VCGs did not change the conclusions made for clinical observations, these were rather further strengthened using the background incidents for salivation, change in feces, and high stepping gait.

The test-substance related changes of several quantitative

parameters present in the original study were not reproduced using virtual controls: the decrease in food intake was only reproduced for males, not for females. The increase in serum protein in females was not reproduced. Rather, a decrease in protein was seen in both males and females (monotonic dose-response correlation was only seen in males). Of note is that a significant increase in protein/creatinine ratio was seen in the high dose in males in the original study which was, however, not reproduced using virtual controls. The finding in calcium was not reproduced. The significant decrease in chloride observed in female animals in the original study was now also present in male animals of all dose groups. The increase in absolute liver weight in females was not reproduced after replacing CCGs with VCGs but relative liver weight increases were still observed. However, assessing the whole picture of all findings including the histopathology and the clinical observations, the altered clinical pathology in the table of noteworthy findings did not change the conclusions of the study director, i.e., the NOAEL remained at the mid dose.

4. Discussion

As shown in our analyses, the reproducibility of statistical results of the measured endpoints after replacing the CCGs with VCGs is low to moderate. While 72%–85% of all results which were not significant in the original study could be reproduced using VCGs, parameter values with statistically significant findings, were only reproducible in 12%–30% of the cases. One possible reason could be the sensitivity to handling stress of certain parameter values (e.g., glucose, lactate dehydrogenase, alanine aminotransferase, aspartate aminotransferase, creatine kinase) which might result in higher variability in the HCD (Ohta et al., 2009). Another probably more important reason might be related to the statistical evaluation itself. Generally, the statistical tests are underpowered in animal studies (Sena et al., 2014; Munoz-Muriedas, 2021) and no adjustment for multiplicity is applied (i.e., an α-error of 0.05 for roughly 100 parameters x 2 sexes may lead to up to 10 parameters showing a significant change just by chance) (Kluxen et al., 2021). To address the former problem, HCD could be used to increase statistical power in animal studies (Bonapersona et al., 2021; Gurjanov et al., 2023). This would improve detection of rare findings, especially in lower dose treatment groups. But the benefits of high-powered studies in the context of regulatory toxicology remain to be assessed. The latter problem—a lack in multiplicity-adjustment for statistical outcomes—has been addressed with the proposal of shifting the interpretation from *p*-values towards effect sizes (Schmidt et al., 2016; Gigerenzer, 2018; Kluxen, 2020). Currently, the use of effect sizes is done only partly when the study directors assess a toxicological relevance of a significant effect against reference values from HCD.

The reproducibility of statistical results might be improved by a selection of suitable animals matched by more relevant endpoints such as blood parameter values, rather than initial body weight alone. However, with the current design of sub-chronic studies in rodents this is not feasible. To select suitable animals by matching blood values, blood sampling prior to start of test substance would be required, and the relatively small blood volume in rodents does not allow for sampling blood twice within such a short period. However, the concept of selecting animals by matching blood values could be explored for larger animal species, where usually pre-values are analyzed before the start of the study.

Despite the inconsistencies in the statistical evaluations, the changes in parameter values that were considered test-substance related in the original study were generally well reproduced by the VCGs. Those statistical changes in the parameter values that were not reproduced ultimately did not change the core conclusion of the toxicological test. In none of the three studies which were analyzed, the study conclusion changed after replacing the concurrent controls with the virtual controls, i.e., the NOAEL, the STD_{10}, the target organs, and the relevant clinical pathology markers for monitoring remained unchanged. The determination of the threshold doses takes all findings into consideration: quantitative changes usually need to be dose-dependent and significant, and if non-significant, these changes need to have a causal relationship with the test substance (Park and Cho, 2011). The *p*-values obtained from significance tests for quantitative parameters alone—while being a useful tool—should not be used as single criteria for decision making and for scientific reasoning (Gigerenzer, 2018; Kluxen, 2020).

A recent paper by Mecklenburg et al. (2023) confirms our observation that the threshold doses (NOAEL, STD_{10}) are relatively insensitive to changes in the controls because they are mainly determined based on severe toxicological findings in the next higher dose groups. The authors showed that the reproduction of the NOAELs in 20 analyzed 90-day toxicity studies in non-human primates was even possible after completely omitting control-group animals (Mecklenburg et al., 2023). The authors state that a limitation in their approach was the difficulty to distinguish between effects caused by the treatment vehicle from those which were caused by the test substance itself, an issue which could be addressed by using VCGs with matching vehicles.

Historical control data was identified as being useful to calculate the background incidences of qualitative parameters such as histopathological findings. For instance, the extramedullary hematopoiesis observed in legacy study B was classified as non-adverse because this phenomenon is often seen in younger rats, and more often in females than in males (Suttie, 2006; Hobbie et al., 2013). This could be already confirmed based on 245 historical control animals. More rare findings may be further confirmed by increasing the number of HCD animals.

The currently small number of animals in our HCD pool (126 male animals and 119 female animals) admittedly limits the generalization of our conclusions particularly with regard to background incidences for histology findings. The number even decreased after further filtering the animals down to match the initial body weight values of the legacy studies' dose groups, especially for females. This issue should be addressed by further analyses of how far certain study parameters could be grouped together. E.g., whether the use of certain vehicles with comparably physicochemical constitution result in similar changes or lack of changes of the parameters measured in animals. If this is confirmed, certain vehicles could be grouped together.

A sustainable implementation of VCGs will require a continuous update of the control animal data pool to avoid a genotypic or phenotypic separation of the HCD from the animals used in future studies. This might be achieved by either conducting experiments with a full set of CCGs once in a while (Golden et al., 2023), or by replacing control animals only partially (i.e., keeping sentinel animals) (Gurjanov et al., 2023). Sentinel animals might additionally alleviate the uncertainty of the study director to attribute effects to the treatment, while they might have actually been caused by sudden unnoted changes in the environment, undetected infections or handling stress through less experienced animal care takers.

Our results of the VCG performance need to be interpreted in the light of lacking data regarding general reproducibility of systemic toxicity studies. Whereas the robustness of individual analytical parameters is determined in ring trials, ethical consideration prevent the systematic analysis of intra- and interlaboratory variability for the entire study with the same test substance. For less complex studies like the local lymph node assay performed in mice to determine the hazard of skin sensitization (OECD, 2010) the variability of results for an individual test substance tested in several laboratories spans over a concentration range of three orders of magnitude (Dumont et al., 2016). Given this high variability already seen in less complex in vivo studies we reckon, that the replacement of CCGs with VCGs will contribute if at all only to a minor extent to lowering the reproducibility of study outcome.

In summary, we were able in our study to reproduce the overall study conclusions of the three 4-week rat toxicity legacy studies after replacing VCGs with CCGs. As stated above, we are aware of the limitations of our

analyses regarding data size both in terms of the number of studies analyzed as well as the size of the HCD pool, i.e., we consider our study as exploratory mainly providing a path forward in the validation of the VCG concept.

We are aware that the unblinded process where the study directors or SMEs had full knowledge of the legacy studies and the pharmacological mode of action of the respective test substances introduced a certain level of subjectivity in the overall assessment of VCG feasibility. However, we considered this procedure to be necessary in order to fully recapitulate the process by which toxicological study reports are prepared. Prior knowledge of the test substance and safety assessment is required to truly differentiate an observed effect from an observed *adverse* effect.

5. Conclusion

This study explored the ability to reproduce the analysis results of three 4-week rat toxicity studies after replacing concurrent control animals with virtual control groups (VCGs). Though, when using VCGs statistical results of around 60% of the quantitative parameters were reproducible, biological reasoning and expert knowledge used to further classify whether the statistical results can be attributed to the test substance, resulted in overall good concordance with the original study results in term of NOAEL, STD_{10}, identified target organs and relevant clinical pathology biomarkers. Future work is needed to increase the pool of HCD which can be used for VCG. Of particular interest regarding generalization of our results will be the assessment of non-rodent studies with considerably smaller group sizes. Additionally, a larger pool of HCDs may facilitate the transition from relying on *p*-values from significance tests to employing effect sizes as a criterion for test-substance relatedness.

Funding source

The described research has been partly performed under the Innovative Medicine Initiative (IMI) Enhancing TRANslational SAFEty Assessment through Integrative Knowledge Management, (eTRANSAFE) project. eTRANSAFE has received support from IMI2 Joint Undertaking under Grant Agreement No. 777365. This Joint Undertaking received support from the European Union's Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations (EFPIA).

CRediT authorship contribution statement

Alexander Gurjanov: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Carlos Vieira-Vieira:** Writing – review & editing, Supervision, Software, Investigation, Data curation. **Julia Vienenkoetter:** Writing – review & editing, Validation, Resources, Investigation, Formal analysis. **Lea A.I. Vaas:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Thomas Steger-Hartmann:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors are employees of Bayer AG. They declare that they do not have further conflicts of interest.

Data availability

Control data and code can be found on GitHub (https://github.com/Bayer-Group/VCG-study-reproducibility). Due to confidentiality reasons, dose-group data cannot be shared.

6 Acknowledgements

A big thanks to Joerg Wichard for supporting and reviewing during the development of this article and to Hannes Friedrich Ulbrich for statistical support. We would like to thank Prof. Andrea Volkamer from the Data Driven Drug Design group of Saarland University for her contributions to the conceptual development of the model presented in this paper and her data science insights. We are deeply grateful for the work of Annika Kreuchwig and Adam Zalewski for setup and maintenance of the data base from which the data used in this publication has been sourced.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2024.105592.

References

Baird, T.J., Caruso, M.J., Gauvin, D.V., Dalton, J.A., 2019. NOEL and NOAEL: a retrospective analysis of mention in a sample of recently conducted safety pharmacology studies. J. Pharmacol. Toxicol. Methods 99, 106597. https://doi.org/10.1016/j.vascn.2019.106597.

Baldrick, P., Cosenza, M.E., Alapatt, T., Bolon, B., Rhodes, M., Waterson, I., 2020. Toxicology paradise: sorting out adverse and non-adverse findings in animal toxicity studies. Int. J. Toxicol. 39 (5), 365–378. https://doi.org/10.1177/1091581820935089.

Bonapersona, V., Hoijtink, H., Sarabdjitsingh, R., Joëls, M., 2021. Increasing the statistical power of animal experiments with historical control data. Nat. Neurosci. 24 (4), 470–477. https://doi.org/10.1038/s41593-020-00792-3.

Carroll, E.E., 2016. Going GLP: Conducting Toxicology Studies in Compliance with Good Laboratory Practices. US Army Medical Department Journal.

CDISC, C.D.I.S.C, 2022. *SEND* [Online]. CDISC: CDISC. Available: https://www.cdisc.org/standards/foundational/send. (Accessed 23 September 2023).

Dumont, C., Barroso, J., Matys, I., Worth, A., Casati, S., 2016. Analysis of the Local Lymph Node Assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches. Toxicol. Vitro 34, 220–228. https://doi.org/10.1016/j.tiv.2016.04.008.

EMA, 2010. *CPMP/SWP/1042/99 Rev 1 Corr* - Guideline on Repeated Dose Toxicity* [Online]. European Medicines Agency. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-repeated-dose-toxicity-revision-1_en.pdf. (Accessed 23 September 2023).

FDA, F.a.D.A, 2009. Drug-induced Liver Injury: Premarketing Clinical Evaluation. Guidance for industry.

FDA, U.F.a.D.A, 2020. Good Review Practice: Clinical Review of Investigational New Drug Applications.

Gad, S.C., 2023. Maximum tolerated dose. In: Wexler, P. (Ed.), Encyclopedia of Toxicology, fourth ed. Academic Press, pp. 43–44. https://doi.org/10.1016/B978-0-12-824315-2.00532-7.

