101172693 – VICT3R

Developing and implementing Virtual Control Groups to reduce animal use in Toxicology Research

– WP6 –

VCG characterization and generation (statistics)

D6.1 Identification of Key Covariates

Lead contributor:	Paolo Piraino – Organon SRL [11]
Other contributors:	Guillemette Duchateau-Nguyen – Roche [29] Jorge Lemos Portela – UPF [1] Manuel Pastor – UPF [1] Matteo Piraino – Organon SRL [11] Stefan Anca – Organon SRL [11] Arianna Capalbo – Organon SRL [11] Wiebke Dammann – TU Dortmund [5] Jörg Rahnenführer – TU Dortmund [5]
Due date:	31 August 2025
Delivery date:	17 October 2025
Dissemination level:	Public

Contents

Document History 3

Definitions 4

Executive Summary 6

Introduction 7

Sections 7

1. Covariate Analysis [UPF] 7

1.1 Methods 7

1.2 Results 12

1.3 References 18

2. Variance explained by key covariates and analysis of residuals [ORGANON] 19

2.1 Analysis objectives 19

2.2 Source Data and Definition of the Analytical Cohort 19

2.3 Data Curation, Transformation, and Pre-processing 20

2.4 Statistical Modeling Strategy 22

2.5 Results: Variance explained by key covariates 25

2.6 Effects of Key covariates 29

2.7 Variance explained by other covariates 33

2.8 References 36

3 Changepoint analysis [TUDO] 37

3.1 Changepoint analysis 37

3.2 References 38

4 Discussion 39

Conclusion 41

Appendix 42

Document History

Version	Date	Description
V 0.1	11 Aug 2025	First Version
V 0.2	27 Aug 2025	Resolution of comments in live review in WP6 Meeting
V 0.3	17 Sept 2025	Version Implementing Formal Review Comments
V 0.4	02 Oct 2025	Version for Consortium Review
V 1.0	17 Oct 2025	Final Version

Definitions

Participants of the VICT3R Consortium are referred to herein according to the following codes:

1. UNIVERSIDAD POMPEU FABRA (UPF)
2. SYNAPSE RESEARCH MANAGEMENT PARTNERS SL (SYNAPSE)
3. FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV (ITEM)
4. GRITSYSTEMS AS (grit42)
5. TECHNISCHE UNIVERSITAT DORTMUND (TUDO)
6. PHUSE CUIDEACHTA FAOI THEORAINN RATHAIOCHTA (PHUSE)
7. MEDBIOINFORMATICS SOLUTIONS SL (MBIS)
8. BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION (BSC CNS)
9. UNIVERSITAT KONSTANZ (UKON)
10. DECIPHEX LIMITED (Deciphex)
11. ORGANON SRL (Organon)
12. SYNCWORK AKTIENGESELLSCHAFT (Syncwork)
13. VRIJE UNIVERSITEIT BRUSSEL (VUB)
14. BAYER AKTIENGESELLSCHAFT (Bayer)
15. MERCK KOMMANDITGESELLSCHAFT AUF AKTIEN (MKDG)
16. AMGEN RESEARCH (MUNICH) GMBH (Amgen)
17. ASTRAZENECA UK LIMITED (AZ)
18. BASF SE (BASF SE)
19. BOEHRINGER INGELHEIM INTERNATIONALGMBH (BII)
20. BRISTOL-MYERS SQUIBB COMPANY CORP (BMS)
21. INCYTE BIOSCIENCES DISTRIBUTION B.V (Incyte BD)
22. INSTEM SCIENTIFIC LIMITED (Instem)
23. IPSEN INNOVATION SAS (IPSEN)
24. JANSSEN PHARMACEUTICA NV (Janssen)
25. NOVARTIS PHARMA AG (Novartis)
26. NOVO NORDISK A/S (Novo)
27. ORION OYJ (Orion)
28. PFIZER INC (Pfizer)
29. F. HOFFMANN-LA ROCHE AG (ROCHE)
30. SANOFI-AVENTIS DEUTSCHLAND GMBH (Sanofi)
31. TAKEDA PHARMACEUTICALS INTERNATIONAL AG (TPIZ)
32. UCB BIOPHARMA (UCB)
33. ABBVIE INC (AbbVie)
42. CAATEVENTS gGmbH (CAATevents)
43. GLAXOSMITHKLINE RESEARCH & DEVELOPMENT LIMITED (GSK)
44. INSTITUT SERVIER DE MEDECINE TRANSLATIONNELLE (ISMT)
45. ZOETIS BELGIUM SA (ZOETIS)

Grant Agreement	The agreement signed between the beneficiaries and the IHI for the undertaking of the VICT3R project (101172693).
Project	The sum of all activities carried out in the framework of the Grant Agreement.
Consortium	The VICT3R Consortium, comprising the above-mentioned legal entities.
Consortium Agreement	Agreement concluded amongst VICT3R beneficiaries for the implementation of the Grant Agreement (GA). Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the GA.

Executive Summary

This deliverable reports the results from the statistical analysis undertaken to identify and characterize key covariates essential for the development and implementation of Virtual Control Groups (VCGs) in nonclinical toxicology studies. In this first exercise, the effect of the covariates is estimated using linear regression models. The analysis was conducted on the VICT3R database (v2.1.1), focusing on data from species rat (due to its prevalence).

The work confirms that a subset of covariates can account for a significant portion of the variability observed in controls, in particular sex, strain, and the year of the study. A significant portion of the variability in the data was attributed to general study-level effects. Depending on the type of measure (domain), the importance of a covariate may vary, as it was observed for organ weight measures with the covariate “body weight at baseline”. A substantial amount of "residual" or unexplained variability remains, indicating that other unmeasured factors influence the results. Preliminary analysis showed that operational covariates such as sponsor or pH of Vehicle, can contribute substantially to inter-study variability.

Looking across the endpoints, some were quite stable from study to study while others showed high variability. With this initial set of covariates, in some cases we were able to reproduce the same level of precision and variability as in the study level, in others we observed that more covariates will be needed, and lastly there are some cases where the endpoint variable is unstable even at the study level, which proves that there may be some external factors besides the covariates affecting them.

A changepoint analysis was run to find a past timepoint where covariate measurements are similar enough to form virtual control groups. The preliminary analysis done on Alkaline phosphatase in male Sprague Dawley helped to determine the magnitude of covariate changes required for reliable detection.

The report highlights the need for further curation of the data to build more comprehensive and generalizable models.

Introduction

Obtaining Virtual Control Groups (VCGs) by random sampling pooled data requires this data to be homogeneous, so the variability of relevant toxicological endpoints between the VCG components can be attributed to random sampling effects, thus being comparable to the variability observed in Concurrent Controls. A higher variability is often indicative of a lower homogeneity in the pooled sample, due to the effect of certain variables (covariates) which exert influence in the considered endpoint.

This analysis aims to identify the covariates necessary for matching VCGs with treated groups. The core principle of this matching procedure involves the identification of the study-specific variables from the treatment arms. These variables are then utilized to select the most "similar" animals within the VICT3R database, which are subsequently designated as matched controls.

To create VCGs, it is critical to determine which covariates are essential for the toxicological endpoints of interest. Many covariates may limit the pool of available animals due to data availability constraints. Conversely, ignoring certain covariates might compromise the reliability of the study using VCGs, as the VCGs may not exhibit sufficient similarity to the treated group.

The selection of the covariates for exploration depends on the data availability. The set of covariates considered in this analysis comes primarily from (Palazzi et al, 2024) who identified 52 potential covariates and was further expanded based on insights gathered by the partners. (Table A-1 in appendix).

Sections

1. Covariate Analysis [UPF]

1.1 Methods

This analysis aims to identify the list of covariates that are needed for the VCG generation to achieve the same level of precision and variability that is observed in concurrent controls from real life studies.

A step-by-step process was followed, starting with a general list of potential covariates, and focusing on some scenarios that maximize the available data. Then, on one hand we calculated the variability and precision observed at the study level in some endpoints from the Laboratory test results (LB) and Organ measurements (OM) SEND domains of the VICT3R dataset. On the other hand, an iterative process was followed using linear regression models to quantify the variability and precision of the endpoints after accounting for the selected covariates. These results are later compared to decide if VCGs generated with these covariates could replicate the conditions seen on concurrent controls.

General decisions

Using the most recent version of the VICT3R dataset at the time of the analysis (v2.1.1), the analysis focused on Rats, as they are the species with the most available data (74885 available animals), although it will be extended to other species in the future.

To ensure comparability of the studies and given that there is no standardized definition for the days of data collection, the corresponding variables LBDY and OMDY, indicating the day of the study where the data was collected, was fixed as 28 days with a margin of error of +/-3 days. This was chosen as it was the time interval with most available data in the VICT3R database. For each subject, if multiple measurements were available within the collection window, the observation closest to the data collection day was selected. Only the most common measurement unit per endpoint was retained to maintain unit homogeneity across the dataset

Only studies performed after the year 2000 were used for this analysis. This is being done to reduce the effect of the changes in procedures and instruments that have taken place over the years. A more specific threshold will be applied in the future once the corresponding analysis (currently being performed by TU Dortmund) is completed and incorporated here.

Covariate Selection

A list of potential covariates was extracted from Palazzi et al [*Palazzi 2024*], where they identified 52 covariates that may be important for the VCG generation.

Of these 52 covariates, only the ones that were available in the VICT3R database and were also properly curated in version 2.1.1 were selected for this analysis. Other potential covariates could be included if they are incorporated in future versions of the database.

After this filtering process, the list of potential covariates included both biological and experimental design variables:

> Numerical: BW_Start (starting body weight), AGE, START_YEAR, STUDY_DURATION, LBDY/OMDY
>
> Categorical: SPECIES, SEX, STRAIN, ROUTE (route of administration), COMPANY (sponsor), HOUSEGRP (housing group)

In the current version of the dataset for rats, these are the available datapoints per covariate (excluding LBDY/OMDY which are specific to the endpoints):

Covariate name	Number of animals	Percentage of total
Species, Sex, Strain, Sponsor	74885	100.0%
Route	73878	98.7%
Start Year	64174	85.7%
BW Start	53262	71.1%
Study Duration	53025	70.8%
Age	32197	43.0%
Housing group	18039	24.1%
Complete cases	8757	11.7%

Table 1.1.1: Number of rats with available data for each covariate and the corresponding percentage of the total number

Most of them were directly available in the VICT3R database. It was not possible to use the study length (SLENGTH) captured in TS domain because they are very few records capturing this information, instead it was calculated as the difference between study end date and start date (it is called STUDY_DURATION). The starting body weight (BW_Start) was chosen as the value of body weight closest to day 1, with a margin of +/-3 days. And the housing group was classified into three categories (individual, paired and grouped).

The variables SPECIES and LBDY/OMDY were fixed as described above.

Endpoint selection

For the LB domain, we focused on the 30 most frequent endpoints (Table A.3 in the Appendix), defined as unique combinations of LBTEST (test name) and LBSPEC (specimen type). For the OM domain we focused on the 8 most frequent endpoints (Table A.4 in the Appendix), in this case corresponding to organ weights.

For each of the selected endpoints, the corresponding dataset was created using either the LB or OM domains, and this was merged with subject-level covariates using STUDYID and USUBJID. This main dataset contained a column (either LBSTRESN or OMSTRESN depending on the domain) corresponding to the values of the given endpoint, followed by columns with the information of the selected covariates.

Rows with missing values in the outcome (LBSTRESN or OMSTRESN respectively) were excluded. Moreover, values further than 5 standard deviations away from the mean were considered biologically implausible and excluded (with the goal of removing only the values that were errors).

Only rows without missing values in numerical covariates were retained. For categorical variables, missing values were recoded as "Missing" to preserve data.

Description of the model used

To evaluate whether a defined set of covariates can account for the variability observed in laboratory endpoints across multiple studies, we adopted a linear regression framework, which assumes the endpoint variables can be explained as a linear combination of the potential covariates plus a residual error. The objective of this modelling step was not to predict individual values with maximal accuracy, but rather to assess whether key covariates capture the level of variability and precision observed at the study level. The linear model provides direct estimates of covariate importance (through regression coefficients and p-values), which can be interpreted and compared across endpoints.