Gigerenzer, G., 2018. Statistical rituals: the replication delusion and how we got there. Adc. Method Pract. Psychol. sci. 1 (2), 198–218. https://doi.org/10.1177/2515245918771329.

Golden, E., Allen, D., Amberg, A., Anger, L.T., Baker, E., Baran, S.W., et al., 2023. Toward implementing virtual control groups in nonclinical safety studies: workshop report and roadmap to implementation. ALTEX-Alternative Anim. Exp. https://doi.org/10.14573/altex.2310041.

Gurjanov, A., Kreuchwig, A., Steger-Hartmann, T., Vaas, L., 2023. Hurdles and signposts on the road to virtual control groups—a case study illustrating the influence of anesthesia protocols on electrolyte levels in rats. Front. Pharmacol. 14, 1142534.

Gurjanov, A., Vieira e Vieira, C., 2023. VCG Study Reproducibility [Software]. GitHub. https://github.com/Bayer-Group/VCG-study-reproducibility.

Haseman, J.K., Huff, J., Boorman, G.A., 1984. Use of historical control data in carcinogenicity studies in rodents. Toxicol. Pathol. 12 (2), 126–135. https://doi.org/10.1177/019262338401200203.

Hobbie, K., Elmore, S.K., Kolenda-Roberts, H.M., 2013. *Spleen - Extramedullary Hematopoiesis* [Online]. National Toxicology Program U.S. Department of Health and Human Services. Available: https://ntp.niehs.nih.gov/atlas/nnl/immune-system/spleen/ExtramedullaryHematopoiesis. (Accessed 15 August 2023).

Howard, B., 2002. Control of variability. ILAR J. 43 (4), 194–201. https://doi.org/10.1093/ilar.43.4.194.

Hukkanen, R.R., Moriyama, T., Patrick, D.J., Werner, J., 2023. Toxicologic pathology forum: opinion on approaches for reporting toxic and adverse dose levels in nonclinical toxicology studies supporting the development of anticancer pharmaceuticals. Toxicol. Pathol., 01926233221146937 https://doi.org/10.1177/01926233221146937.

ICH, 2009. ICH guideline M3(R2) on non-clinical safety studies for the conduct of human clinical trials and marketing authorisation for pharmaceuticals. In: International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use.

ICH, 2010. Nonclinical evaluation for anticancer pharmaceuticals S9. In: *International Conference On Harmonization*).

Ioannidis, J.P., 2005. Why most published research findings are false. PLoS Med. 2 (8), e124. https://doi.org/10.1371/journal.pmed.0020124.

Kluxen, F.M., 2020. "New statistics" in regulatory toxicology? Regul. Toxicol. Pharmacol. 117, 104763 https://doi.org/10.1016/j.yrtph.2020.104763.

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

Maziasz, T., Kadambi, V.J., Silverman, L., Fedyk, E., Alden, C., 2010. Predictive toxicology approaches for small molecule oncology drugs. Toxicol. Pathol. 38 (1), 148–164. https://doi.org/10.1177/0192623309356448.

Mecklenburg, L., Lenz, S., Hempel, G., 2023. How important are concurrent vehicle control groups in (sub) chronic non-human primate toxicity studies conducted in pharmaceutical development? An opportunity to reduce animal numbers. PLoS One 18 (8), e0282404. https://doi.org/10.1371/journal.pone.0282404.

Munoz-Muriedas, J., 2021. Large scale meta-analysis of preclinical toxicity data for target characterisation and hypotheses generation. PLoS One 16 (6), e0252533. https://doi.org/10.1371/journal.pone.0252533.

OECD, 2008. Test No. 407: Repeated Dose 28-day Oral Toxicity Study in Rodents. https://doi.org/10.1787/9789264070684-en.

OECD, 2010. Test No. 429: Skin Sensitisation. https://doi.org/10.1787/9789264071100-en.

Ohta, Y., Kaida, S., Chiba, S., Tada, M., Teruya, A., Imai, Y., et al., 2009. Involvement of oxidative stress in increases in the serum levels of various enzymes and components in rats with water-immersion restraint stress. J. Clin. Biochem. Nutr. 45 (3), 347–354. https://doi.org/10.3164/jcbn.09-59.

Opitz, D., Maclin, R., 1999. Popular ensemble methods: an empirical study. J. Artif. Intell. Res. 11, 169–198. https://doi.org/10.1613/jair.614.

Palazzi, X., Burkhardt, J.E., Caplain, H., Dellarco, V., Fant, P., Foster, J.R., et al., 2016. Characterizing “adversity” of pathology findings in nonclinical toxicity studies: results from the 4th ESTP international expert workshop. Toxicol. Pathol. 44 (6), 810–824. https://doi.org/10.1177/0192623316642527.

Park, Y.-C., Cho, M.-H., 2011. A new way in deciding NOAEL based on the findings from GLP-toxicity test. Toxicol. Res. 27, 133–135. https://doi.org/10.5487/TR.2011.27.3.133.

Poland, C.A., Miller, M.R., Duffin, R., Cassee, F., 2014. Part. Fibre Toxicol. 11 (42) https://doi.org/10.1186/s12989-014-0042-8, 2014.

Robles-Diaz, M., Garcia-Cortes, M., Medina-Caliz, I., Gonzalez-Jimenez, A., Gonzalez-Grande, R., Navarro, J.M., et al., 2015. The value of serum aspartate aminotransferase and gamma-glutamyl transpetidase as biomarkers in hepatotoxicity. Liver Int. 35 (11), 2474–2482. https://doi.org/10.1111/liv.12834.

Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., eTOX Cases, M., et al., 2017. Legacy data sharing to improve drug safety assessment: the eTOX project. Nat. Rev. Drug Discov. 16 (12), 811–812. https://doi.org/10.1038/nrd.2017.177.

Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., Asakura, S., Amberg, A., et al., 2023. eTRANSAFE: data science to empower translational safety assessment. Nat. Rev. Drug Discov. https://doi.org/10.1038/d41573-023-00099-5.

Schmidt, K., Schmidtke, J., Kohl, C., Wilhelm, R., Schiemann, J., van der Voet, H., et al., 2016. Enhancing the interpretation of statistical P values in toxicology studies: implementation of linear mixed models (LMMs) and standardized effect sizes (SESs). Arch. Toxicol. 90, 731–751. https://doi.org/10.1177/0192623313517771.

Sena, E.S., Currie, G.L., McCann, S.K., Macleod, M.R., Howells, D.W., 2014. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. J. Cerebr. Blood Flow Metabol. 34 (5), 737–742. https://doi.org/10.1038/jcbfm.2014.28.

Steger-Hartmann, T., Kreuchwig, A., Vaas, L., Wichard, J., Bringezu, F., Amberg, A., et al., 2020. Introducing the concept of virtual control groups into preclinical toxicology testing. ALTEX-Alternative Anim. Exp. 37 (3), 343–349. https://doi.org/10.14573/altex.2001311.

Steger-Hartmann, T., Clark, M., 2023. Can historical control group data be used to replace concurrent controls in animal studies? Toxicol. Pathol. 51 (6), 361–362.. https://doi.org/10.1177/01926233231208987.

Suttie, A.W., 2006. Histopathology of the spleen. Toxicol. Pathol. 34 (5), 466–503. https://doi.org/10.1080/01926230600867750.

Trost, D.C., 2014. Hepatotoxicity. In: Lawrence Gould, A. (Ed.), *Statistical Methods for Evaluating Safety in Medical Product Developmen*t. Wiley VCH, Hoboken, New Jersey, pp. 229–270. https://doi.org/10.1002/9781118763070.ch9.

Wood, F., 2008. The CDISC Study Data Tabulation Model (SDTM): History, Perspective, and Basics. https://lexjansen.com/wuss/2009/cdi/CDI-Wood.pdf. (Accessed 9 October 2023).

Wright, P.S., Smith, G.F., Briggs, K.A., Thomas, R., Maglennon, G., Mikulskis, P., et al., 2023. Retrospective analysis of the potential use of virtual control groups in preclinical toxicity assessment using the eTOX database. Regul. Toxicol. Pharmacol. 138, 105309 https://doi.org/10.1016/j.yrtph.2022.105309.

3.3 Supplementary Material

Table of Contents

1 Methods

1.1 Software

Data gathering, statistical calculations, model development, evaluation, and visualization was performed using the statistical programming software R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria). The details of R packages used are listed in Table below.

Table A1: R packages used in the VCG qualification procedure.

Package name_version	**Usage**
parallel (R Core Team, 2021)	Divide calculations on several cores for faster computing
data.table_1.14.8 (Dowle and Srinivasan, 2021)	Fast loading and writing of large tables
tidyverse_2.0.0 (Wickham, 2017)	Efficient processing and visualizing of data
openxslx_4.2.5.2 (Schauberger and Walker, 2023)	Reading and writing of Excel (XLSX) sheets
PMCRplus_1.9.6 (Pohlert, 2022)	Pipe friendly computing of significance tests
rstatix_0.7.2 (Kassambara, 2023)	Pipe friendly calculation of the Wilcox test
effsize_0.8.1 (Torchiano, 2020)	Computing the effect size
pwr_1.3-0 (Champely, 2020)	*Post-hoc* power calculations of significance tests for sensitivity analysis
gt_0.9.0 (Iannone et al., 2023)	Visualization of tables
webshot2_0.1.0 (Winston, 2019)	Exporting of gt-tables as static images
mice_3.15.0 (van Buuren and Groothuis-Oudshoorn, 2011)	Imputation of missing data
janitor_2.2.0 (Firke, 2023)	Cleaning corrupt table entries
grid (R Core Team, 2021)	Adding text elements to plots
gridExtra_2.3 (Auguie, 2017)	Combining plots with plot titles
cowplot_1.1.1 (Wilke, 2020)	Combining several plots together

1.2 Data visualization

To characterize the data and to test for the presence of subjects with unusually high or low parameter values, each quantitative parameter of the HCD was visualized as a histogram and as a boxplot (numeric value of the parameter with respect to the study, sorted by the study

year). Values whose median value were outside of the threshold (i.e., outside of the 1.5 • interquartile range (IQR) across all studies) were highlighted in color.

1.3 Missing data imputation

1.3.1 Methods

The number of missing data in the HCD was calculated for each quantitative parameter separately. A subject in a study might miss the quantification of a parameter for several reasons related to sample and animal handling, including clotted blood samples or premature sacrifice of the animals. However, parameters could also be missing simultaneously in several subjects in a study. Reasons for multi-subject missing data might include the omission of a particular parameter in the study per decision of the study director given the test substance's mode of action and the required standard test batteries (OECD, 2008; 2018a). Since missing values in HCD are unavoidable, various imputation methods were tested and the performance of the VCGs in their ability to reproduce statistical results of the legacy study was used as a benchmark to select the most suitable imputation method:

- Available cases (i.e., no imputation): Missing data were ignored accepting the potentially smaller sample size for some parameters.
- Median imputation: missing values were replaced with the median of the non-missing values.
- Random sampling: missing values were replaced with randomly drawn non-missing values.
- Predictive mean matching (PMM): missing values were imputed with the PMM function (Kleinke, 2018).

1.3.2 Imputation results

To assess the benefits of imputing missing data, the change of VCG performance in its ability to reproduce statistical results was taken as a benchmark. For the studies the performance was as follows:

Legacy study A:

- Available cases: Statistical results of 73 out of 108 parameters (68 %) reproduced.
- Imputation median: Statistical results of 69 out of 108 parameters (64 %) reproduced.
- Imputation RS: Statistical results of 72 out of 108 parameters (67 %) reproduced.
- Imputation PMM: Statistical results of 72 out of 108 parameters (67 %) reproduced.

Legacy study B

- Available cases: Statistical results of 63 out of 107 parameters (59 %) reproduced.

- Imputation median: Statistical results of 63 out of 107 parameters (59 %) reproduced.
- Imputation RS: Statistical results of 57 out of 107 parameters (53 %) reproduced.
- Imputation PMM: Statistical results of 61 out of 107 parameters (57 %) reproduced.

Legacy study C

- Available cases: Statistical results of 61 out of 115 parameters (53 %) reproduced.
- Imputation median: Statistical results of 57 out of 115 parameters (50 %) reproduced.
- Imputation RS: Statistical results of 55 out of 115 parameters (48 %) reproduced.
- Imputation PMM: Statistical results of 60 out of 115 parameters (52 %) reproduced.

The performance of VCGs did not change substantially in any of the imputation methods used. Therefore, further performance tests (i.e., "test-substance relatedness" and "study conclusions") were only assessed for the HCD of the "available cases" scenario.

Table B1: List of statistical tests used for each parameter. dunnett: Dunnett's Exact Homogeneous Test; het_dunn: Dunnett's Exact Heterogeneous Test; u_test: Bonferroni adjusted Mann Whitney U-test; log_trans_dunnett: Dunnett's Exact Homogenous Test after logarithmic transformation; no test: no statistical test performed.

Parameter	**Parameter short name**	**Specimen**	**Significance test used**
Body weight day *n*	BW_D*n*	Body weight	dunnett
Food consumption week *n*	FC_W*n*	Food and water consumption	u_test
Water consumption week *n*	WC_W*n*	Food and water consumption	u_test
Carnitine acetyl transferase	ECOD	Liver Bio Chemie	het_dunn
Glutathione S transferase	GST	Liver Bio Chemie	het_dunn
UDP Glucuronosyltransferase	UGC	Liver Bio Chemie	het_dunn
Weight organ *X*	WEIGHT_*X*	Organ weight	dunnett
Organ weight to body weight ratio *X*	OWBW_*X*	Organ weight to body weight ratio	dunnett
Activated partial thromboplastin time	APTT	Plasma	u_test
Atypical lymphocytes	ATYP	Plasma	u_test
Basophils	BASO	Plasma	u_test
Eosinophils	EOS	Plasma	u_test
Erythrocytes	ERY	Plasma	dunnett
Fibrinogen	FIBRINO	Plasma	u_test
Hematocrit	HCT	Plasma	dunnett
Hemoglobin	HGB	Plasma	dunnett
Lymphocytes	LYM	Plasma	u_test
Mean corpuscular hemoglobin	MCH	Plasma	dunnett

Mean corpuscular hemoglobin concentration	MCHC	Plasma	dunnett
Mean corpuscular volume	MCV	Plasma	dunnett
Monocytes	MONO	Plasma	u_test
Neutrophils	NEUT	Plasma	u_test
Prothrombin time	PT	Plasma	u_test
Reticulocytes	RETCULOCYTES	Plasma	u_test
Thrombocytes	THROMNUC	Plasma	dunnett
Alanine aminotransferase	ALT	Serum	het_dunn
Albumin	ALB	Serum	dunnett
Alkaline phosphatase	ALP	Serum	het_dunn
Aspartate aminotransferase	AST	Serum	het_dunn
Bilirubin	BILI	Serum	u_test
Calcium	CA	Serum	log_trans_dunnett
Chloride	CL	Serum	dunnett
Cholesterin	CHOL	Serum	log_trans_dunnett
Creatine kinase	CK	Serum	het_dunn
Creatinine	CREAT	Serum	dunnett
Gamma glutamyl transferase	GGT	Serum	u_test
Glutamate dehydrogenase	GLDH	Serum	het_dunn
Inorganic phosphates	PHOS	Serum	het_dunn
Lactate dehydrogenase	LDH	Serum	het_dunn
Potassium	K	Serum	log_trans_dunnett
Protein	PROT	Serum	dunnett
Sodium	SODIUM	Serum	dunnett
Thyroid stimulating hormone	TSH	Serum	u_test
Thyroxine	T4	Serum	u_test
Total bilirubin	BILI	Serum	u_test
Triglycerides	TRIG	Serum	dunnett
Triiodothryonine	T3	Serum	u_test
Urea	UREA	Serum	dunnett
7-Ethoxycoumarin deethylase	ECOD	Liver Biochemistry	het_dunn
7-Ethoxyresorfurin deethylase	EROD	Liver Biochemistry	het_dunn
Creatinine excretion rate	CREATEXR	Urine	het_dunn
Protein	PROT	Urine	u_test
Protein creatinine ratio	PROCRC	Urine	het_dunn
Protein per sampling period urinary volume	PROUVOC	Urine	het_dunn
Urinary volume	VOLUME	Urine	u_test
Urine creatinine	CREAT	Urine	u_test
Urine osmolality	OSMLTY	Urine	u_test
Glucose	GLUC	Whole blood	dunnett

2 Results

2.1 Missing data

The same parameter is not measured in all studies, leading to missing data in the VCGs (see **Fehler! Verweisquelle konnte nicht gefunden werden.**). For quantitative values in general, more than 80% of most values were present while body weight was the only parameters without any missing data. The following parameters had more than 20 % of missing data:

- Hormones (T3, T4, TSH) since hormones are not regularly measured and are not included into regulatory test battery.
- Organ weights specific to animal sex (epididymis, prostate gland, testis, seminal vesicles, uterus, ovary) as an approx. 1:1 ratio of males and females is present, 50 % of missing data is a given.
- Body weight after day 28 since after termination of the dosing period, body weights are only measured on the day of sacrifice.
- Liver biochemistry parameters (CARNITAT, EROD, ECOD, GST, UGT) since these parameters are not regularly measured and are not included into regulatory test battery.

Macroscopic findings and microscopic findings have only data from 225 of 322 animals in. The reason for this is that the remaining 97 HCD animals belonged to recovery groups whose histopathology measurements were taken after day 35 (which is the cutoff for our data). Clinical observation data, i.e., data which was taken during the in-life phase of the study was included.

Table C1: Missing values of quantitative parameters.

Parameter and specimen	**Missing values**	**Total population**	**Available data [%]**
Alkaline phosphatase in serum	0	245	100
Alanine aminotransferase in serum	0	245	100
Aspartate aminotransferase in serum	0	245	100
Body weight day 1-28	0	245	100
Calcium in serum	0	245	100
Cholesterin in serum	0	245	100
Creatinine kinase in serum	0	245	100
Chloride in serum	0	245	100
Creatinine in serum	0	245	100
Glutamate dehydrogenase in serum	0	245	100
Glucose in serum	0	245	100

Potassium in serum	0	245	100
Lactate dehydrogenase in serum	0	245	100
Inorganic phosphate in serum	0	245	100
Sodium in serum	0	245	100
Triglycerides in serum	0	245	100
Urea in serum	0	245	100
Bilirubin in serum	3	245	99
Atypical lymphocytes in plasma	5	245	98
Basophils in plasma	5	245	98
Body weight day 29	5	245	98
Eosinophils in plasma	5	245	98
Hematocrit in plasma	5	245	98
Hemoglobin in plasma	5	245	98
Lymphocytes in plasma	5	245	98
Mean corpuscular hemoglobin concentration in plasma	5	245	98
Mean corpuscular volume in plasma	5	245	98
Monocytes in plasma	5	245	98
Neutrophils in plasma	5	245	98
Erythrocytes in plasma	5	245	98
Thrombocytes in plasma	5	245	98
Lymphocytes in plasma	5	245	98
gamma-Glutamyl transferase in serum	12	245	95
Albumin in serum	20	245	92
Protein in serum	20	245	92
Kidney relative weight	21	245	91
Liver relative weight	21	245	91
Spleen relative weight	21	245	91
Thymus relative weight	21	245	91
Adrenal glands absolute weight	22	245	91
Kidney absolute weight	21	245	91
Liver absolute weight	21	245	91
Spleen absolute weight	21	245	91
Thymus absolute weight	21	245	91

Mean corpuscular hemoglobin in plasma	24	245	90
Reticulocytes in plasma	24	245	90
Osmolality in urine	26	245	89
Brain relative weight	27	245	89
Heart relative weight	27	245	89
Protein creatinine ration in urine	26	245	89
Protein in urine	26	245	89
Urinary volume	26	245	89
Brain absolute weight	27	245	89
Heart absolute weight	27	245	89
Activated partial thromboplastin time in plasma	40	245	84
Food and water consumption week 2-4	40	245	84
Fibrinogen in plasma	40	245	84
Prothrombin time in plasma	40	245	84
Creatinine excretion rate in urine	46	245	81
Food and water consumption week 1	60	245	76
Body weight day 30	81	245	67
Creatinine in urine	86	245	65
Triiodothryonine in serum	85	245	65
Thyroxine in serum	85	245	65
Thyroid stimulating hormone in serum	106	245	57
Epididymis relative weight	135	245	45
Prostate gland relative weight	135	245	45
Testis relative weight	135	245	45
Epididymis absolute weight	135	245	45
Prostate gland absolute weight	135	245	45
Testis absolute weight	135	245	45
Uterus relative weight	137	245	44
Uterus absolute weight	137	245	44
Seminal vesicles absolute weight	145	245	41
Body weight day 31	153	245	38
Carnitine acetyl transferase in liver biochemistry	168	245	31
Carnitine acetyl transferase in liver biochemistry	168	245	31

7-Ethoxyresorfurin deethylase in liver biochemistry	168	245	31
Glutathione S transferase in liver biochemistry	168	245	31
UDP Glucuronosyltransferase in liver biochemistry	168	245	31
Ovary relative weight	187	245	24
Ovary absolute weight	187	245	24
Body weight day 32	202	245	18
Food and water consumption week 5	225	245	8
Protein per sampling period urinary volume	225	245	8
Body weight day 33	231	245	6
Body weight day 34-35	240	245	2

2.2 Note to supplementary material D and E

Supplementary material D illustrates the statistical power of each significance test from every iteration. The corresponding abbreviations used in the figures are available in supplementary material E. However, due to large size of the supplementary figures D1-D3, and the corresponding Table E1, they were not included in the version of this thesis. Nevertheless, they are accessible online in the Elsevier data repository associated with this publication:

Supplementary material D: https://ars.els-cdn.com/content/image/1-s2.0-S0273230024000333-mmc4.docx.

Supplementary material E: https://ars.els-cdn.com/content/image/1-s2.0-S0273230024000333-mmc5.docx.

3 References

Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr.

Dowle, M., and Srinivasan, A. (2021). "data.table: Extension of `data.frame`. R package version 1.14.0. https://CRAN.R-project.org/package=data.table.

Firke, S. (2023). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.2.0. https://CRAN.R-project.org/package=janitor.

Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., and Seo, J. (2023). gt: Easily Create Presentation-Ready Display Tables. R package version 0.9.0. https://CRAN.R-project.org/package=gt.

Kassambara, A. (2023). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2. https://CRAN.R-project.org/package=rstatix.

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology*. https://doi.org/10.1027/1614-2241/a000141.

OECD (2008). Test no. 407: Repeated dose 28-day oral toxicity study in rodents. https://doi.org/10.1787/9789264070684-en.

OECD (2018a). Test No. 408: Repeated Dose 90-Day Oral Toxicity Study in Rodents. https://doi.org/10.1787/20745788.