The linear regression model was defined as:

$$Y_j = \beta_{0j} + \sum_{i=1}^{k} X_{ij}\ \beta_{ij}\ + \epsilon_j,$$

Where the dependent variable (Y_j) is the endpoint, the beta values represent the coefficients of the model for that endpoint, and the independent variables are the covariates.

The rationale for using a linear model includes:

- Interpretability: Coefficients provide intuitive insights into covariate effects.
- Transparency: Useful for identifying candidate covariates for VCG modeling.
- Consistency: Aligns with regulatory expectations for reproducible and auditable models.
- Comparability: Since the formula remains stable across endpoints, the results can be easily compared.
- Variability measurement: The variability of the residuals *(Var(ϵ))* measures the unexplained variability, which is the one remaining after accounting for the effect of the covariates. This can be compared to the study-level variability to measure the models' performance.

To satisfy linear modeling assumptions, the response variable LBSTRESN was transformed as needed. If the distribution was approximately symmetric and normally distributed (Shapiro–Wilk $p > 0.05$ and skewness < 1), no transformation was applied. If not, a log transformation was tested. If the log-transformed data passed normality, it was used.

Otherwise, a Box–Cox transformation was applied [*Box 1964*], with λ estimated via maximum likelihood. All transformed values were later back transformed to the original scale before performance evaluation. This step was essential to ensure homoscedasticity and normality of residuals, key assumptions of linear models.

For each study, a Leave One Study Out (LOSO) cross-validation approach was used. In each fold, a specific study was left out, a linear model was trained on all other studies and used to predict outcomes in the held-out study. This strategy mimics real-world VCG application, where we seek to predict for a new study using a historical dataset, and it's a fair comparison to the way the baseline metrics were calculated. The metrics for each of the folds in terms of prediction accuracy, variability and importance of the covariates were calculated and later summarized to have an overview for each of the endpoints.

Metrics at the study-level and from the covariates model

For each endpoint variable there were several metrics that were collected at the study level from the dataset and characterize the variability in the concurrent controls. These metrics were then compared to the metrics obtained with a linear model that only uses the selected covariates for prediction.

On the one hand, the baseline study-level variation was calculated as the weighted mean of the coefficient of variation (CV), considering the number of animals per study, that is:

$$Baseline_{CV} = \frac{\sum_{i=1}^{k} CV_i n_i}{\sum_{i=1}^{k} n_i}, \qquad CV = \frac{SD}{Mean},$$

Where *k* is the number of studies and n_i is the number of animals in study *i*. Moreover, the mean value of the endpoint per study is also calculated. These metrics provide both a measurement of the expected values and the variability for the endpoint in each of the studies and in general.

On the other hand, there are a set of metrics that were calculated using a linear model and cross-validation. In each fold, there are several metrics that were calculated:

- Test CV: Standard deviation of residuals from the LOSO predictions, divided by the mean observed value in the test set.
- NRMSE (Normalized RMSE): Quantify precision of predictions in absolute and relative terms.
- R^2 (Coefficient of Determination): Proportion of variance explained by the model.
- Predictions Mean: Mean value of the test set predictions.
- RME (Real Mean Error): The difference between the observed mean and the predicted mean, divided by the predicted mean.
- Model CV: calculated as the weighted mean of the Test CV values considering the number of animals per study.
- CV ratio: which is the model CV divided by the baseline CV. Ideally this value would be as close to 1 as possible, thus indicating that the variability obtained after removing the effect of the considered co-variates is similar to the experimental variability.

Moreover, in each fold, the covariates importance was measured with the coefficients from the linear model along with their p-values. These were standardized, stored and later summarized using the mean across studies. The values from categorical variables were summarized into one also by using the mean across categories.

Several plots were produced to summarize the results of the metrics from the LB and OM domain variables, as well as the comparison with the baseline values and a summary of the covariates importance in each scenario.

In future analysis, mixed models (e.g. with random effects for study or sponsor) may be considered, but there are some complications to consider, like the low number of animals in some studies and the low number of animals for some sponsors in certain endpoints, as well as the reduced interpretability of the results.

1.2 Results

The analysis uncovered some important initial results that show the potential of using these methods with the VICT3R data but also show some limitations that will need to be solved during the project. The results are grouped by LB domain and OM domain in this section, with some discussion points at the end.

For the LB domain the analysis includes the 30 most frequent combinations of Specimen and Test, and for the OM domain it consists of the 8 most frequent organ weights, all of these selected from the rats' subset of the VICT3R data.

LB domain

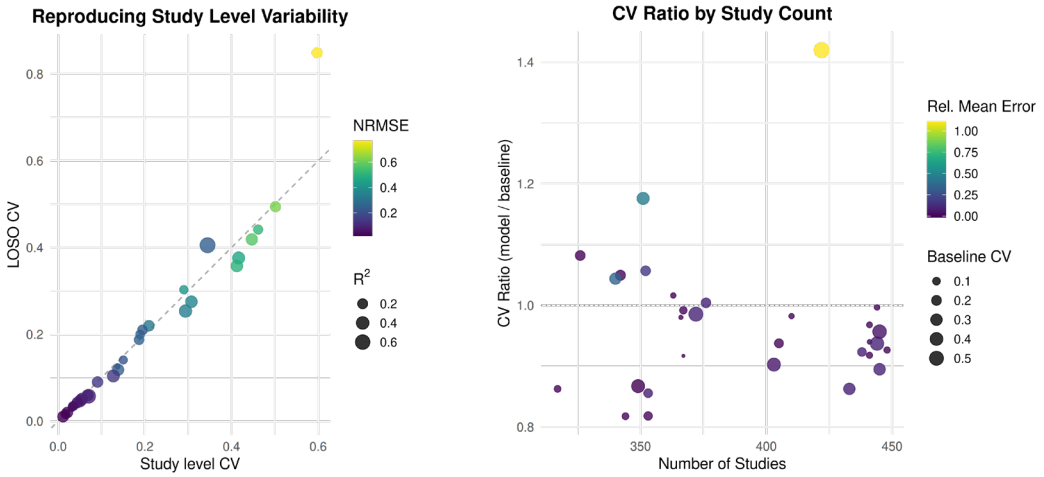


Figure 1.2.1: The left plot shows the comparison of study level CV with the Model's CV using cross validation. The right plot shows the CV ratio with respect to the number of studies. Each point represents a different studyendpoint variable i n the plots.

Figure 1.2.1 (left) shows the comparison of the Baseline CV (x-axis) and the Model's CV (y-axis). Each point represents a different endpoint variable, the size shows the explained variance (R^2) and the color of the points indicates the NRMSE. Most points show an intermediate R^2 of around 0.3, and the NRMSE seems to increase as the Baseline Study level CV increases, indicating that variables with higher variability at the study level show higher NRMSE in the predictions with the model. The closer the points are to the diagonal line $y = x$, the closer the model can reproduce the unexplained variability that is observed in real studies, and most points seem to fall close to this line with only one exception.

This is also present in Figure 1.2.1 (right), where most points fall in the ideal area within 0.8 and 1.2 of CV ratio. It is noticeable how one of the endpoints shows a higher degree of both CV ratio and baseline variability, as well as high levels of error in the estimates. This corresponds to whole blood basophils, and it belongs to a category of endpoints that are more sensitive to changes that the animals suffer during the study, leading to high variability even in animals with the same initial characteristics, and thus are more challenging to predict with our models. There also seems to be a tendency of lower CV ratio in endpoints where the number of available studies is higher, indicating better estimates with higher sample size.

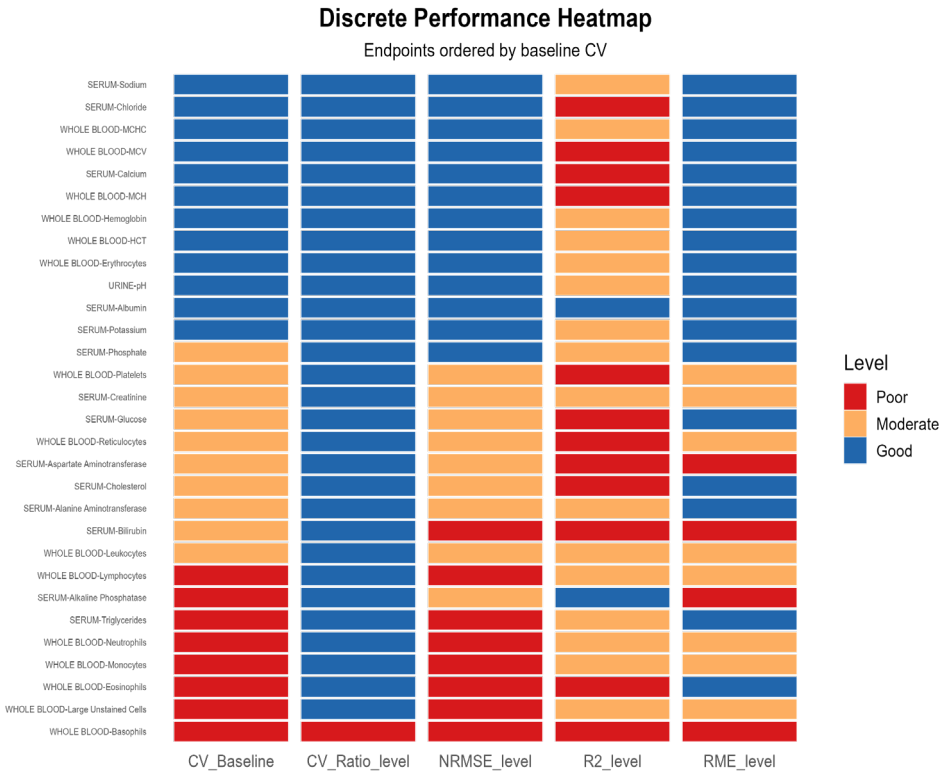


Figure 1.2.2: Heatmap showing the main summarized metrics by endpoint. The colors indicate the categories of the values adapted to each metric.

A summary of the main metrics for each endpoint can be found in the heatmap from Figure 1.2.2. Here, the values have been classified into categories of good, moderate or poor according to usual ranges for each individual metric, after adapting some general recommendations [*Gelman 2008, Gelman 2007, Cohen 1988, Hyndman 2006*] to this specific scenario. The final choice of thresholds can be found in Figure 1.2.3.

Metric	Good	Moderate	Poor
CV_Baseline	*≤ 0.10*	*(0.10, 0.30]*	*> 0.30*
CV_Ratio	*(0.80, 1.20]*	*(0.60, 0.80] or (1.20,1.40]*	*≤ 0.80 or > 1.20*
NRMSE	*≤ 0.20*	*(0.20, 0.40]*	*> 0.40*
R2	*> 0.40*	*(0.20, 0.40]*	*≤ 0.20*
RME	*≤ 0.05*	*(0.05, 0.15]*	*> 0.15*

Figure 1.2.3: Thresholds used to define categories in each metric

As one can see in Figure 1.2.2, the CV at baseline, indicating the variability observed at the study level, seems to dictate the behavior of the NRMSE, CV ratio, and partly the RME. The R^2 shows generally moderate to poor values (R^2>0.2) indicating a moderate to high level of explained variability.

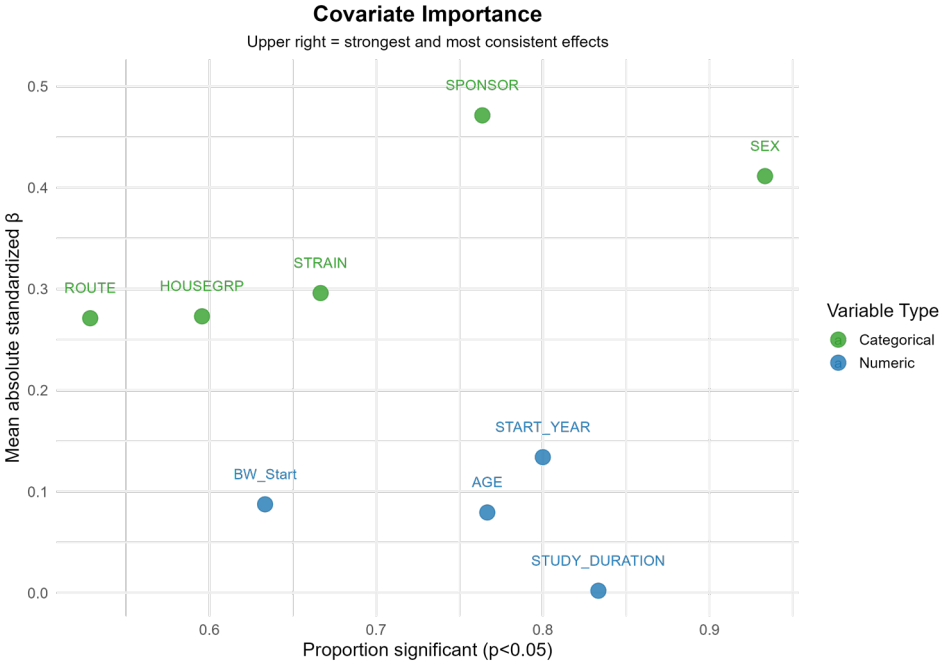


Figure 1.2.4: Importance of the covariates in terms of the coefficients from the linear model and the proportion of endpoints where they were found to be statistically significant for the model in terms of their p-value.

Lastly, Figure 1.2.4 shows the importance of each covariate in the model, after following a standardization process for the coefficients to allow comparison. Although the coefficients of numerical and categorical variables follow different definitions in a linear model, this plot allows us to observe some of the most important covariates in terms of coefficient magnitude in the models, and in terms of p-values after summarizing across the different folds of the cross-validation.

The study duration for example seems to have a very low coefficient in general, and the Route of administration, although it has a higher coefficient, it's found to be generally less significant. On the other hand, variables like sex, sponsor or start year of the study seem to be the most important in the models.

OM domain

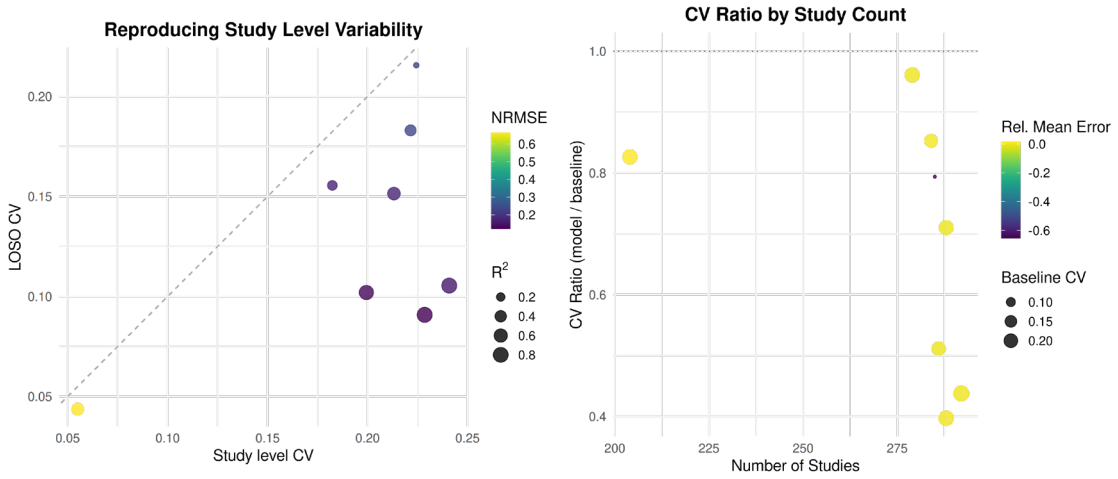


Figure 1.2.5: The left plot shows the comparison of study level CV with the Model's CV using cross validation. The right plot shows the CV ratio with respect to the number of studies. Each point represents a different study in the plots.

The OM domain shows different results with respect to the ones from the LB domain. In Figure 1.2.5 (left), (equivalent to Figure 1.2.1 above) we can see how all points fall in the area where the Models' LOSO CV < Study-level CV, meaning that the model is reducing the level of variability that is observed in real life studies. This is probably due to the high correlation of the covariate BW_Start with the weights of the organs. Moreover, the explained variability given by the R^2 coefficient is higher than before, with values higher than 0.4 in many cases.

In Figure 1.2.5 (right) we can see more clearly how the RME remains very low, and there are cases where the CV ratio is low (reaching values of 0.4), indicating that the model predicts well the values of the organ measurements, but it fails to recognize some factors that cause variability in animals within the same studies.

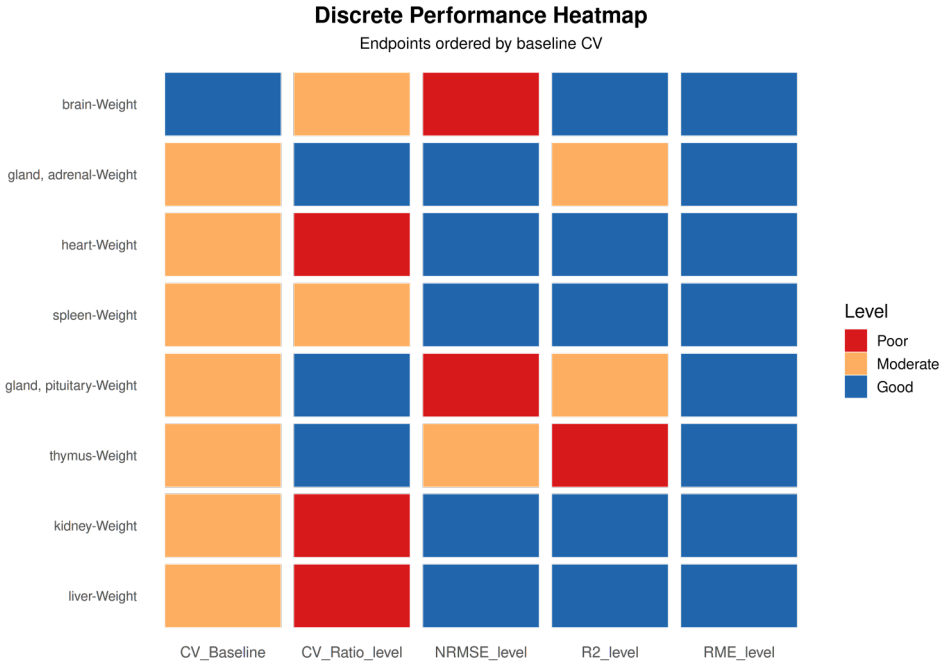


Figure 1.2.6: Heatmap showing the main summarized metrics by endpoint. The colors indicate the categories of the values adapted to each metric.

This is all summarized in Figure 1.2.6, (see Figure 1.2.2 above for the color code) where all the metrics are plotted for each endpoint. In this case the CV at baseline remains mostly in moderate levels (between 0.1 and 0.3) showing that most variables have variability in the endpoints within the same studies. As we saw before, the model predicts accurately the organ measurement mean values but it's not capturing these differences in animals within the same studies, causing the RME levels to be low but the CV ratios to be overfitted (the model has less variability than the baseline).

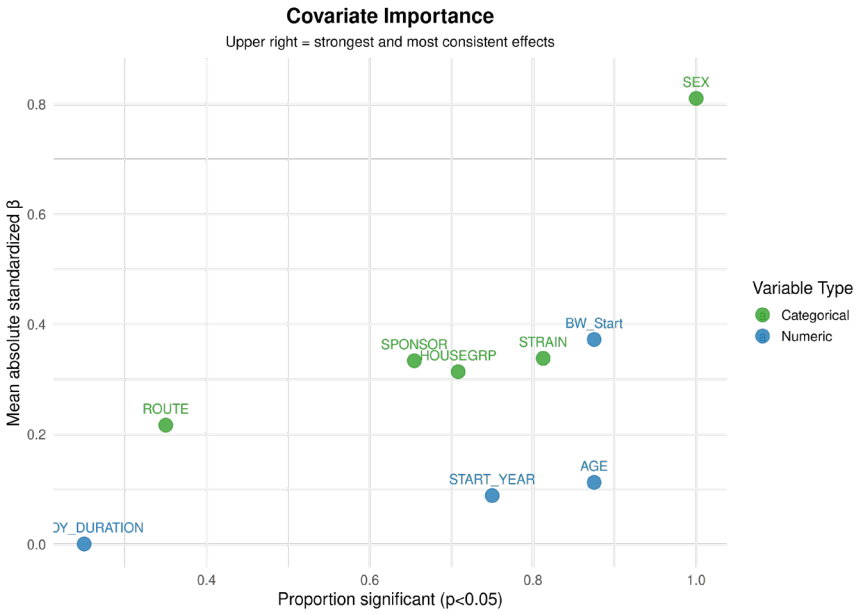


Figure 1.2.7: For OM domain, Importance of the covariates in terms of the coefficients from the linear model and the proportion of endpoints where they were found to be statistically significant for the model in terms of their p-value.

Figure 1.2.7 shows the covariate importance for the OM domain. In this case, sex has a high importance both in terms of the coefficients and in terms of significance, followed by BW_start (which has a very high coefficient even though numerical variables tend to have a lower one) and then strain and age. These are all related to the weight of the animal at the time of death, so it was expected that they would be important in the model. On the other hand, study duration and route are the less important variables in the model, same as it happened with the LB domain.

Discussion points

This analysis gives an initial overview of the covariates that will be needed for the matching process in the VCG generation and shows some limitations of the current dataset. These initial results indicate a positive first step towards the generation of VCG data and show how we could reproduce in many cases the same conditions of variability and precision that are found in real studies by using only a small subset of covariates and a simple approach with a linear model.

However, from the initial list of 52 covariates [*Palazzi 2024]*, only a subset of 11 is currently available and curated in the VICT3R dataset. Moreover, the use of the covariate "SPONSOR" is not a realistic approach, as this would be very sensitive to changes in procedures from the same sponsor and not generalizable to new sponsors.

There are different approaches that may be used to overcome this issue in the future. The covariate of the sponsor contains a lot of information within it about the procedures and the conditions of the study. When we have more available curated covariates in future versions of the dataset (such as information about the facility, diet, treatment vehicle, study type…), we could substitute the SPONSOR variable by these other covariates, which would create a more robust and generalizable approach.

The current dataset also has many gaps for certain domains and/or specific variables. This complicates the current approach, as it requires animals to have complete information at least in the main covariates and endpoints. Even in this analysis where the most frequent species (rats) were used, it was hard in some cases to get enough data to run the models, so it would be extremely difficult to generalize this to new species with less available data.

Although the results using this approach of the linear model look promising and they are easy to interpret, there is a complex scenario of correlations within covariates and with some of the endpoints. Thus, it may be needed to use more complex approaches that account for these interactions, like mixed models.

There are some challenges that need to be addressed in terms of data availability, data quality and some specifics about the models. Nevertheless, the results indicate that it is possible to recreate study specific conditions for a wide range of endpoints by only using a subset of covariates, which shows the possibility of generating robust VCGs.

1.3 References

[1] Palazzi X, Anger LT, Boulineau T, et al (2024) Points to consider regarding the use and implementation of virtual controls in nonclinical general toxicology studies. Regul Toxicol Pharmacol

[2] Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B.

[3] Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine.

[4] Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

[5] Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge.

[6] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting.

[7] Sokal, R. R., & Braumann, C. A. (1980). Significance tests for coefficients of variation and variability profiles. Systematic Zoology

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R (2nd ed.). Springer.

2. Variance explained by key covariates and analysis of residuals [ORGANON]

2.1 Analysis objectives

In this section we further substantiate the findings from the previous section (section 1) and undertake supplementary independent analysis in response to suggestions that emerged from the preliminary observations. Limited to the LB domain data, an independent approach was implemented as a form of independent validation and has the potential to provide additional insights beyond the primary results.