Pohlert, T. (2022). PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended. R package version 1.9.6. https://CRAN.R-project.org/package=PMCMRplus.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schauberger, P., and Walker, A. (2023). openxlsx: Read, Write and Edit xlsx Files. R package version 4.2.5.2. https://CRAN.R-project.org/package=openxlsx.

Torchiano, M. (2020). _effsize: Efficient Effect Size Computation_. doi: 10.5281/zenodo.1480624 (URL: https://doi.org/10.5281/zenodo.1480624), R package version 0.8.1, https://CRAN.R-project.org/package=effsize.

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. https://doi.org10.18637/jss.v045.i03.

Wickham, H. (2017). "tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Wilke, C.O. (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot.

Winston, C. (2019). webshot: Take Screenshots of Web Pages. R package version 0.5.2. https://CRAN.R-project.org/package=webshot.

4 Discussion

The aim of the virtual control group (VCG) concept is to replace concurrent control group (CCG) animals in studies and by that contribute to the 3R concept (Russel and Burch, 1959). Within the scope of this work, the objective was to assess the capability of reproducing outcomes of nonclinical toxicity studies after substituting CCGs with VCGs. The results of legacy studies served as benchmarks comprising statistical results, test-substance-relatedness, and overall study conclusions.

The objective of the VCG consortium is to create VCGs out of historical control data (HCD) from both the respective companies' own laboratories as well as external data. However, this thesis focused on VCGs generated from Bayer AG internal HCD only, as the intention was to establish a proof of concept with data which is assumed to be the most similar to the legacy studies (Grevot et al., 2023).

The VCG performance was tested on rat studies being the type of study with the most data in the internal data set. Furthermore, this work was limited to creating VCGs by resampling methodologies, i.e. randomly sampling animals from the HCD set. This was done for two reasons: (i) endpoints were measured in animals and stored in data repositories under GLP conditions. In contrast to VCGs created by simulation approaches (Steger-Hartmann et al., 2020), sampling animals directly from such repositories would still preserve the GLP status. (ii) Extracting endpoints on an animal-by-animal basis ensures the preservation of potential correlations between endpoints.

A prerequisite for evaluating the VCG performance is that harmonized data and a code framework for assessing the VCG performance is provided and that a benchmark system for evaluating the performance of VCGs against a reference is established. The results of this work suggest that study results can be well-reproduced after substituting CCGs with VCGs given that HCD resembles the legacy study in study design and value distribution to achieve good VCG performance.

4.1 The set up for assessing the VCG performance

This work aims to set a cornerstone for future development and validation of different VCG concepts. Therefore, all harmonization tools, R-software, and the statistical evaluation process used in all publications presented in the results-section were published on respective GitHub

repositories (Gurjanov et al, 2023; Gurjanov et al., 2024a; Gurjanov et al., 2024b) to ensure maximum transparency and reusability of the results.

4.1.1.1 Data harmonization

Ensuring reproducibility of the systematic VCG-performance validation requires a harmonized data set and accessible code for creating and assessing VCGs. Nonclinical toxicity data is stored under SEND structure (standard for exchange of nonclinical data) ensuring a well harmonized data structure similar across different companies. Nevertheless, the parameters themselves show a high variability in their naming, abbreviations, and units of measurement, even if only internal data sets were taken into account. Large harmonization efforts were undertaken to create a data set usable for VCG creation and evaluation and “dictionaries” created consisting of a large variety of different names, abbreviations, and units which are mapped to one parameter term respectively. They were created on the shared VCG data collection provided by the VCG consortium so that their applicability and sustainability is ensured for external data as well.

4.1.1.2 Statistical test battery set-up

Another requirement for VCG-performance-validation is the assessment of the reproducibility of the statistical evaluation procedure, which was performed by in-house GLP-approved software provided by Xybion Digital Inc (Xybion, 2024). As this test battery is not accessible to the public it was reproduced using R software environment. For each endpoint, a specific statistical test was assigned ensuring the comparability across companies by standardizing the methodology. While alternative statistical tests could be employed (see 4.2.2.1) the objective of this work was to closely mirror the established internal test battery. Statistical outcomes reproduced with R were compared to the original results shown in study reports. These statistical outcomes of the internal test system were reproducible in 96% of all cases by the test system created with R. A small fraction of the resulting *p*-values deviated slightly from the original result, which is assumed to be caused by rounding issues.

4.1.1.3 Missing values

Setting up a framework for VCG-performance assessment revealed many missing values in the data set in both, study design parameters and endpoints. If data points are missing completely at random it might be reasonable to fill the gaps through imputation methods (Donders et al., 2006) some of which were examined in the supplementary material of section 3.3. The results of this section suggest that the VCG performance was not influenced notably after imputing missing values in HCD, although imputation could still be beneficial if applying more advanced analysis and simulation methods that can face challenges with missing data (Emmanuel et al., 2021).

VCGs might not be suitable for studies with specific design covering parameters which are rarely measured in other studies. For instance, some liver enzyme endpoints (EROD, ECOD, GST, CAT) were measured in only 31% of all studies. If such rarely measured parameters are missing or were only measured in a small number of historical studies the number of available HCD animals might be insufficient for any further statistical analyses. Therefore, study directors and data scientists should check *a priori* whether all endpoints that are planned to be examined in a study are well covered by HCD before deciding whether VCGs should be used.

4.2 Evaluating the performance of VCGs against a reference.

Assessing the VCG performance requires the use of a benchmark for comparison, and legacy toxicity studies were chosen as the benchmark in this thesis. Reproducing the study results after substituting CCGs with VCGs is a straightforward proof of concept confirming the approach's viability. This approach comprises three parts representing the study assessment workflow of a regulatory study as shown in Figure 2 in the introduction section. In short, substituting CCGs with VCGs should result in:

- reproduced statistical results,
- reproduced test-substance-relatedness,
- and reproduced study conclusions.

4.2.1 Statistical reproducibility

The first step of the VCG performance assessment was to test the ability of VCGs to reproduce the statistical outcomes of the original toxicity study. This validation method is straightforward, easily interpretable, unbiased and is amenable to systematic high-throughput VCG validation. The statistical test-battery which was adapted from the internal GLP validated internal evaluation framework consists of univariate inferential statistical tests with one respective test being assigned to each endpoint.

VCGs were created by randomly sampling animal values from HCD. Randomly sampled values from a certain distribution result in a subset with a similar distribution as the set which was sampled from, and therefore, the closer the HCD-value distribution resembled the CCG, the better VCGs reproduced the original legacy study results. Thus, to achieve good reproducibility of statistical outcomes, HCD should either be selected in a way that they closely resemble the legacy study in its endpoint distributions, or a sampling methodology should be selected which leads to VCGs having value distributions similar to the study. The following section summarizes the approaches for improving the VCG performance which were evaluated in this thesis.

4.2.1.1 Select “more suitable” HCD

Section 3.1 describes—based on the example of different anesthesia protocols—that confounding factors can heavily distort the HCD distribution and negatively affect the VCG performance. Anesthetizing animals with CO_2 results in elevated serum-electrolyte values compared to animals anesthetized with isoflurane, and subsequentially, removing CO_2-anestetized animals from the HCD resulted in improved VCG performance for legacy studies, where isoflurane was used. Additionally, the results of section 3.2 show that matching HCD to the initial body weight of the legacy-study dose groups improve the VCG performance to reproduce statistical body-weight results. Matching by body weight is often a surrogate for age for small animals, where the exact date of birth is often not available. Matching by age—given that this information is provided—is advisable as many physiological parameters correlate with this parameter (Wolford et al., 1987; McCutcheon and Marinelli, 2009).

Understanding the underlying structure of HCD and the study-design parameters affecting animal parameters should be a constant strive in the VCG-project. Aligning HCD to a study by design parameters is a non-biased way for obtaining VCGs which truly represent the study-animal population and improves the performance of VCGs in reproducing statistical results (Wright et al., 2023). To reduce the frequency of mismatches—such as using animals anesthetized with CO_2 or using excessively young/old animals—and to enhance the validity of HCD in future studies, a collaborative verification process between study directors, subject matter experts and data scientists should be established. For instance, an automatic screening process to select suitable HCD followed by manual review by the involved parties could ensure the alignment of study data with suitable historical controls and therefore refine the accuracy and relevance of findings.

Aside from age, body weight and study design, which were discussed as matching parameters in this thesis, other parameters might be explored for HCD matching in future, such as blood and urine parameters. This matching procedure requires however that the respective endpoints are taken prior to study initiation. The results presented in this thesis are based on rodent data and due to small volume, blood is not drawn from rodents at the beginning of a sub-chronic study. Nevertheless, larger animals do not have this restriction (dogs, non-human primates, pigs), allowing for additional matching possibilities. Matching by physiological parameters might result in animals being more similar to those used in the study and lead to better statistical reproducibility. One possible drawback here could be the large multitude of blood parameters compared to a potentially relatively small number of animals obstructing further matching procedures. It needs to be tested whether HCD animals can be found which truly match legacy-study animals in all endpoints or whether only a selected subset of relevant endpoints should be used for matching. Another possibility would be to match animals using

machine-learning methods—such as *k* Nearest Neighbors (Peterson, 2009)—to find the closest neighbors to the study animals in the feature space.

4.2.1.1.1 Limited data availability problem

A disadvantage of rigorous matching of HCD to the study may carry the risk of reducing HCD size to a point where no VCGs can be created anymore. Section 3.2 shows a case where matching HCD by body weight left only one animal in the HCD set, preventing any further statistical analyses showing that VCGs fail if the HCD is too small. Additionally, Figure 2 of section 3.3 shows that, even with 1612 in-house studies at hand, most were found not eligible for HCD because they were outside of the time frame of interest (studies older than 5 years were discarded) or performed in a design different to the legacy study, leaving 14 studies (126 male and 119 female animals) as HCD. This small number of data might be especially problematic for larger animals (e.g., dogs): a quick analysis revealed that already applying relatively lenient filter criteria on 400 internal dog studies (strain, study year, study length, test facility location, route of administration) resulted in only 12 studies applicable for HCD, i.e., 36 dogs per sex—note, in regulatory studies in non-rodents, 3 animals per group per sex are required (ICH, 2009). HCD did not benefit from introducing data from external companies as each performed studies in a somewhat unique design. For instance, Bayer used predominantly a mixture of ethanol, Kolliphor®HS15, and water as a vehicle in rat studies while data from other companies suggest a predominant use of methyl cellulose. Using additional parameters such as initial body weight or blood values (if available) as matching criteria might reduce the available HCD even more.

Overcoming this obstacle would require combining data from studies of different designs. For instance, studies with varying routes of administration might be summarized into two categories: enteral (oral, oral gavage), and parenteral (intravenous, inhalation, subcutaneous, etc.).

This approach was not pursued in this thesis as the goal was to create a proof of concept using data from studies performed under maximum similar conditions. Combining data of different study designs might be controversial, as small deviations from study design and environment can lead to distorted endpoints (Vandenberg et al., 2020). Moreover, incorporating data from external companies might partially violate guideline requirements stating that HCD should originate “from the same laboratory” (OECD, 2018). Nevertheless, statistical evaluation methods exist that account for HCD being “not as good as CCG”. For instance, clinical studies frequently incorporate external data through propensity score, meta-analysis, and Bayesian methods (Rosenbaum and Rubin, 1985; Austin, 2011; Sawamoto et al., 2022). The impact of study-design parameters is analyzed and weighted accordingly so

that historical data does not impact the statistical outcomes as much as more relevant concurrent controls (Ghadessi et al., 2020). Ultimately, expert judgement by statisticians, study directors, and SMEs in dialogue with regulatory bodies might be needed to determine which study design parameters can be combined without significantly impairing endpoint variability.
Another possibility to overcome the problems of little data availability, which was out of scope in this thesis, is the introduction of simulated endpoints (Steger-Hartmann et al., 2020). Future studies might explore simulation methodologies, for VCGs which might address the missing value problem, and could be beneficial for studies falling outside of the HCD value ranges, i.e., situations where resampling methods fail. A wide array of methods can be considered for simulating VCG values, from straightforward random number generators producing normally distributed values with specified location parameters to advanced generative neural networks possibly capable of accounting all potential correlations among endpoints. The large collaboration of members of the VCG project facilitates the data gathering necessary for this purpose.