The supplementary analytical work and its specific objectives are:

1. Conduct a Variance Component Analysis to partition the total variability observed in laboratory tests commonly used in toxicology studies. This will serve to identify and rank the principal drivers of variance, apportioning it across known biological covariates and study-level effects.
2. Employ Linear Mixed-Effects Models (LMEMs) to precisely quantify the magnitude, direction, and statistical significance of "key" biological covariates (e.g., Strain, Sex, Body Weight) while appropriately accounting for the hierarchical, nested structure of the data (i.e., Body Weight within SEX, or Studies within Study Year).
3. Investigate the latent sources of inter-study variability by performing an Analysis of Variance (ANOVA) on the study-level random effects extracted from the LMEMs, explicitly testing their association with other “operational” covariates such as housing group, facility, animal supplier.

2.2 Source Data and Definition of the Analytical Cohort

All data for this analysis were sourced from the VICT3R database (V 2.1.1). To have comparable results with previous section, a similar homogenous and analytically tractable dataset was derived through the sequential application of inclusion and exclusion criteria, as specified below. Each filtering step was designed to control for potential major sources of confounding and to guarantee sufficient data for robust statistical modeling.

- **Species:** The analysis was restricted to studies involving only the Norway rat (*Rattus norvegicus*) to eliminate inter-species biological variation.
- **Study Sex Composition:** Only studies that included both male and female animals were retained. This criterion is critical to ensure that the effect of sex is directly estimable and not completely confounded with study-level effects, which would occur in single-sex studies.

- **Route of Administration:** To focus the analysis on the most common study designs and ensure adequate statistical power, the cohort was limited to studies using the two most frequently occurring routes of administration in the database (ORAL, SUBCUTANEOUS).
- **Strain:** Similarly, the analysis was restricted to the two most abundant rat strains present in the database to ensure sufficient sample sizes within each genetic background (WISTAR HAN, SPEGUE DAWLEY).
- **Time Point:** The analysis was focused on a standardized cross-sectional time window corresponding to Day 28 (+/−3 days) of the studies. This time point is representative of sub-chronic toxicity assessments and provides a consistent basis for comparison across studies.
- **Data Density:** For the final stage of filtering, only laboratory tests (LBTESTs) for each specimen (LBSPEC) with a minimum of 100 individual animal measurements remaining after the application of all preceding criteria were included in the modeling phase. This ensures that the statistical models for each LBTEST are stable and well-powered.

The resulting attrition of studies and animals from the application of these criteria are documented in Appendix.

2.3 Data Curation, Transformation, and Pre-processing

Following cohort selection, a multi-step data processing pipeline was executed to prepare the data for statistical modeling.

Laboratory (LB) Data

For each clinical pathology endpoint, the following steps were performed:

1. **Unit Standardization:** When multiple measurement units were recorded for a single LBTEST, only values from the most frequently occurring standard unit were utilized
2. **Outlier Removal:** To minimize the impact of probable erroneous data entries or technical errors, a conservative outlier removal strategy was employed similarly to what was done in section 1. For each LBTEST, values exceeding ±5 standard deviations from the mean of that test were excluded from the analysis.
3. **Logarithmic Transformation:** Many biological measurements exhibit right-skewed distributions. To better satisfy the normality and homoscedasticity assumptions of linear models, a skewness-based rule was applied. For laboratory tests with distributions exhibiting significant skew, a natural logarithmic transformation ($y'=\ln(y)$) was applied.
4. **Scaling and Centering:** Finally, all processed LBTEST values were standardized by subtracting the mean and dividing by the standard deviation. This transformation results in a variable with a mean of 0 and a standard deviation of 1. This step is crucial as it places all tests on a common

scale, allowing for the direct comparison of effect sizes (i.e., regression coefficients) across different models.

By scaling LB data, we are directly comparing how many standard deviations from the mean the outcome variable is predicted to be for different groups or conditions defined by the covariates. This is done to standardize the effect sizes and compare the relative importance of different covariates. For log-transformed LB values, changes detected on the log-transformed scale, correspond to a multiplicative decrease on the original scale. While we won't be able to state the exact percentage change in the original units due to the scaling step on top of the log-transformation, we will still compare the relative strength and direction of the effects of the covariates.

Bodyweight (BW) and Change in Bodyweight (DeltaBW)

Baseline bodyweight (BW0) and the change in bodyweight from baseline to the Day 28 time point (DeltaBW) were also processed:

1. **Outlier Removal:** The same ±5 standard deviation exclusion rule was applied to both BW0 and DeltaBW
2. **Categorization of Continuous Variables:** this was done for the variance decomposition model (described later in statical session), which accepts categorical variables only, while the linear mixed model used continuous BW data. The continuous BW0 and DeltaBW variables were converted into three-levels categorical factors (BW0LoMeHi, DeltaBW-LoMeHi), thus allowing to handle potential non-linear relationships and enhance interpretability. This was achieved by calculating tertiles (Low, Medium, High). This categorization was performed separately for males and females as rats exhibit significant sexual dimorphism in adult (28 weeks) size. By defining the tertiles *within* each sex, the BWLoMeHi variable represents an animal's relative size compared to its same-sex peers, thus isolating the biological effect of size from the effect of sex. Accordingly, the variance decomposition model that uses categorical body weight will capture any variability in LB values attributable to the animal being classified as "lighter," "normal," or "heavier" relative to its sex-specific peers.

Other Covariates

For operational covariates stored as free text (e.g., Housing Group, Breeder/Supplier, Facility, pH of vehicle, GLP flag, …), preliminary harmonization and categorization were performed using methods based on Large Language Models (LLMs) mainly to standardize inconsistent entries (e.g., "Charles River Labs" vs. "CRL") and create few homogeneous categories. This transformation is arbitrary and performed to have a preliminary estimate of the potential effect and importance of other covariates. When the VICT3R database is further curated and harmonized, VICT3R data could be directly used.

2.4 Statistical Modeling Strategy

A three-stage statistical modeling strategy, depicted in figure 2.1, was employed to systematically identify the sources of variability in the data. This sequence was designed to progress logically from a high-level, descriptive partition of variance to a detailed, inferential quantification of specific covariate effects.

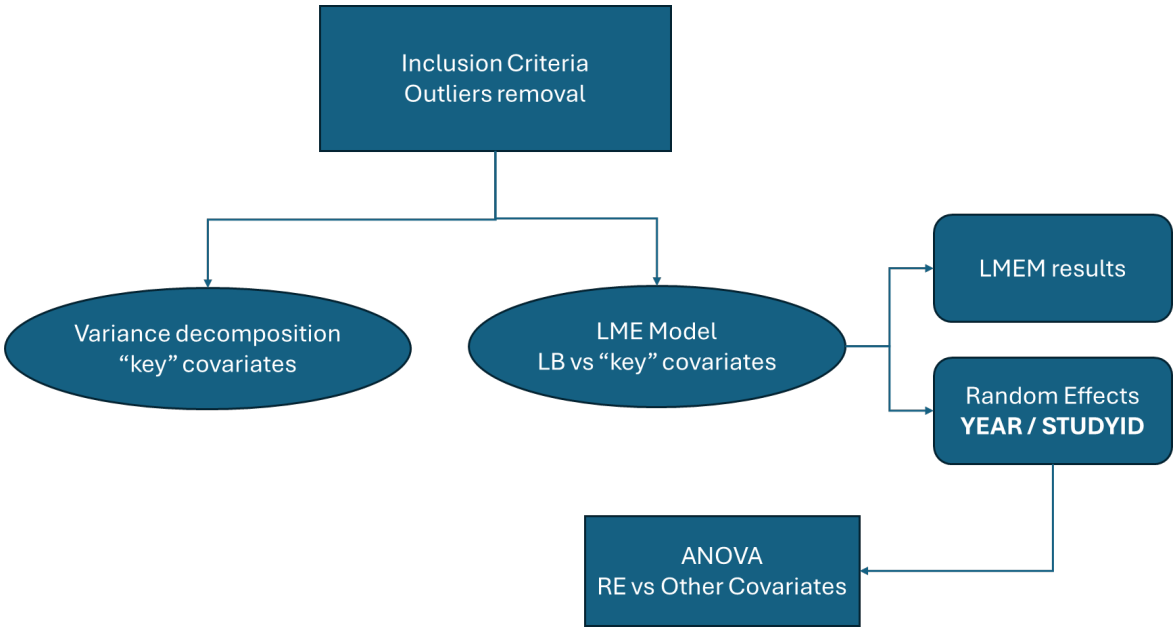


Figure 2.4.1: Schema of the statistical modeling strategy. Data from the VICT3R database is filtered according to inclusion criteria, outliers are removed, laboratory values are scaled and used in different models

Stage 1: Variance decomposition Analysis

The initial analytical objective was to determine the relative importance of each potential major covariate in explaining the total observed variability of a given TEST. We aim to have a population wise approximation of the covariates effects, given that the VICT3R database itself is a subset of the real population and that not all existing covariates values are recorded in the database. To achieve this, a variance component analysis was conducted. This technique uses a random-effects-only model to partition the total variance into a sum of variance components, with each component corresponding to a covariate in the model. This provides a powerful, high-level "map" of the variance structure, highlighting which covariates are the dominant sources of animal-to-animal variation before proceeding to more detailed modeling.

Model Specification: For each scaled and transformed laboratory endpoint (LBSTRESN), the following linear model with only random effects was fitted:

$$LB_p = \mu + \alpha_{i(p)} + \beta_{j(p)} + \gamma_{k(p)} + \delta_{l(p)} + \zeta_{m(p)} + \eta_{n(p),o(p)} + \varepsilon_p$$

Where each term is defined as follows:

- **LBp**: The scaled and transformed value of the laboratory endpoint for the p-th animal.
- **μ**: The overall mean of the endpoint across all animals and studies (the fixed intercept).
- $\alpha_{i(p)}$: The random effect associated with the STRAIN of animal p.
- $\beta_{j(p)}$: The random effect associated with the SEX of animal p.
- $\gamma_{k(p)}$: The random effect for the baseline body weight category (BW_LoMeHi) of animal p.
- $\delta_{l(p)}$: The random effect for the change in body weight category (DeltaBW_LoMeHi) of animal p.
- $\zeta_{m(p)}$: The random effect for the ROUTE of administration used for animal p.
- $\eta_{n(p)},o_{(p)}$: The nested random effect for the specific STUDYID (indexed by o) within the study YEAR (indexed by n) to which animal p belongs.
- **ϵp**: The residual error for animal p, which captures all remaining unexplained variability (e.g., individual biological randomness, measurement error).

Assuming the independence across the covariates, the total variance of the endpoint (Var(LBp)) can be decomposed into the sum of the individual variance components, which then allows for the calculation of the proportion of the total variance that can be attributed to each factor.

A key output of this analysis is the magnitude of the residual variance itself. This value represents the "noise floor" of the system — the amount of variability that persists even after accounting for all major known biological and study-level factors. Its magnitude provides a crucial, data-driven reality check on the theoretical limits of a VCG. A large residual variance implies that a significant portion of inter-animal variability is either truly random or driven by unmeasured factors, thereby setting a practical upper bound on the achievable precision of any VCG matching algorithm.

Stage 2: Linear Mixed-Effects Modeling (LMEM) of Key Covariate Influence

Having established the relative contributions of the variance sources, the second stage employed Linear Mixed-Effects Models (LMEMs) for formal hypothesis testing and estimation of effect sizes. This approach is essential for analyzing data with a hierarchical or clustered structure, such as animals nested within studies. Traditional methods like ordinary least squares regression assume independence of all observations, an assumption that is violated in this context. Failure to account for this non-independence leads to underestimated standard errors and an inflation of the Type I error rate (an increased risk of false-positive findings). LMEMs resolve this by incorporating random effects to model the correlation among observations within the same cluster (study). This approach has become the standard for robust analysis of clustered and longitudinal biological data (Ruiz et al., 2023; Yu et al., 2022).

Model Specification: For each endpoint, the following LMEM was fitted, treating key biological factors as fixed effects and the study identifier nested within year as a random effect:

Let LBijk be the LBSTRESN (the response variable) for the k-th individual animal, from the j-th study, conducted in the i-th year. The model is specified as:

$$LB_{jik} = \eta_{ijk} + u_i + v_{ij} + \varepsilon_{ijk}$$

Where the term η_{ijk} is the linear predictor for the fixed effects:

$$\eta_{ijk} = \beta_0 + \beta_1 * STRAIN_k + \beta_2 * SEX_k + \beta_3 * (SEX_k * BW0_k) + \beta_4 * ROUTE_k + \beta_5 * DeltaBW_k$$

u_i is the random intercept for the **YEAR**. It represents the deviation of the i-th year's average LBSTRESN from the overall average defined by the fixed effects. v_{ij} is the random intercept for the **STUDYID**, nested within **YEAR**. It represents the deviation of the j-th study's average LBSTRESN from the average of the i-th year.

ε_{ijk} is the residual error term for each observation, which accounts for the random, unexplained variation within each study group

The model estimates a single, population-average coefficient (a β parameter) for each of the covariates. These coefficients represent the estimated magnitude and direction of each factor's effect on the LB endpoint, adjusted for all other factors in the model.

The random intercepts are crucial components as they estimate the variance of the study-specific intercepts (σ^2_{study}), assuming they are drawn from a normal distribution with a mean of zero. This accounts for inter-study heterogeneity as a source of random variation in the data, which will be used in the subsequent step.

Stage 3: Deconstruction of Inter-Study Variability via ANOVA on Random Effects

The LMEM in Stage 2 effectively accounts for study-to-study variability but does not explain its origins. The random effects term isolates the systematic deviation of each study's mean from the overall population mean, adjusted for the fixed effects. These study-specific deviations, known as Best Linear Unbiased Predictors (BLUPs), can be extracted from the fitted LMEM for each study. The central hypothesis of Stage 3 is that this "unexplained" study-level variance is not entirely random but may be systematically driven by the "other" operational covariates (e.g., Housing, Breeder/Supplier, Facility) that were not included in the primary LMEM.

To test this hypothesis, this analysis employs an innovative approach. The BLUPs for each study are treated as a new response variable, and an Analysis of Variance (ANOVA) is used to determine if the mean BLUP value differs significantly across the levels of these operational factors.

Model Specification: For each laboratory endpoint, the study-level BLUPs were extracted from the corresponding LMEM fitted in Stage 2. Subsequently, a series of one-way ANOVAs were performed, with each model taking the form:

$$BLUP_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Where:

- $BLUP_{ij}$ is the BLUP value for the j-th observation within the i-th group of the "Other".
- μ is the overall grand mean of BLUP across all observations.
- α_i is the effect of being in the i-th group of "Other" covariate. It represents the deviation of that group's mean from the grand mean μ.
- ϵ_{ij} is the random error term for each observation, assumed to be normally distributed with a mean of 0 and variance σ^2, written as $\epsilon_{ij} \sim N(0,\sigma^2)$.

where "Other" covariate was, in turn, each of the operational variables of interest (e.g., Housing group, Facility, Breeder/Supplier, pH of Vehicle, GLP flag, Sponsor …).

A statistically significant result from this analysis (e.g., a low p-value for Housing group) carries profound and actionable implications. It provides empirical evidence for a causal pathway: the choice of an operational factor (e.g., animal supplier) leads to a systematic and predictable shift in the baseline physiology of the animals in a study, a shift that is captured by the random effect in the LMEM. This finding would demonstrate that a VCG matching algorithm based solely on biological covariates like sex and strain would be insufficient. To create a truly comparable virtual control, the algorithm must also incorporate key operational factors as either matching or stratification variables. This final analytical stage is therefore designed to provide the direct evidence needed to justify these critical design decisions for the future VCG platform.

2.5 Results: Variance explained by key covariates

The variance decomposition step provides a population-level ("global") assessment of the relative contribution of each covariate to overall variability, under the assumption that the **VICT3R** dataset is representative of the broader population. To mitigate the impact of temporal changes in laboratory techniques, study-to-study variability was estimated within each study year.

For each model, the variance components were calculated and expressed as percentages of the total variance.

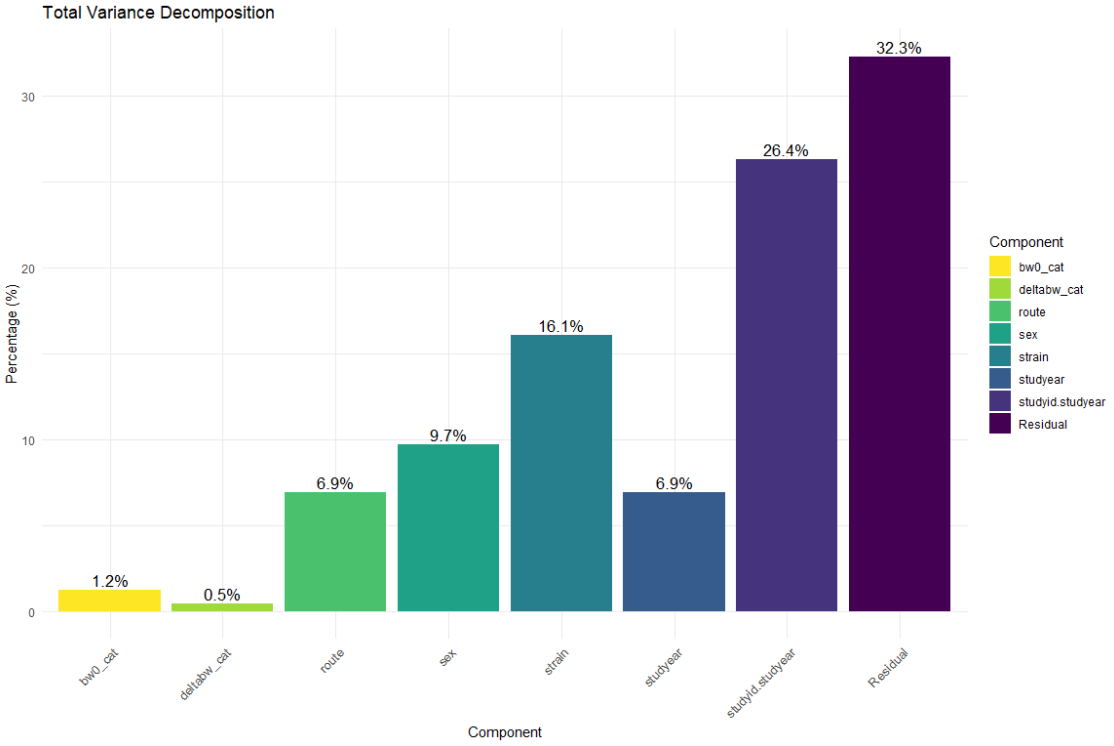


Figure 2.5.1: Percentages of variance associated with the key covariates (components) across all laboratory tests.

The pie chart of figure 2.5.1 depicts the percentages of variance across all tests explained by the different covariates included in the model described in section 2.4. Key covariates (STRAIN, SEX, ROUTE, BW) account for less than 50% of the variance, while the remaining variance is attributed to study-to-study differences and residuals, reflecting variability that has not been explained by the model.

Among the explanatory factors, strain emerges as one of the most significant contributors to the variance, it confirms what is described by several publications (Weber, 2011; Kort, 2020). The strain component likely captures key biological differences, making it a critical factor to consider for matching.

Other components, such as route, and sex, play smaller but still notable roles in explaining the variance, suggesting that these factors have some influence and should be considered as well when creating VCGs.

Bodyweight components, analysed as categorical classes (Low-Medium-High), do not appear to have a significant impact on the variance after controlling for other covariates, at least not when examined within sex (nesting was considered during variable transformation, not included in the model). Taking into account that the bodyweight of the variance decomposition model is categorical, the true variability and impact of body weight could be underestimated here.

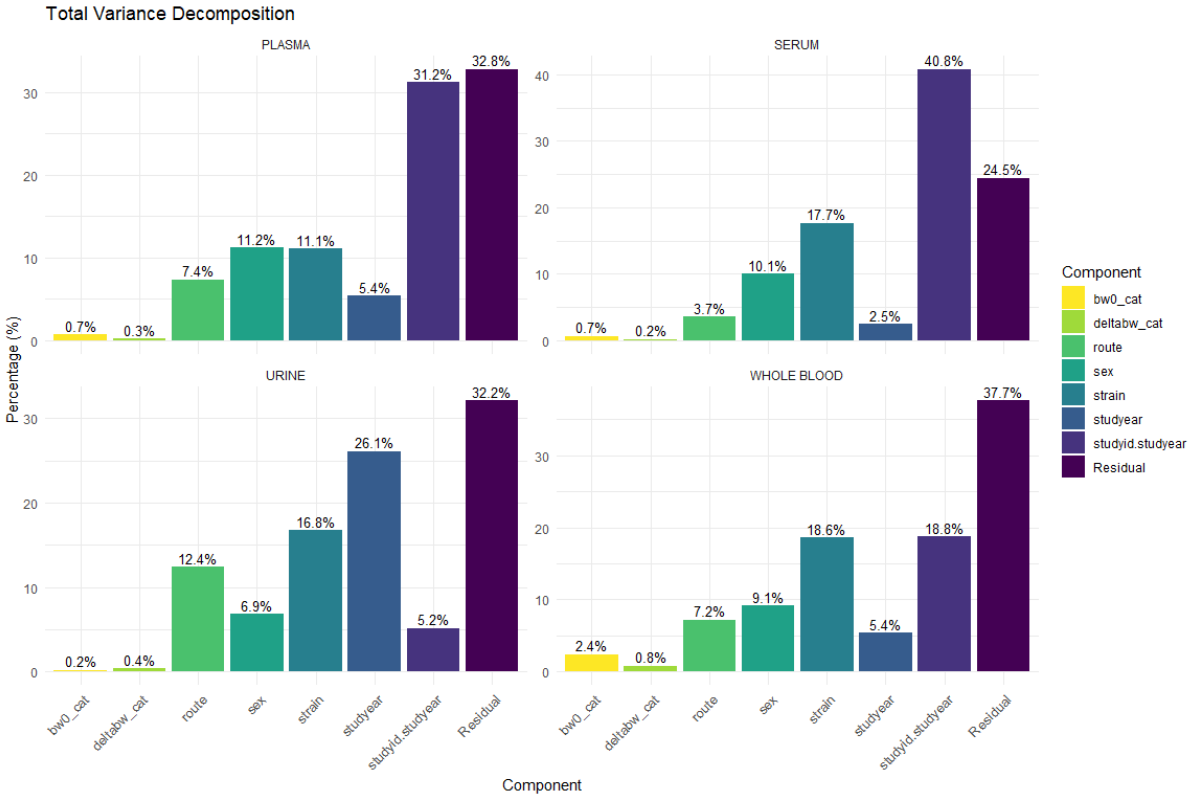


Figure 2.5.2: Percentage of variance associated with the key covariates across all laboratory tests by spicemen.

Upon stratifying the analyses by specimen type, the results were largely consistent with the combined, unstratified analysis. A noteworthy exception was observed for assays conducted on urine samples, which demonstrated low inter-study variability but, conversely, high variability across years. Furthermore, body weight accounted for a significantly greater portion of the variance in measurements from whole blood compared to those from other biological matrices.

This finding for urine is interesting. Low variation between different studies might suggest that the analytical methods are well-standardized methods. However, high variation from year to year could reflect systematic changes in diet, laboratory protocols or equipment that occur over the years (Barr et al., 2005).

The differences in variance patterns across specimens highlight the need for specimen-specific modeling approaches. For example, in Plasma and Serum, efforts may need to focus on identifying additional factors to reduce residual variance, while in Whole Blood and Urine, existing factors like strain and study year already explain substantial variability.