4.2.1.2 Keeping sentinel animals.

Beside rigorous HCD matching, a method to improve the performance of VCGs is presented in section 3.1: keeping a fraction of CCG animals in the set (called sentinel animals). This publication highlighted the reproducibility of electrolyte parameters, and using sentinel-animal electrolyte values as additional parameters to match HCD increased the VCG performance to 100%. Having sentinel animals is essential for a bioassay as this ensures its validity. For instance, distinguishing a test-substance related effect from a possible infectious disease would not be possible with virtual controls only, therefore, at least a small number of concurrent controls is indispensable (SOLAS, 1999; Steger-Hartmann et al., 2020).
Another advantage of sentinel animals is that they can counter the impact of potential confounders. In the case of electrolyte values, even in the presence of animals anesthetized with CO_2, the VCG performance was high. Still, it should be kept in mind that the goal of VCGs is to reduce the number of animals and to contribute to the 3R-concept. Therefore, future studies should aim to determine the minimum number of animals needed in a study to preserve the integrity of the assay, while VCGs can be used to replace the removed animals.

4.2.1.3 Weighted sampling

Matching HCD by initial body weight does not always lead to normally distributed HCD similar to the legacy study. Figure 1B of section 3.2 shows one such example, where body-weight matching resulted in a skewed distribution. After reflecting on potential ways of improving the VCG performance it became apparent that the already available body-weight density

distribution might be used for assigning statistical weights to the HCD, and by that, refine the resampling procedure.

Section 3.2 presents the performance of VCGs not sampled from HCD purely at random, but instead, where HCD data points were assigned a statistical weight based on the probability density distribution of the legacy study's dose groups. This sampling method proved to have a superior average VCG performance compared to the other methods, in particular for cases where the legacy study was on the far-ends of the HCD-distribution. This method ensures VCGs having a distribution similar to that of the legacy study in its initial body weight. Furthermore, statistical weights are adjustable so that the resulting distribution of virtual control groups can be fine-tuned. Nonetheless, this approach may lead to biased results since the VCGs are not sampled at random from the HCD anymore, possibly resulting in animals not representative to the HCD population. Additionally, this procedure allowed for repeated sampling, which means that certain animals might be overrepresented in the VCG set.

4.2.1.4 Summary

Summarizing these results, all methodologies bear potential advantages in terms of VCG performance. A prerequisite of HCD is that it should have at least as much available animals as the study. Further, HCD which closely resembles the CCG in value distributions is the main driver for good VCG performance. Matching HCD by body weight (and other blood parameters if possible) is advisable as it is an effective, non-biased way to improve the VCG-performance and keeping a small fraction of CCG animals is essential to ensure the validity of the bioassay. Weighted sampling in turn, while being effective for studies which are dissimilar to the HCD in their values, should be considered with caution as it might introduce bias.

4.2.2 Beyond statistical reproducibility: test-substance-relatedness and study conclusion

The performance of VCGs created in ways mentioned above was assessed on statistical reproducibility alone, but judging the VCG performance only by the reproducibility of univariate inferential statistical results is not sufficient. It is widely accepted in life sciences that statistical significance does not inherently mean biological relevance or evidence for treatment-related changes (Kluxen et al., 2021; Steger-Hartmann and Clark, 2023).

The legacy studies used in section 3.3 showed that from all parameters with a significant difference between control and dose group, only 50%-70% were classified as test-substance related. This thesis introduced a method to recapitulate the decision-making process of a study director in a semi-systematic way, allowing them also to provide the corresponding rationale

of their decisions. This approach was introduced with the aim of replicating the decision-making process regarding relevance of findings as accurately as possible.

Moving even further, an effect caused by the test substance is not necessarily an *adverse* effect (Baird et al., 2019). Defining a NOAEL is based on a holistic assessment comprising qualitative (histopathology, clinical observations) findings as well as quantitative (blood, urine, body/organ weight). Qualitative findings are often the main drivers in determining adversity due to their pivotal role in the assessment process, i.e., they provide the most evidence for damage caused by a test substance. Quantitative parameters in contrast can suggest adverse effects, but these indications often require additional evidence for a conclusive assessment of adversity, i.e., quantitative parameters often contribute to the adversity determination when they are a part of an interaction of several changes and/or are associated with qualitative findings (Palazzi et al., 2016; Baird et al., 2019). The results in this thesis are in line with this observation showing that most findings in quantitative parameters were likely to be seen as non-adverse if they were not associated with histopathological ones. Historical control data was used to calculate background incidences for qualitative findings and therefore, their influence is rather limited compared to the impact of HCD on quantitative findings—through the usage of VCGs—explaining the high reproducibility of the studies. Future evolvement of the VCGs might include histopathological slides of HCD for which might support pathologists in the evaluation of findings compared to the background incidences (Grevot et al., 2023), yet the necessity and applicability of this needs further exploration (Golden et al., 2023).

4.2.2.1 Alternative data-driven evaluation methods

The goal of the presented benchmark system was the careful reproduction of a toxicological evaluation process, including statistical testing and further interpretation and conclusion drawing from toxicologists, so that the original toxicity-study results can be used as a benchmark. But in future, the implementation of VCGs might also open doors for alternative statistical evaluation frameworks (Ioannidis, 2005; Kluxen, 2020; Fornacon-Wood et al., 2022). The good VCG performance on the reproducibility of study conclusions, even though only approximately 60% of statistical results could be reproduced, highlights the challenges of applying univariate inferential tests on a large number of endpoints. As discussed in section 3.3, the complexity of a sub-chronic animal study and a biological variability of endpoints measured in animals limits the ability to achieve an absolute reproduction of toxicological results. Moreover, computing significance tests for each parameter and sex leads to the so-called “multiple comparison problem”, i.e., can result in an inflated α error rate (Pocock et al., 1987; Kluxen, 2020). While alternatives to traditional univariate inferential statistics exist, such as selecting a primary endpoint for a study (Pocock et al., 1987), adjusting for multiplicity

across all tests (Pocock et al., 1987; Benjamini and Hochberg, 1995), global statistics (Pocock et al., 1987) or mixed-effect models (Hoffman et al., 2002; Huang and Walker, 2019), none of the proposed statistics have been adapted in regulatory toxicity studies.

Since it is realistic to expect some discrepancies in statistical outcomes between the original study and the study using VCGs it is important to have these results interpreted by study directors and SMEs to determine test-substance-relatedness of effects and subsequently a study conclusion. It needs to be mentioned however that the study conclusions were drawn in a non-blinded fashion, i.e., full knowledge of the previous study outcome and the test substance mode of action was given. This is an unavoidable requirement as the knowledge of the test substance is necessary to robustly distinguish NOEL from NOAEL (Baird et al., 2019). Nonetheless, this level of subjectivity might be partially addressed by introducing alternative data-driven evaluation models.

For instance, distinguishing an irrelevant statistical effect from a relevant one could be facilitated through the incorporation of effect sizes in the regulatory test framework (Schmidt et al., 2016) as they provide additional information on the difference of values between control and dose groups (Kluxen et al., 2021). In the context of the VCG project, HCD could be used to estimate effect sizes, i.e., ranges of HCD endpoint values could define a “normal” range, and effects higher than the observed range could then be interpreted as a toxic effect. This is already informally done by study directors (Kluxen et al., 2021), as seen in section 3.3, where out of all reasons for discarding a statistically significant finding, a “small effect size” was provided 31% of all cases (i.e., the values of dosed and control groups were within the relevance limits of HCD). Nonetheless, integrating HCD into an effect-size estimation could open a path for regulated standardized effect sizes and reduce the potential bias introduced by subjective decision making.

Alternatively, decision making might be supported by data-driven Bayesian models. Numerous studies recommend a shift from inferential to Bayesian statistics in nonclinical settings highlighting possible benefits (Kramer and Font, 2017; Lim et al., 2018; Kluxen, 2020; Fornacon-Wood et al., 2022; Ruberg et al., 2023):

- Introducing HCD into the statistical evaluation, the number of control-group animals could be reduced while preserving statistical power (Kramer and Font, 2017).
- Prior knowledge in form of historical control data could be integrated into the statistical evaluations directly (Kramer and Font, 2017; Lim et al., 2018; Ruberg et al., 2023).
- Different types of data (including qualitative histopathology data) can be implemented, and correlations can be accounted for. Bayesian statistics can be created as a holistic

model comprising all information in one, and by that addressing the multiple-comparison problem (Ruberg et al., 2023).

In the end, choosing an optimal statistical framework for nonclinical studies is subject to a broad discussion within the scientific community (Kluxen, 2020), but this aspect is beyond the scope of the current thesis. The statistical test battery followed by a subsequent study conclusion as presented in this work mirrors a well-established process in regulatory framework. Introducing novel data-driven models for decision-making might reduce bias introduced by subjective decision making but this shift means that the original study results can no longer serve as a benchmark. Thus, the manner in which VCGs were evaluated in this work is best suited to validate this research's hypothesis.

Replicating legacy studies in the evaluation process presented in this thesis is a logical initial step, but for further validation of the VCG concept, VCGs might run in parallel to CCGs on ongoing studies, allowing to assess the VCG performance on real-life and real-time conditions (Golden et al., 2023).

5 Conclusion

Within the scope of this work, the performance of virtual control groups (VCGs) has been tested in its ability to reproduce the results of nonclinical toxicity studies. While it was shown that VCGs generated from HCD by a resampling approach generally can reproduce the results—in particular the conclusions of toxicity study reports—the performance of VCGs to reproduce statistical results is largely dependent on the provided historical control data (HCD). Legacy studies who highly resemble the HCD in their parameters have a higher performance in reproducing statistical results. Therefore, well-performing VCGs require HCD which optimally mirrors the study at hand in study-design parameters (strain, route of administration, age, initial body weights, etc.). Apart from statistical reproducibility it was shown that VCGs generally can reproduce the overall conclusions of toxicity study reports. This was confirmed on three legacy studies, each representing one of a spanning spectrum from “no test-substance related findings”, over “some findings which were not seen as severe”, to “severe findings including premature sacrifice of animals”. This range of outcomes highlights the good performance and robustness of VCGs across varying scenarios, but future research should explore the limitations of the applicability of VCGs further to define boundaries of the VCG applicability.

This work was performed with Bayer AG internal data of sub-chronic rat studies. Collaborative data collection efforts ensure that control data from external companies will be implemented and harmonized so that potential benefits from a large set of historical control data can be explored. Ultimately, this work aims to set the cornerstone for the adaptation of VCGs in future regulatory toxicity studies.

6 Bibliography

Altholtz, L. Y., Fowler, K. A., Badura, L. L., and Kovacs, M. S. (2006). Comparison of the stress response in rats to repeated isoflurane or CO2: O2 anesthesia used for restraint during serial blood collection via the jugular vein. J. Am. Assoc. Laboratory Animal Sci. 45 (3), 17–22. https://pubmed.ncbi.nlm.nih.gov/16642965/.

Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3), 399-424. https://doi.org/10.1080%2F00273171.2011.568786.