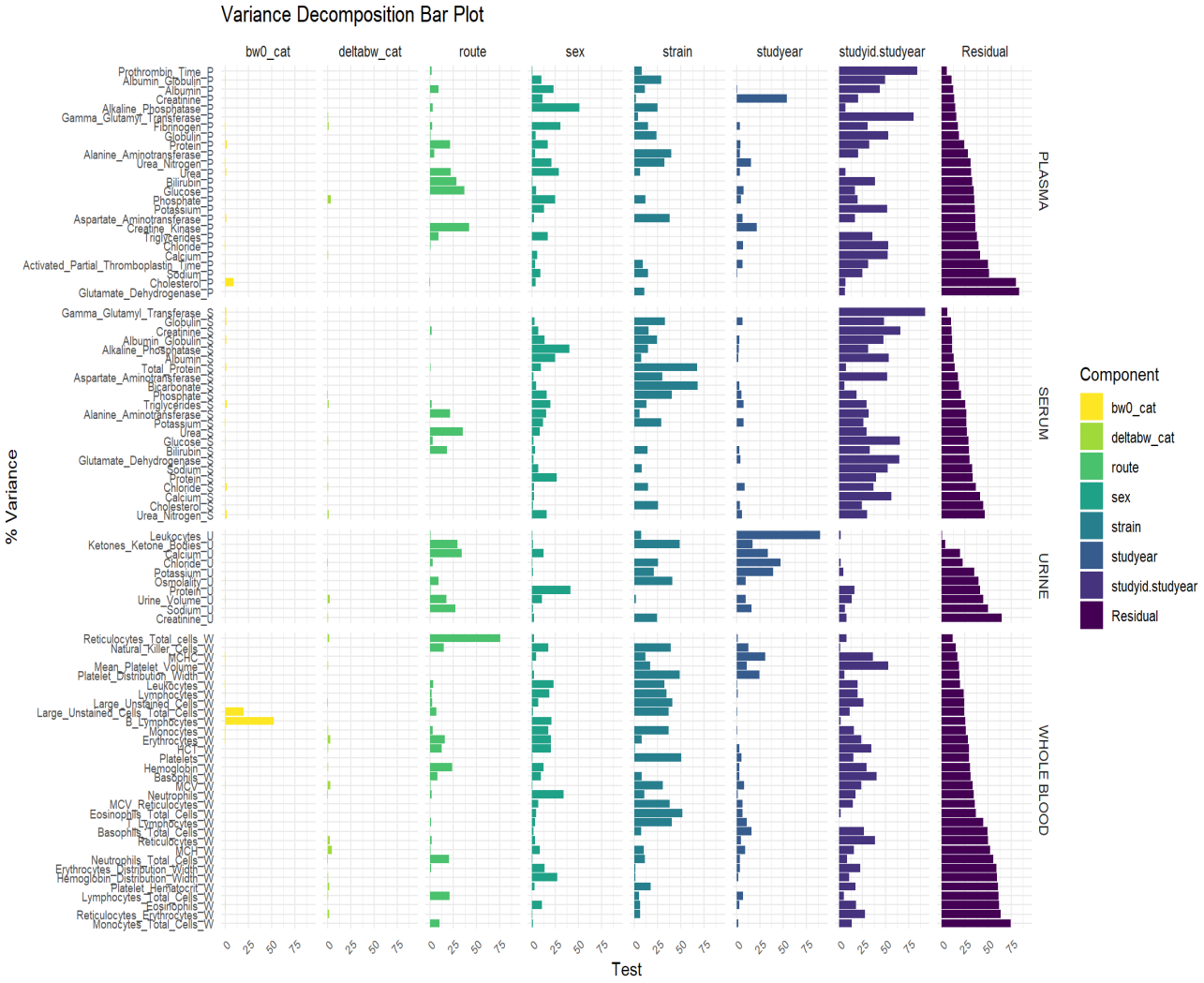


Figure 2.5.3: Percentages of variance associated with the key covariates by laboratory tests by specimen.

A detailed breakdown of the variance decomposition for each individual laboratory test across the four specimen types (Plasma, Serum, Urine, and Whole Blood) is given in figure 2.5.3. It complements the insights from the figure 2.5.2 by allowing us to examine how the contribution of each component varies for each laboratory test.

The proportion of residual variance differs across laboratory tests. In some cases, such as Prothrombin Time in Plasma or Reticulocytes Total in Whole Blood, the residual component is smaller, suggesting that the main covariates account for much of the variance in these tests. For other tests, including Cholesterol or Glutamate Dehydrogenase in Plasma and Monocytes Total in Whole Blood, the residual component represents most of the variance in the model. This indicates that these laboratory tests may be influenced by additional sources of variation not included in the current model.

The variability observed at the test level highlights the necessity of carefully selecting animals for use as VCG based on the specific set of tests intended for the study. The results suggest that, for certain

laboratory tests, it is essential for the VCG to closely match the characteristics of the treated animals and CCG, if included in the study, in order to minimize sources of variation.

2.6 Effects of Key covariates

Results from the linear effects model from stage 2 are presented in figure 2.6.1

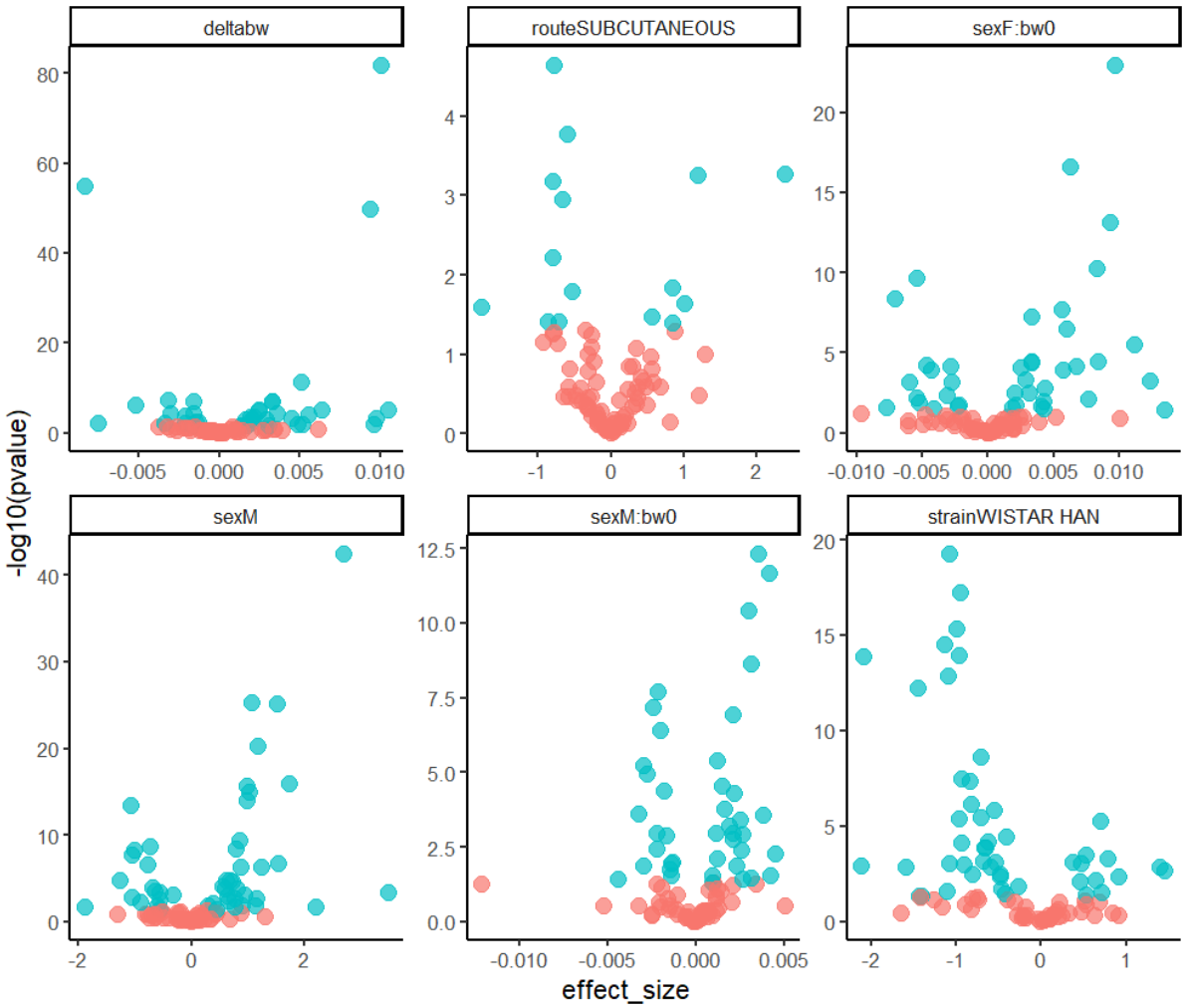


Figure 2.6.1: Vulcano plot of fixed effects from LME models for laboratory test assessed in this work. Each dot represents a laboratory test. All dots in blue green indicate tests with significant p-values (< 0.05) for each of the covariate tested as fixed effects.

While many of these findings are statistically significant, considering that they are obtained from scaled data, their biological importance needs further assessment, as it would depend on the size of the change (magnitude) compared to established reference intervals for the species and strain. For example, the effects associated with bw0 within sex and deltabw have very small magnitudes, they may be biologically meaningful, but data should be read on the original scale, whereas effects for sex or strain, which are categorical variables, are much larger and easily interpretable.

The percentage of test (by specimen) where a statistically significant effect of a covariate was observed in the linear mixed model ranged from 9 to 74 (table 2.6.1-A). SEX and STRAIN were the most frequent, especially for measurements in serum, followed by body weight parameters, and ultimately route.

SPECIMEN	BW_0 (F)	BW_0 (M)	Δ BW	SEX	STRAIN	ROUTE
PLASMA	36%	44%	28%	44%	40%	20%
SERUM	74%	61%	52%	70%	70%	9%
URINE	10%	10%	10%	40%	30%	22%
WHOLE BLOOD	41%	47%	59%	53%	59%	19%
GRAND TOTAL	44%	46%	43%	53%	53%	17%

Table 2.6.1-A: percent of laboratory tests within each specimen for which a covariate in the LME model showed a p-value (not adjusted) less than 0.05.

Despite the effect size for the numeric covariates related to body weight were smaller compared to the categorical variables, it is important to acknowledge that the effect was statistically significant (not adjusted p-value <0.05) for almost half of the laboratory measures tested here. As illustrated in Figure 2.6.2, variations in LB test results due to body weight can be observed. For instance: B lymphocytes reflect bodyweight changes during the study period; total protein is linked to initial bodyweight and body weight changes over the course of the study. Considering that for such parameter values were log transformed, the effect is multiplicative and cannot be neglected despite the small effect size.

The tests most sensitive to the covariates are summarized in table 2.6.1-B where substantial findings (effect size > 1.5 and $p < 0.05$) from the series of linear mixed models are shown. Among all the key covariate tested, the results highlight route of administration, rat strain, and sex as significant drivers of variation in specific tests.

Subcutaneous **route of administration** (compared to oral) affects reticulocytes (immature red blood cells) levels, suggesting a mild, localized stress response, possibly stimulating red blood cell production (erythropoiesis). Urinary ketones as well may vary upon the route, eventually reflecting differences in energy pathways.

The **rat strain** had a strong impact on few measurements. The decreases in protein, ketones, and certain immune cells (leukocytes, B-cells) highlight inherent variability in metabolism and immune function between strains.

The strong effects of **sex** on several analytes suggest a profound physiological dimorphism that impacts multiple organ systems. The concurrent elevation of liver-related enzymes (Alkaline Phosphatase) and lipids (Cholesterol, Triglycerides) points to significant baseline differences in hepatobiliary function and lipid metabolism, likely regulated by sex hormones. The increases in both B-Lymphocytes and Natural Killer cells indicate a sex-based difference in the immune system's baseline state.

SPECIMEN / LB TEST	ROUTE (SUBCUTANEOUS)	SEX(M)	STRAIN (WISTAR HAN)
PLASMA			
ALKALINE_PHOSPHATASE		1.7	
TRIGLYCERIDES		1.6	
SERUM			
CHOLESTEROL		2.7	
TOTAL_PROTEIN			-2.1
TRIGLYCERIDES		1.5	
URINE			
CALCIUM		-1.9	
KETONES_KETONE_BODIES	-1.8		-2.1
LEUKOCYTES			-1.6
WHOLE BLOOD			
B_LYMPHOCYTES		3.5	-1.6
NATURAL_KILLER_CELLS		2.2	
RETICULOCYTES_TOTAL_CELLS	2.4		

Table 3.6.1-B: Coefficients obtained from mixed models where any of the effects tested were larger than 1.5 (standard deviation units) and p values (not adjusted) less than 0.05.