Avila, A.M., Bebenek, I., Bonzo, J.A., Bourcier, T., Bruno, K.L.D., Carlson, D.B., et al. (2020). An FDA/CDER perspective on nonclinical testing strategies: Classical toxicology approaches and new approach methodologies (NAMs). Regulatory Toxicology and Pharmacology 114, 104662. https://doi.org/10.1016/j.yrtph.2020.104662.

Baird, T.J., Caruso, M.J., Gauvin, D.V., Dalton, J.A., 2019. NOEL and NOAEL: a retrospective analysis of mention in a sample of recently conducted safety pharmacology studies. J. Pharmacol. Toxicol. Methods 99, 106597. https://doi.org/10.1016/j.vascn.2019.106597.

Baldrick, P. (2008). Safety evaluation to support first-in-man investigations II: Toxicology studies. Regul. Toxicol. Pharmacol. 51 (2), 237–243. https://doi.org/10.1016/j.yrtph.2008.04.006.

Baldrick, P., Cosenza, M.E., Alapatt, T., Bolon, B., Rhodes, M., Waterson, I., 2020. Toxicology paradise: sorting out adverse and non-adverse findings in animal toxicity studies. Int. J. Toxicol. 39 (5), 365–378. https://doi.org/10.1177/1091581820935089.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood cancer J. 6 (9), e473. Frontiers in Pharmacology 15 frontiersin.org. https://doi.org10.1038/bcj.2016.84.

Bode, G. (2020). “Regulatory guidance: ICH, EMA, FDA,” in Drug discovery and evaluation: Methods in clinical Pharmacology, 1085–1138.

Bonapersona, V., Hoijtink, H., Sarabdjitsingh, R., Jo¨els, M., 2021. Increasing the statistical power of animal experiments with historical control data. Nat. Neurosci. 24 (4), 470–477. https://doi.org/10.1038/s41593-020-00792-3.

Carroll, E.E., 2016. Going GLP: Conducting Toxicology Studies in Compliance with Good Laboratory Practices. US Army Medical Department Journal. CDISC, C.D.I.S.C, 2022. SEND [Online]. CDISC: CDISC. Available: https://www.cdisc.org/standards/foundational/send. (Accessed 23 September 2023).

CDISC, C.D.I.S.C. (2022). SEND [Online]. CDISC: CDISC. Available: https://www.cdisc.org/standards/foundational/send (Accessed March 29th 2024).

Chair, I. (2021). 4446 Clinical Signs of Pain and Disease in Laboratory Animals [Online]. Yale University: Yale Office of Animal Research Support (OARS). Available: https://your.yale.edu/policies-procedures/guides/4446-clinical-signs-pain-and-disease-laboratory-animals (Accessed August 19, 2023 2023).

Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr.

Charan, J., and Kantharia, N. (2013). How to calculate sample size in animal studies? J. Pharmacol. Pharmacother. 4 (4), 303–306. https://doi.org10.4103/0976-500X.119726.

de Kort, M.,Weber, K.,Wimmer, B.,Wilutzky, K., Neuenhahn, P., Allingham, P., et al. (2020). Historical control data for hematology parameters obtained from toxicity studies performed on different wistar rat strains: Acceptable value ranges, definition of severity degrees, and vehicle effects. Toxicol. Res. Appl. 4, 239784732093148. https://doi.org/10.1177/2397847320931484.

Deckardt, K., Weber, I., Kaspers, U., Hellwig, J., Tennekes, H., and van Ravenzwaay, B. (2007). The effects of inhalation anaesthetics on common clinical pathology parameters in laboratory rats. Food Chem. Toxicol. 45 (9), 1709–1718. https://doi.org/10.1016/j.fct.2007.03.005.

Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T., and Moons, K.G. (2006). A gentle introduction to imputation of missing values. Journal of clinical epidemiology 59(10), 1087-1091. https://doi.org/10.1016/j.jclinepi.2006.01.014.

Dowle, M. and Srinivasan, A. (2023). Data.Table: Extension of `data.Frame`. R package version 1.14.8. Https://cran.R-project.Org/package=data.Table.

Du Sert, N. P., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., ... & Würbel, H. (2020). Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLoS biology, 18(7), e3000411. https://doi.org/10.1371/journal.pbio.3000411.

Dumont, C., Barroso, J., Matys, I., Worth, A., Casati, S., 2016. Analysis of the Local Lymph Node Assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches. Toxicol. Vitro 34, 220–228. https://doi.org/10.1016/j.tiv.2016.04.008.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 50, 1096-1121. https://doi.org/10.1080/01621459.1955.10501294.

EFSA Scientific Committee (2011). Guidance on conducting repeated-dose 90-day oral toxicity study in rodents on whole food/feed. EFSA J. 9 (12), 2438. https://doi.org/10.2903/j.efsa.2011.2438.

EMA (2010). “CPMP/SWP/1042/99 Rev 1 Corr* - guideline on repeated dose toxicity,” in European Medicines agency. [Online]. Available: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-repeated-dose-toxicity-revision-1_en.pdf (Accessed September 23 2023).

EMA (2013). ICH guideline M3(R2) on non-clinical safety studies for the conduct of human clinical trials and marketing authorisation for pharmaceuticals. 5 ed. European Union: European Medicines Agency EMA. M3(R2) Step 5 Non-clinical safety studies for conduct of human clinical trials for pharmaceuticals (europa.eu) (Accessed April 7, 2024).

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big data 8, 1-37. https://doi.org/10.1186/s40537-021-00516-9.

EPA (2000). Health Effects Test Guidelines: OPPTS 870.3050 Repeated Dose 28–Day Oral Toxicity Study in Rodents [Online]. Washington D.C., U.S.: EPA. Available: https://ntp.niehs.nih.gov/sites/default/files/iccvam/suppdocs/feddocs/epa/epa_870_3050.pdf (Accessed Apr 6, 2024 2024).

FDA (2000). IV. C. 3. a Short-term toxicity studies with rodents. Toxicological principles for the safety assessment of food ingredients: FDA Redbook. https://www.fda.gov/files/food/published/Toxicological-Principles-for-the-Safety-Assessment-of-Food-Ingredients.pdf.

FDA (2009). Drug-induced Liver Injury: Premarketing Clinical Evaluation. Guidance for industry. https://www.regulations.gov/docket/FDA-2008-D-0128. (Accessed August 8, 2023).

FDA (2020). Good Review Practice: Clinical Review of Investigational New Drug Applications. https://www.fda.gov/media/116737/download (Accessed August 8, 2023).

Firke, S. (2023). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.2.0. https://CRAN.R-project.org/package=janitor.

Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J.P., and Price, G.J. (2022). Understanding the differences between Bayesian and frequentist statistics. International journal of radiation oncology, biology, physics 112(5), 1076-1082. https://doi.org/10.1016/j.ijrobp.2021.12.011.

Gad, S. C. (1994). Routes in toxicology: An overview. J. Am. Coll. Toxicol. 13 (1), 34–39. https://doi.org/10.3109/10915819409140653.

Gad, S.C., (2023). Maximum tolerated dose. In: Wexler, P. (Ed.), Encyclopedia of Toxicology, fourth ed. Academic Press, pp. 43–44. https://doi.org/10.1016/B978-0-12-824315-2.00532-7.

Ghadessi, M., Tang, R., Zhou, J., Liu, R., Wang, C., Toyoizumi, K., et al. (2020). A roadmap to using historical controls in clinical trials–by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). Orphanet journal of rare diseases 15, 1-19.

Gigerenzer, G., 2018. Statistical rituals: the replication delusion and how we got there. Adc. Method Pract. Psychol. sci. 1 (2), 198–218. https://doi.org/10.1177/2515245918771329.

Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., et al. (2016). Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood cancer journal 6(9), e473-e473. https://doi.org/10.1038/bcj.2016.84.

Golden, E., Allen, D., Amberg, A., Anger, L.T., Baker, E., Baran, S.W., et al., (2023). Toward implementing virtual control groups in nonclinical safety studies: workshop report and roadmap to implementation. ALTEX-Alternative Anim. Exp. https://doi.org/10.14573/altex.2310041.

Greim, H., Gelbke, H., Reuter, U., Thielmann, H., and Edler, L. (2003). Evaluation of historical control data in carcinogenicity studies. Hum. Exp. Toxicol. 22 (10), 541–549. https://doi.org/10.1191/0960327103ht394oa.

Grevot, A., Boisclair, J., Guffroy, M., Hall, P., Pohlmeyer-Esch, G., Jacobsen, M., et al. (2023). Toxicologic Pathology Forum Opinion Piece: Use of Virtual Control Groups in Nonclinical Toxicity Studies: The Anatomic Pathology Perspective. Toxicologic Pathology, 01926233231224805. https://doi.org/10.1177/01926233231224805.

Gurjanov, A., Kreuchwig, A., Steger-Hartmann, T., Vaas, L., 2023. Hurdles and signposts on the road to virtual control groups—a case study illustrating the influence of anesthesia protocols on electrolyte levels in rats. Front. Pharmacol. 14, 1142534. https://doi.org/10.3389/fphar.2023.1142534.

Gurjanov, A., Vieira e Vieira, C., Vaas L. A. I., (2023). VCG Resampling [Software]. GitHub. https://github.com/Bayer-Group/VCG-resampling.

Gurjanov, A., Vieira e Vieira, C., (2024a). VCG Study Reproducibility [Software]. GitHub. https://github.com/Bayer-Group/VCG-study-reproducibility.

Gurjanov, A., Vieira e Vieira, C., Vaas L. A. I., (2024b). VCG initial body weight [Software]. GitHub. https://github.com/Bayer-Group/VCG-INITBW.

Gurjanov, A., Vieira-Vieira, C., Vienenkoetter, J. et al. (2024). Replacing concurrent controls with virtual control groups in rat toxicity studies. Regulatory Toxicology and Pharmacology, 105592. https://doi.org/10.1016/j.yrtph.2024.105592.

Hamada, C. (2018). Statistical analysis for toxicity studies. Journal of toxicologic pathology 31, 15-22. https://doi.org/10.1293/tox.2017-0050.

Haseman, J.K., Huff, J., Boorman, G.A., 1984. Use of historical control data in carcinogenicity studies in rodents. Toxicol. Pathol. 12 (2), 126–135. https://doi.org/10.1177/019262338401200203.

Hobbie, K., Elmore, S.K., Kolenda-Roberts, H.M., 2013. Spleen - Extramedullary Hematopoiesis [Online]. National Toxicology Program U.S. Department of Health and Human Services. Available: https://ntp.niehs.nih.gov/atlas/nnl/immune-system/spleen/ExtramedullaryHematopoiesis. (Accessed 15 August 2023).

Hoffman, W. P., Ness, D. K., van Lier, R. B. L.. (2002). Analysis of Rodent Growth Data in Toxicology Studies, Toxicological Sciences, 66, 2, 313–319, https://doi.org/10.1093/toxsci/66.2.313.

Horii, I. (2016). The principle of safety evaluation in medicinal drug-how can toxicology contribute to drug discovery and development as a multidisciplinary science? The Journal of Toxicological Sciences 41(Special), SP49-SP67. https://doi.org/10.2131/jts.41.sp49.

Hotchkiss, C., Brommage, R., Du, M., and Jerome, C. (1998). The anesthetic isoflurane decreases ionized calcium and increases parathyroid hormone and osteocalcin in cynomolgus monkeys. Bone 23 (5), 479–484. https://doi.org/10.1016/s8756-3282(98)00124-0.

Hothorn, L. A. (2016). The two-step approach—a significant anova f-test before dunnett's comparisons against a control—is not recommended. Communications in Statistics-Theory and Methods 45, 3332-3343. https://doi.org/10.1080/03610926.2014.902225.

Hothorn, L. A., Kluxen, F. M., and Hasler, M. (2019). Pseudo-data generation allows the statistical re-evaluation of toxicological bioassays based on summary statistics. bioRxiv, Preprint. https://doi.org/10.1101/810408.

Hothorn, L.A. (2014). Statistical evaluation of toxicological bioassays–a review. Toxicology Research 3(6), 418-432. https://doi.org/10.1039/c4tx00047a.

Howard, B., 2002. Control of variability. ILAR J. 43 (4), 194–201. https://doi.org/10.1093/ilar.43.4.194.

Howard, S. C., Jones, D. P., and Pui, C.-H. (2011). The tumor lysis syndrome. New England J. Med. 364 (19), 1844–1854. https://doi.org/10.1056/NEJMra0904569.

Huang, K., and Walker, C.A. (2019). Comparisons of statistical models for growth curves from 90-day rat feeding studies. Archives of toxicology 93(8), 2397-2408. https://doi.org/10.1007/s00204-019-02496-5.

Hukkanen, R.R., Moriyama, T., Patrick, D.J., Werner, J., 2023. Toxicologic pathology forum: opinion on approaches for reporting toxic and adverse dose levels in nonclinical toxicology studies supporting the development of anticancer pharmaceuticals. Toxicol. Pathol., 01926233221146937 https://doi.org/10.1177/01926233221146937.

Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., and Seo, J. (2023). gt: Easily Create Presentation-Ready Display Tables. R package version 0.9.0. https://CRAN.R-project.org/package=gt.

ICH (2009). Guidance on nonclinical safety studies for the conduct of human clinical trials and marketing authorization for pharmaceuticals m3 (r2). International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-m3r2-non-clinical-safety-studies-conduct-human-clinical-trials-and-marketing-authorisation-pharmaceuticals-step-5_en.pdf (Accessed Apr 7 2024).

ICH (2010). Nonclinical evaluation for anticancer pharmaceuticals S9. In: International Conference On Harmonization). Ioannidis, J.P., 2005. Why most published research findings are false. PLoS Med. 2 (8), e124. https://doi.org/10.1371/journal.pmed.0020124.

ICH (2020). Harmonised guideline: Detection of reproductive and developmental toxicity for human pharmaceuticals S5 (R3). International conference on the harmonisation of technical requirements for registration of pharmaceuticals for human use. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-s5-r3-guideline-detection-reproductive-and-developmental-toxicity-human-pharmaceuticals-step-5-revision-4_en.pdf (Accessed Apr 7, 2024).

ICH (2023). Safety Guidelines [Online]. Geneva, Switzerland: ICH. Available: https://www.ich.org/page/safety-guidelines (Accessed Apr 2, 2024).

Igl, B.-W., Bitsch, A., Bringezu, F., Chang, S., Dammann, M., Frötschl, R., et al. (2019). The rat bone marrow micronucleus test: Statistical considerations on historical negative control data. Regul. Toxicol. Pharmacol. 102, 13–22. https://doi.org/10.1016/j.yrtph.2018.12.009.

Ioannidis, J.P. (2005). Why most published research findings are false. PLoS medicine 2(8), e124. https://doi.org/10.1371/journal.pmed.1004085.

Jacob Filho, W., Lima, C. C., Paunksnis, M. R. R. et al. (2018). Reference database of hematological parameters for growing and aging rats. The Aging Male 21, 145-148. https://doi.org/10.1080/13685538.2017.1350156.

Jourdan, T. (2013). Empfehlungen der Berliner Tierschutzbeauftragten zu Score Sheets und Abbruchkriterien [Online]. Unterfranken: Regierung Unterfranken. Available: https://www.regierung.unterfranken.bayern.de/mam/aufgaben/bereich5/sg54/score_sheet_empfehlungen-der-berliner-tschb-zu-abbruchkriterien.pdf (Accessed January 2 2024).

Kacew, S. (1996). Invited review: role of rat strain in the differential sensitivity to pharmaceutical agents and naturally occurring substances. J. Toxicol. Environ. Health Part A 47 (1), 1–30. https://doi.org/10.1080/009841096161960-2840.

Kassambara, A. (2023). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2. https://CRAN.R-project.org/package=rstatix.

Keenan, C., Elmore, S., Francke-Carroll, S., Kemp, R., Kerlin, R., Peddada, S., et al. (2009). Best practices for use of historical control data of proliferative rodent lesions. Toxicologic pathology 37(5), 679-693. https://doi.org/10.1177/0192623309336154.

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. Methodology. https://doi.org/10.1027/1614-2241/a000141.

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

Kluxen, F.M., (2020). "New statistics" in regulatory toxicology? Regul. Toxicol. Pharmacol. 117, 104763 https://doi.org/10.1016/j.yrtph.2020.104763.

Kolker, E., Stewart, E., and Ozdemir, V. (2012). Opportunities and challenges for the life sciences community. OMICS A J. Integr. Biol. 16 (3), 138–147. https://doi.org/10.1089/omi.2011.0152.

Kramer, M., and Font, E. (2017). Reducing sample size in experiments with animals: historical controls and related strategies. Biological Reviews 92(1), 431-445. https://doi.org/10.1111/brv.12237.

Langford, N. J. (2005). Carbon dioxide poisoning. Toxicol. Rev. 24 (4), 229–235. https://doi.org/10.2165/00139709-200524040-00003.

Lecoq, A.-L., Livrozet, M., Blanchard, A., and Kamenický, P. (2021). Drug-related hypercalcemia. Endocrinol. Metabolism Clin. 50 (4), 743–752. https://doi.org/10.1016/j.ecl.2021.08.001.

Lim, J., Walley, R., Yuan, J., Liu, J., Dabral, A., Best, N., et al. (2018). Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities. Ther. Innovation Regul. Sci. 52 (5), 546–559. https://doi.org/10.1177/2168479018778282.

Maziasz, T., Kadambi, V.J., Silverman, L., Fedyk, E., Alden, C., 2010. Predictive toxicology approaches for small molecule oncology drugs. Toxicol. Pathol. 38 (1), 148–164. https://doi.org/10.1177/0192623309356448.

McCutcheon, J.E., and Marinelli, M. (2009). Age matters. European Journal of Neuroscience 29(5), 997-1014. https://doi.org/10.1111/j.1460-9568.2009.06648.x.

Mecklenburg, L., Lenz, S., Hempel, G., 2023. How important are concurrent vehicle control groups in (sub) chronic non-human primate toxicity studies conducted in

pharmaceutical development? An opportunity to reduce animal numbers. PLoS One 18 (8), e0282404. https://doi.org/10.1371/journal.pone.0282404.

Menssen, M. (2023). The calculation of historical control limits in toxicology: Do's, don'ts and open issues from a statistical perspective. Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 503695. https://doi.org/10.1016/j.mrgentox.2023.503695.

Munoz-Muriedas, J., 2021. Large scale meta-analysis of preclinical toxicity data for target characterisation and hypotheses generation. PLoS One 16 (6), e0252533. https://doi.org/10.1371/journal.pone.0252533.

Murphy , K. (2023). Regulating and Authorizing Medicines: A Comparison of the FDA and EMA [Online]. Xtelligent Healthcare Media, U.S.: PharmaNewsIntelligence. Available: https://pharmanewsintel.com/features/regulating-and-authorizing-medicines-a-comparison-of-the-fda-and-ema (Accessed Apr 3, 2024 2024).

Namdari, R., Jones, K., Chuang, S. S. et al. (2021). Species selection for nonclinical safety assessment of drug candidates: Examples of current industry practice. Regulatory Toxicology and Pharmacology 126, 105029. https://doi.org/10.1016/j.yrtph.2021.105029.

Nicklas, W., Homberger, F. R., Illgen-Wilcke, B., Jacobi, K., Kraft, V., Kunstyr, I., et al. (1999). Implications of infectious agents on results of animal experiments: Report of the Working Group on Hygiene of the Gesellschaft für Versuchstierkunde-Society for Laboratory Animal Science (GV-SOLAS). Lab. Anim. 33 (1), 39–87. https://doi.org/10.1258/002367799780639987.

OECD (2008). Test no. 407: Repeated dose 28-day oral toxicity study in rodents. https://doi.org/10.1787/9789264070684-en.

OECD (2016). Test No. 487: In vitro mammalian cell micronucleus test. https://doi.org/10.1787/20745788.

OECD (2018a). Test No. 408: Repeated Dose 90-Day Oral Toxicity Study in Rodents. https://doi.org/10.1787/20745788.

OECD (2018b). Test No. 453: Combined chronic toxicity/carcinogenicity studies. https://doi.org/10.1787/20745788.

OECD (2023). OECD Test Guidelines for Chemicals [Online]. Paris, France: OECD. Available: https://www.oecd.org/chemicalsafety/testing/oecdguidelinesforthetestingofchemicals.htm (Accessed Aprl 2, 2024).

Ohta, Y., Kaida, S., Chiba, S., Tada, M., Teruya, A., Imai, Y., et al., (2009). Involvement of oxidative stress in increases in the serum levels of various enzymes and components in rats with water-immersion restraint stress. J. Clin. Biochem. Nutr. 45 (3), 347–354. https://doi.org/10.3164/jcbn.09-59.

Opitz, D., Maclin, R., (1999). Popular ensemble methods: an empirical study. J. Artif. Intell. Res. 11, 169–198. https://doi.org/10.1613/jair.614.

Palazzi, X., Burkhardt, J.E., Caplain, H., Dellarco, V., Fant, P., Foster, J.R., et al. (2016). Characterizing “adversity” of pathology findings in nonclinical toxicity studies: Results from the 4th ESTP international expert workshop. Toxicologic Pathology 44(6), 810-824. https://doi.org/10.1177/0192623316642527.

Parasuraman, S. (2011). Toxicological screening. Journal of pharmacology & pharmacotherapeutics 2(2), 74. https://doi.org/10.4103%2F0976-500X.81895.

Parasuraman, S., Raveendran, R., and Kesavan, R. (2010). Blood sample collection in small laboratory animals. J. Pharmacol. Pharmacother. 1 (2), 87–93. https://doi.org/10.4103/0976-500X.72350.

Park, Y.-C., Cho, M.-H., (2011). A new way in deciding NOAEL based on the findings from GLP-toxicity test. Toxicol. Res. 27, 133–135. https://doi.org/10.5487/TR.2011.27.3.133.

PBL, P.B. (2024). Regulatory Toxicology Studies [Online]. San Francisco Bay Area 551 Linus Pauling Drive, Hercules CA 94547: PBL, Pacific Biolabs. Available: https://pacificbiolabs.com/regulatory-tox (Accessed Apr 1, 2024 2024).

Peterson, L.E. (2009). K-nearest neighbor. Scholarpedia 4(2), 1883. http://dx.doi.org/10.4249/scholarpedia.1883.

Pinches, M. D., Thomas, R., Porter, R., Camidge, L., and Briggs, K. (2019). Curation and analysis of clinical pathology parameters and histopathologic findings from eTOXsys, a large database project (eTOX) for toxicologic studies. Regul. Toxicol. Pharmacol. 107, 104396. https://doi.org/10.1016/j.yrtph.2019.05.021.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. J. chronic Dis. 29 (3), 175–188. https://doi.org/10.1016/0021-9681(76)90044-8.

Pocock, S.J., Geller, N.L., and Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials. Biometrics, 487-498. https://pubmed.ncbi.nlm.nih.gov/3663814/.

Pognan, F., Steger-Hartmann, T., Díaz, C., Blomberg, N., Bringezu, F., Briggs, K., et al. (2021). The eTRANSAFE project on translational safety assessment through integrative knowledge management: Achievements and perspectives. Pharmaceuticals 14 (3), 237. https://doi.org/10.3390/ph14030237.