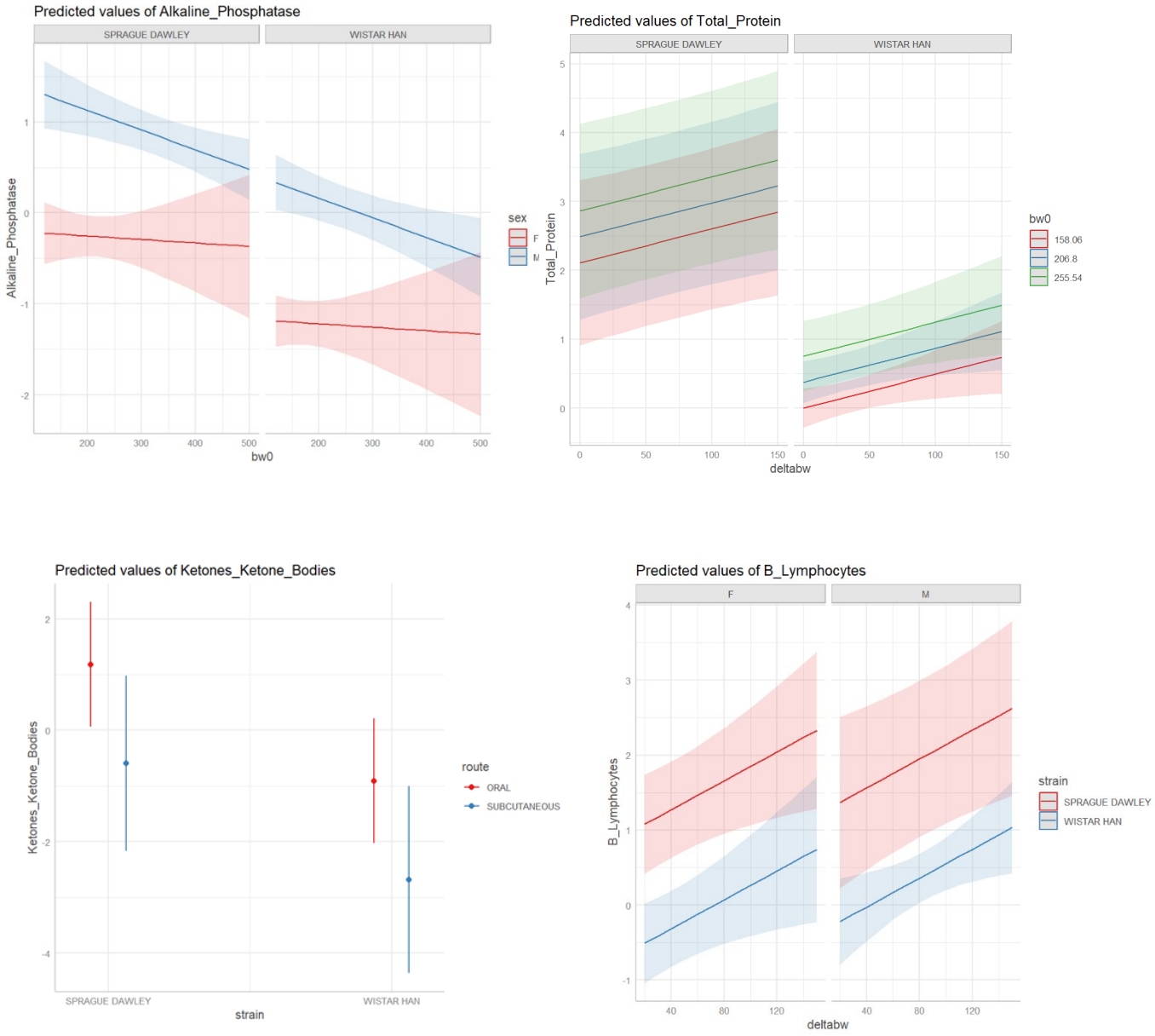


Figure 2.6.2: Visualization of the main effects for some of the tests reported in table 2.6.1. In this caseLB test values are log-transformed and scaled.

2.7 Variance explained by other covariates

Typically, details regarding the study setup are reported within the TS domain. However, unlike “key” covariates, certain study-specific information tends to have a lower percentage of availability in the current version of the VICT3R database. This is the case for Housing Group, GLP flag, pH of vehicle, animal supplier, and study type, which are the “other” covariates object of the present analysis obtained using the approach described in stage 3.

For some of them, such study-specific information is typically provided as open text, frequently using non-standardized formats, which can result in inconsistencies and multiple unique values. Therefore, variables were recoded to simplify the number of categories that exist for that specific covariate. To categorize the text variables, we utilized *ChatGPT* (GPT-5; OpenAI, 2025), a large language model accessed via the OpenAI platform using an API. The model was prompted with structured instructions to classify textual responses into predefined categories based on semantic similarity and context. The model’s outputs were reviewed manually to ensure reliability and accuracy, and any discrepancies were resolved. This approach allowed efficient handling of large textual datasets while maintaining human oversight, however, the outcome remains linked to the arbitrary classification obtained using LLMs whereas the definition of the categories needs a consensus from the scientific community. Other approaches might be considered in the future. Following results must be considered merely exploratory and preliminary.

As outlined in stage 3, a new model is built using the selected "other" covariate to determine if and how much it could explain the variability across studies measured as BLUPs from LME.

Housing conditions

Models results (52 converged) obtained from housing conditions (grouped, individual, pair-housed) indicated up to 9 blood and serum laboratory tests with low p-values across comparisons, implying that social environment plays a measurable role in physiological outcomes. However, the vast majority of studies are in the “grouped” category—often over 85% of the total. This skews the dataset and may inflate the statistical power for comparisons involving grouped housing. “Individually” and “Pair-housed” animals are severely underrepresented (1–4 studies or none at all). With such imbalance the robustness of p-values is compromised, and the risk of false positives or exaggerated effects is high.

For this reason, we will withhold reporting and interpreting the results until an updated version of the database containing more comprehensive records on housing conditions becomes available.

Animal Supplier

Similarly to housing conditions, data availability from animal suppliers was limited, and in the subset of data according to our inclusion criteria only two suppliers were represented, and with great imbalance between groups (0-3 studies for the underrepresented group), which posed challenges for conducting any statistical analyses under standard assumptions. The effect of animal suppliers on the study-to-study variability will be undertaken when additional data become available.

Sponsor

The role of the covariate "sponsor" in determining the variability of laboratory tests and organ measures has been also discussed in section 1. Results from the approach adopted in this section reflect previous findings, with a significant effect of sponsor on many endpoints (65% of the ones tested).

Among the others, model results for Serum Albumin were particularly robust and are shown in Figure 2.7.1. The variability observed across studies for Serum Albumin (Best Linear Unbiased Predictors derived from LME — R2 = 0.86) was largely explained by the sponsor (ANOVA R2 of 0.80, p<0.001) in a subset of data that included 72 studies and 1,363 animals. Two companies seem to systematically record lower values than the other companies.

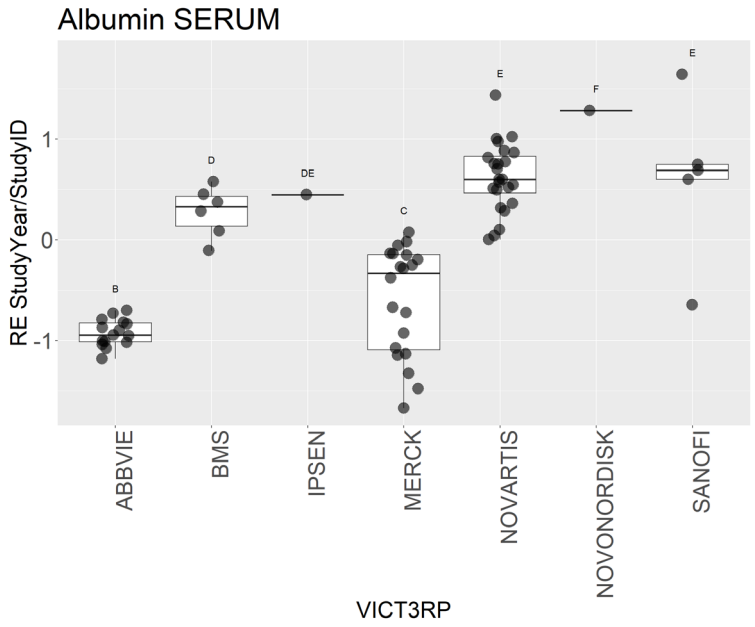


Figure 2.7.1: Differences of the random effects for STUDYID (within STUDY YEAR) by sponsor.

The residuals (ϵ_{ijk}) from the LME model (stage 2) were also analyzed in parallel using ANOVA in the same manner as the random effects, to assess the hypothesis that some of the variability in the residuals—after accounting for study-to-study differences—could be attributed to the sponsor. None of the tests yielded statistically significant results, thereby confirming that the sponsor influences only the component of variability associated with the study.

GLP studies

In the VICT3R database, studies conducted according to GLP standards were approximately five times more numerous than non-GLP studies, with data available for up to 100 studies in several laboratory tests. However, no significant differences between the two groups were identified that could be attributed to study-to-study variability.

pH of vehicle

Based on our inclusion criteria, between 3 and 9 studies were available for each endpoint, with the pH of the treatment vehicle reported in the database. Of the 45 linear regression models tested, 4 demonstrated a plausible association between BLUPs and the pH of the treatment vehicle ($p < 0.001$, $R^2 > 0.15$, with at least 8 data points), all observed on red blood cell indices, including hematocrit (HCT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and mean corpuscular hemoglobin concentration (MCHC).

Results from the models for MCHC and MCV are depicted in figure 2.7.2. MCHC shows a clear positive correlation with the pH of the vehicle (in the Acidic Range pH 2-6), and in the same range there is a visible downward trend for MCV values. A single study at pH 8, then in the alkaline range, shows a different trend, reversed.

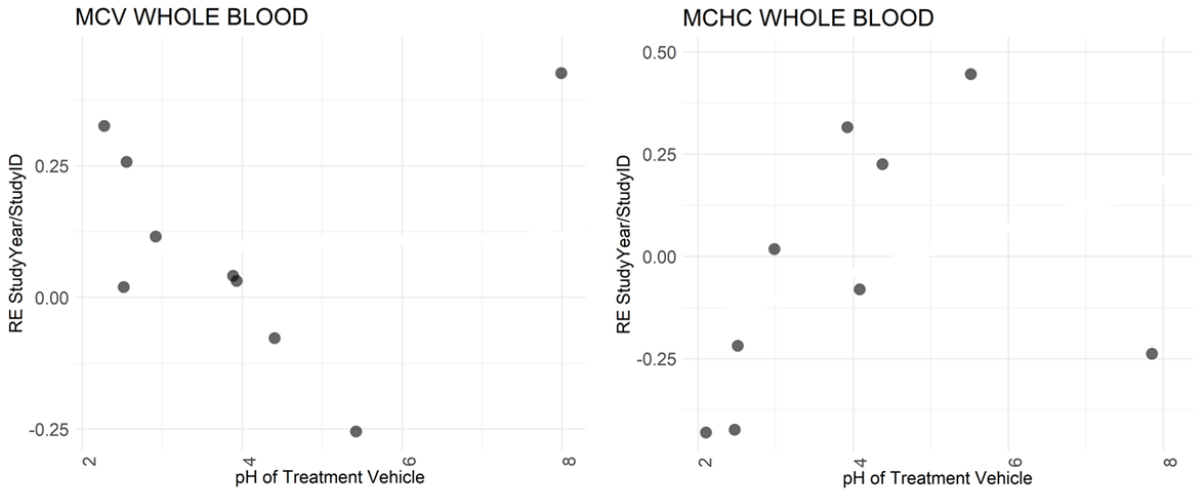


Figure 2.7.2: Differences of the random effects for STUDYID (within STUDY YEAR) by pH of vehicle.