Pohlert, T. (2022). Pmcmrplus: Calculate pairwise multiple comparisons of mean rank sums extended. R package version 1.9.6. Https://cran.R-project.Org/package=pmcmrplus.

Poland, C.A., Miller, M.R., Duffin, R., Cassee, F., 2014. Part. Fibre Toxicol. 11 (42) https://doi.org/10.1186/s12989-014-0042-8.b

Prior, H., Haworth, R., Labram, B., Roberts, R., Wolfreys, A., and Sewell, F. (2020). Justification for species selection for pharmaceutical toxicity studies. Toxicology Research 9(6), 758-770. https://doi.org/10.1093/toxres/tfaa081.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robles-Diaz, M., Garcia-Cortes, M., Medina-Caliz, I., Gonzalez-Jimenez, A., Gonzalez Grande, R., Navarro, J.M., et al., (2015). The value of serum aspartate aminotransferase and gamma-glutamyl transpetidase as biomarkers in hepatotoxicity. Liver Int. 35 (11), 2474–2482. https://doi.org/10.1111/liv.12834.

Rosenbaum, P. R., and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Statistician 33-39 (1), 33–38. https://doi.org/10.2307/2683903.

Ruberg, S.J., Beckers, F., Hemmings, R., Honig, P., Irony, T., LaVange, L., et al. (2023). Application of Bayesian approaches in drug development: starting a virtuous cycle. Nature Reviews Drug Discovery 22(3), 235-250. https://doi.org/10.1038/s41573-023-00638-0.

Russel, W. M. S. and Burch, R. L. . The principles of humane experimental technique. Methuen, (1959). http://117.239.25.194:7000/jspui/bitstream/123456789/1342/1/PRILIMINERY%20%20AND%20%20CONTENTS.pdf.

Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., Asakura, S., Amberg, A., et al., (2023). eTRANSAFE: data science to empower translational safety assessment. Nat. Rev. Drug Discov. https://doi.org/10.1038/d41573-023-00099-5.

Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., Cases, M., et al., (2017). Legacy data sharing to improve drug safety assessment: the eTOX project. Nat. Rev. Drug Discov. 16 (12), 811–812. https://doi.org/10.1038/nrd.2017.177.

Sawamoto, R., Oba, K., and Matsuyama, Y. (2022). Bayesian adaptive randomization design incorporating propensity score-matched historical controls. Pharmaceutical Statistics. https://doi.org/10.1002/pst.2203:

Schauberger, P. and Walker, A. (2023). Openxlsx: Read, write and edit xlsx files. R package version 4.2.5.2. Https://cran.R-project.Org/package=openxlsx.

Schenck, P. A., Chew, D. J., Nagode, L. A., and Rosol, T. J. (2006). Disorders of calcium: Hypercalcemia and hypocalcemia. Fluid, electrolyte, acid-base Disord. Small animal Pract. 4, 120–194. http://dx.doi.org/10.1016/B0-72-163949-6/50009-6.

Schmidt, K., Schmidtke, J., Kohl, C., Wilhelm, R., Schiemann, J., van der Voet, H., et al., (2016). Enhancing the interpretation of statistical P values in toxicology studies: implementation of linear mixed models (LMMs) and standardized effect sizes (SESs). Arch. Toxicol. 90, 731–751. https://doi.org/10.1177/0192623313517771.

Sena, E.S., Currie, G.L., McCann, S.K., Macleod, M.R., Howells, D.W., (2014). Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. J. Cerebr. Blood Flow Metabol. 34 (5), 737–742. https://doi.org/10.1038/jcbfm.2014.28.

Shrimanker, I., and Bhattarai, S. (2016). “Electrolytes,” in StatPearls. Treasure Island (FL); StatPearls Publishing. 2022. PMID: 31082167.

Sievert, C. (2018). plotly for R. https://plotly-r.com.

Signorell, A., and al., e.m. (2019). DescTools: Tools for descriptive statistics. R package version 0.99.28. https://cran.r-project.org/package=DescTools.

SOLAS, G. (1999). Implications of infectious agents on results of animal experiments-Oxyurina (Pinworms). Lab. Anim 33(suppl 1), 85-86. https://doi.org/10.1258/002367799780639987.

Steger-Hartmann, T. and Clark, M. (2023). Can historical control group data be used to replace concurrent controls in animal studies? Toxicologic Pathology 01926233231208987. https://doi.org/10.1177/01926233231208987.

Steger-Hartmann, T., Kreuchwig, A., Vaas, L., Wichard, J., Bringezu, F., Amberg, A., et al. (2020). Introducing the concept of virtual control groups into preclinical toxicology testing. ALTEX-Alternatives to animal experimentation 37(3), 343-349. https://doi.org/10.14573/altex.2001311.

Stephens, M.L., and Mak, N.S. (2013). History of the 3Rs in toxicity testing: From Russell and Burch to 21st century toxicology.

Stokes, A. H., Kemp, D. C., Faiola, B., Jordan, H. L., Merrill, C. L., Hailey, J. R., et al. (2013). Effects of Solutol (Kolliphor) and cremophor in polyethylene glycol 400 vehicle formulations in Sprague-Dawley rats and beagle dogs. Int. J. Toxicol. 32 (3), 189–197. https://doi.org/10.1177/1091581813485452.

Strayhorn, J.M. (2021). Virtual controls as an alternative to randomized controlled trials for assessing efficacy of interventions. BMC Medical Research Methodology 21(1), 1-14.

Suttie, A.W., (2006). Histopathology of the spleen. Toxicol. Pathol. 34 (5), 466–503. https://doi.org/10.1080/01926230600867750.

Talbot, S. R., Biernot, S., Bleich, A. et al. (2020). Defining body-weight reduction as a humane endpoint: A critical appraisal. Laboratory animals 54, 99-110. https://doi.org/10.1177/0023677219883319.

Tinawi, M. (2021). Disorders of calcium metabolism: Hypocalcemia and hypercalcemia. Cureus 13 (1), e14619. doi:10.7759/cureus.14619

Torchiano, M. (2020). _effsize: Efficient Effect Size Computation_. doi: 10.5281/zenodo.1480624 https://doi.org/10.5281/zenodo.1480624, R package version 0.8.1, https://CRAN.R-project.org/package=effsize.

Traslavina, R. P., King, E. J., Loar, A. S., Riedel, E. R., Garvey, M. S., Ricart-Arbona, R., et al. (2010). Euthanasia by CO2 inhalation affects potassium levels in mice. J. Am. Assoc. Laboratory Animal Sci. 49 (3), 316–322. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2877304/.

Trost, D.C., (2014). Hepatotoxicity. In: Lawrence Gould, A. (Ed.), Statistical Methods for Evaluating Safety in Medical Product Development. Wiley VCH, Hoboken, New Jersey, pp. 229–270. https://doi.org/10.1002/9781118763070.ch9.

Turner, P., Hickman, D., van Luijk, J., Ritskes-Hoitinga, M., Sargeant, J., Kurosawa, T., et al. (2020). Welfare impact of carbon dioxide euthanasia on laboratory mice and rats: A systematic review. Front. Vet. Sci. 7, 411. https://doi.org/10.3389/fvets.2020.00411.

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. https://doi.org/10.18637/jss.v045.i03.

Vandenberg, L.N., Prins, G.S., Patisaul, H.B., and Zoeller, R.T. (2020). The use and misuse of historical controls in regulatory toxicology: lessons from the CLARITY-BPA study. Endocrinology 161(5), bqz014. https://doi.org/10.1210/endocr/bqz014.

Verzicco, I., Regolisti, G., Quaini, F., Bocchi, P., Brusasco, I., Ferrari, M., et al. (2020). Electrolyte disorders induced by antineoplastic drugs. Front. Oncol. 10, 779. https://doi.org/10.3389/fonc.2020.00779.

White, W. J., and Cham, S. (1998). The development and maintenance of the crl: CH!! J (SD) IGSBR rat breeding System. Biol. Ref. Data CD (SD) IGS Rats., 8–14.

Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Wilke, C. O. (2020). Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.1.1. https://cran.R-project.Org/package=cowplot.

Winston, C. (2019). webshot: Take Screenshots of Web Pages. R package version 0.5.2. https://CRAN.R-project.org/package=webshot.

Wolford, S., Schroer, R., Gallo, P., Gohs, F., Brodeck, M., Falk, H., et al. (1987). Age related changes in serum chemistry and hematology values in normal Sprague-Dawley rats. Fundam. Appl. Toxicol. 8 (1), 80–88. https://doi.org/10.1016/0272-0590(87)90102-3.

Wong, D., Makowska, I. J., and Weary, D. M. (2013). Rat aversion to isoflurane versus carbon dioxide. Biol. Lett. 9 (1), 20121000. https://doi.org/10.1098/rsbl.2012.1000.

Wood, F. K., and Lou, A. (2011). The standard for the Exchange of nonclinical data (SEND): History and basics. https://www.pharmasug.org/proceedings/2011/CD/PharmaSUG-2011-CD14.pdf. (Accessed Janury 3, 2022).

Wood, F., (2008). The CDISC Study Data Tabulation Model (SDTM): History, Perspective, and Basics. https://lexjansen.com/wuss/2009/cdi/CDI-Wood.pdf. (Accessed October 9 2023).

Wright, P. S., Smith, G. F., Briggs, K. A., Thomas, R., Maglennon, G., Mikulskis, P., et al. (2023). Retrospective analysis of the potential use of virtual control groups in preclinical toxicity assessment using the eTOX database. Regul. Toxicol. Pharmacol. 138, 105309. https://doi.org/10.1016/j.yrtph.2022.105309.

Xybion (2024). XYBION All-In-One Cloud LIMS for Life Sciences, R&D, and Labs with a Full QMS Suite [Online]. 105 College Road East, Princeton, New Jersey 08540: XYBION. Available: https://www.xybion.com/ (Accessed Apr. 1, 2024).

Zhan, T., Zhou, Y., Geng, Z., Gu, Y., Kang, J., Wang, L., et al. (2022). Deep historical borrowing framework to prospectively and simultaneously synthesize control information in confirmatory clinical trials with multiple endpoints. J. Biopharm. Statistics 32 (1), 90–106. https://doi.org/10.1080/10543406.2021.1975128.

7 List of Publications

Gurjanov A., Kreuchwig A., Steger-Hartmann T., Vaas L. A. I. (2023), Hurdles and signposts on the road to virtual control groups—A case study illustrating the influence of anesthesia protocols on electrolyte levels in rats. Front. Pharmacol. 14:1142534. https://doi.org/10.3389/fphar.2023.1142534.

Gurjanov A., Vieira-Vieira C., Vienenkoetter J., Vaas L. A. I., Steger-Hartmann T., Replacing concurrent controls with virtual control groups in rat toxicity studies, Regulatory Toxicology and Pharmacology, Volume 148, 2024, 105592, ISSN 0273-2300, https://doi.org/10.1016/j.yrtph.2024.105592. (https://www.sciencedirect.com/science/article/pii/S0273230024000333)

unpublished:

Gurjanov A., Vaas L. A. I., Steger-Hartmann T., The Road to Virtual Control Groups and the Importance of proper Body-Weight Selection. This article was submitted to the journal *ALTEX – Alternatives to animal experimentation* on March 14, 2024.

Gurjanov, A., Vaas, L. A. I. and Steger-Hartmann, T. (2024) “The road to virtual control groups and the importance of proper body weight selection”, ALTEX - Alternatives to animal experimentation, 41(4), pp. 660–665. https://doi.org/10.14573/altex.2403141.