These alterations may be an indirect consequence of local irritation within the gastrointestinal tract, The administration of acidic vehicles can induce a dose-dependent inflammatory response and physiological stress, which in turn can trigger systemic effects, including fluid and electrolyte shifts. The specific pattern of a decreased MCV with a concurrent increase in MCHC is the classic hematological signature of red blood cell dehydration. Such interesting finding will be eventually confirmed when and updated version of the database containing more records becomes available.

Study type – dose regimen.

Information regarding study type, and dose regimen in particular, was available for only a limited number of studies within the subset of data that met our inclusion criteria. For most laboratory tests, comparisons were made between 40 studies with repeated dose regimens and 10 studies with single dose regimens. Statistically significant differences were observed, but studies from the underrepresented group were all from a single sponsor, hence unreliable.

2.8 References

Barr, D. B., Wilder, L. C., Caudill, S. P., Gonzalez, A. J., Needham, L. L., & Pirkle, J. L. (2005). A urinary biomarker-based survey of human exposure to selected organophosphate pesticides in the United States. *Environmental Health Perspectives*, *113*(2), 186–191.

Kort M de, Weber K, Wimmer B, 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

Ruíz JS, López OAM, Ramírez GH, Hiriart JC (2023) Generalized Linear Mixed Models with Applications in Agriculture and Biology. https://doi.org/10.1007/978-3-031-32800-8

Yu Z, Guindani M, Grieco SF, et al (2022) Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. Neuron 110:21–35. https://doi.org/10.1016/j.neuron.2021.10.030

Weber K, Razinger T, Hardisty JF, et al (2011) Differences in rat models used in routine toxicity studies. Int J Toxicol 30:162–73. https://doi.org/10.1177/1091581810391818

3 Changepoint analysis [TUDO]

3.1 Changepoint analysis

The goal of this analysis is to determine a timepoint in the past such that measurements for a covariate are similar enough to build virtual control groups. If such a timepoint is identified, measurements from that point onward can be considered as exchangeable with respect to the corresponding covariate. Further, this approach can be used for data curation, since a changepoint is typically associated with a change in the study design.

We follow an established statistical approach that identifies both the optimal number of changepoints and their locations within a given dataset. This approach uses a likelihood-based model while penalizing larger numbers of changepoints. In the underlying model, a constant function is assumed between changepoints [1].

In the real-world scenario, the most recent changepoint indicates the start of the time interval with similar values.

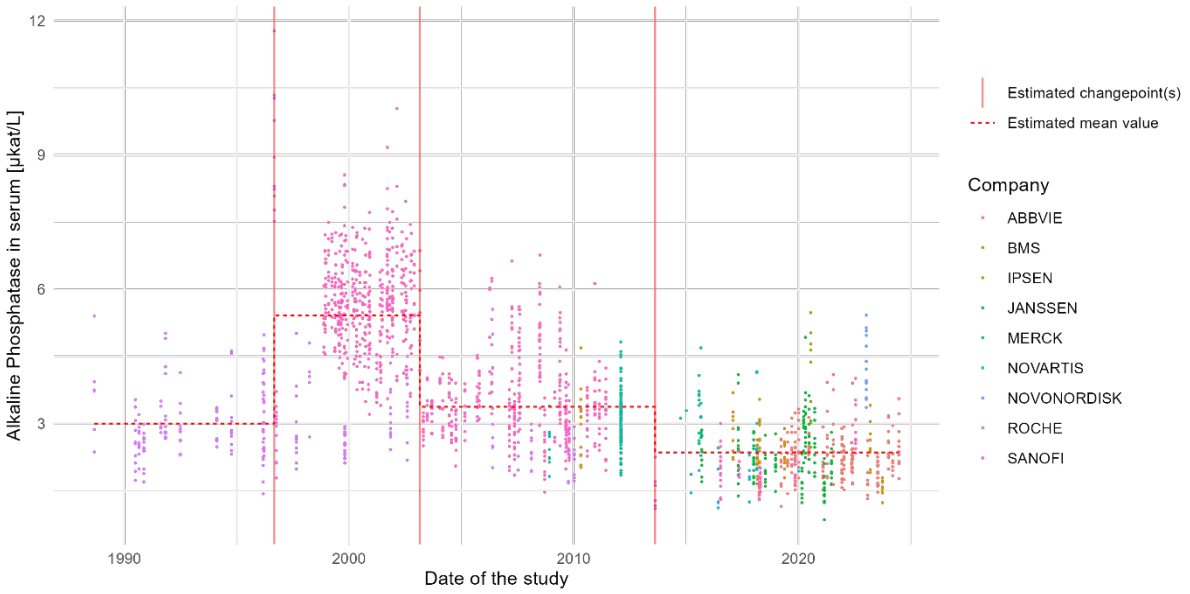


Figure 3.1: Alkaline Phosphatase in serum of male Sprague Dawley rats measured on days 25-35, considering a total of 1,799 rats, using data from the VICT3R database (v2.0.4).

Figure 3.1 shows values for Alkaline Phosphatase from different studies over time, with three detected changepoints, possibly affected by different diets or a change in the animal supplier. An analysis of the body weight of the rats in the segments revealed that the rats in segment two have considerably lower body weights compared to the other segments, while rats in segment 4 show the highest body weight, which may be attributed to the age of the animals.

We have performed an initial simulation study to quantify how large changes in covariate measurements between two neighbouring segments must be in order to reliably detect this changepoint.

3.2 References

[1] Killick, R., & Eckley, I. A. (2014). changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software*, 58(3), 1–19. https://www.jstatsoft.org/article/view/v058i03

4 Discussion

This report presents a comprehensive evaluation of the covariates influencing biological variability in preclinical toxicology studies, with the ultimate goal of establishing an empirical and defensible framework for the generation of robust Virtual Control Groups (VCGs). The findings are derived from several distinct, yet complementary, analytical workstreams.

The first workstream, detailed in Section 1, employed a pragmatic approach using linear regression models combined with a Leave-One-Study-Out (LOSO) cross-validation strategy on a wide range of laboratory (LB) and organ (OM) measurements. The second workstream, presented in Section 2, implemented a multi-stage statistical investigation focused exclusively on the LB domain aiming to understand what the principal sources of variance are. Despite their methodological differences, both analytical workstreams produced convergent evidence identifying a core set of non-negotiable covariates that are fundamental to VCG construction.

Both analyses unequivocally identified core biological characteristics as primary drivers of variability across endpoints. SEX and STRAIN as highly influential in both the LB and OM domains. This finding was quantitatively substantiated and refined in Section 2, and can be considered foundational, non-negotiable matching or stratification variables for any VCG algorithm.

The study's timeline—captured as START_YEAR in Section 1 and STUDYYEAR in Section 2— was identified as a significant predictor of endpoint values giving empirical evidence of systemic, low-frequency drift in the preclinical research ecosystem over time. Temporal drift could be attributable to a confluence of slowly evolving factors, including subtle genetic shifts in commercial rodent colonies, gradual changes in standard operating procedures (SOPs), evolving diet formulations, and periodic upgrades to analytical instrumentation. The observation in Section 2 of high inter-annual variability for urine samples, which could reflect systematic changes in laboratory protocols or equipment, serves as a concrete example of this phenomenon. This result provides data-driven support for the widely held best practice, also highlighted in recent literature, of constructing VCGs from recent historical data, typically within a 3- to 5-year window, to minimize the confounding effects of temporal drift and ensure maximal comparability.

To further investigate the effect of time, the changepoint analysis in section 3 provides a method to identify timepoints in the past where measurements for a covariate change. Identifying these points can define time periods during which measurements are similar enough to build virtual control groups. An example with Alkaline Phosphatase levels demonstrates how this method works in practice. The detected changepoints may be influenced by factors such as changes in diet, variations in animal suppliers, or differences in the age or body weight of the animals. A simulation study was performed to investigate how large changes in covariate measurements between neighboring segments must be to be detected reliably.

The analysis of the OM domain demonstrates that initial body weight is the primary covariate for predicting organ mass. Body weight may not be the predominant factor in LB domain assessments,

but it remains essential to include due to its established significance in many LB assessments and throughout the entire OM domain.

The COMPANY (Sponsor) variable was identified as a covariate of high importance, particularly for the LB domain. The Sponsor variable acts as a powerful composite proxy, implicitly capturing the myriad of unmeasured but internally consistent factors that define a given organization's research environment. These include harmonized protocols, single-source animal suppliers, specific diet formulations, consistent vivarium conditions, and standardized analytical platforms. The statistical significance of this variable is, in itself, a validation of the profound impact of inter-laboratory variability on preclinical data. However, Sponsor is a methodological dead-end for a generalizable VCG platform, as it cannot be applied to new sponsors or studies from organizations not present in the historical database. LMEM's nested random effects structure could be an appropriate statistical tool to capture this inter-study heterogeneity without relying on a non-generalizable label.

This investigation in stage 3 of section 2 aimed to identify sources of inter-study variability by modeling several covariates against the Best Linear Unbiased Predictors (BLUPs) derived from the primary Linear Mixed-Effects model. The analysis revealed that the study sponsor was the most significant and pervasive factor, explaining a substantial portion of the variability across 65% of the tested endpoints. Crucially, this effect was confined to the study-level random effects and did not extend to the residual error, confirming that the sponsor's influence is a systematic, inter-study source of heterogeneity rather than a source of random individual animal variation. Another noteworthy finding in this section was the association between the pH of the treatment vehicle and several red blood cell indices according to a pattern of a classic hematological signature of red blood cell dehydration. Several other covariates could not be meaningfully interpreted due to significant limitations in the available data. Housing conditions, animal suppliers, and study type (dose regimen) were severely hampered by extreme data imbalance. The analysis of GLP status showed no notable differences in terms of study level endpoints. This indicates that, for this dataset, adherence to Good Laboratory Practice did not introduce bias. However, whether GLP status affects the magnitude of this variability may require further investigation with alternative statistical methods.

A substantial proportion of the total variance in LB endpoints (often greater than 50%) is attributable to study-level effects and, most notably, unexplained residual variance, and can be viewed as stochasticity stemming from inherent biological variation between individual animals and unmeasured experimental fluctuations. It establishes a data-driven, realistic upper boundary on the precision that can be expected from any VCG matching algorithm. The magnitude of this inherent noise suggests that VCGs will be most reliable and defensible for detecting toxicological findings with strong signals that rise clearly above this floor. As a consequence, the use of VCG in presence of high-variable endpoints should be investigated more in detail in the coming years.

The analytical workstreams presented here are limited by the depth and quality of available data. Experimental confounders that may account for random and residual variance are not consistently included in standard SEND (Standard for Exchange of Nonclinical Data) datasets. This requires further efforts to enrich the VICT3R database, including manual extraction of information from individual study reports.

Conclusion

In this preliminary statistical effort of data analysis on the version 2.1.1 of the VICT3R database, which is undergoing curation and harmonization, some requirements for constructing reliable VCGs were identified. A core set of non-negotiable biological and temporal covariates— namely Sex and Strain, and Study Year— must anchor any VCG matching algorithm. Around 25% of the variance of the animal data arises from study-specific heterogeneity. Such heterogeneity was associated with the Sponsor variable, which is a composite proxy that cannot be exploited in VCG generation. The quantification of a substantial "noise floor" of irreducible variance sets realistic expectations for the use of VCGs, highlighting their probable strength in detecting clear toxicological signals over subtle, marginal changes. Advanced statistical methods to account for this variance were tested as proof of concept, and results were very much promising. However, the principal barrier to progress is not the sophistication of available statistical models, but the profound gaps in the availability of deeply curated, granular operational data, which is able to fully characterize the experimental conditions of each study, ideally to extend beyond current standardization efforts in SEND.

Although the previous generalizations provide useful context, it is important to maintain a clear distinction between endpoints. The selection of covariates for VCG generation should consider the specific endpoints planned for a study. This consideration supports the development of algorithms or software tools that enable tailored covariate selection, rather than relying solely on a predefined set. Alternatively, an approach combining essential, non-negotiable covariates with those specific to each endpoint may be employed.

Appendix

Table A.1: Table of potential covariates

In Palazzi 2024	Category	Variable description	In VICT3R	Name in VICT3R data	DM	Available values (VICT3R 2.1.1)	SEND variab. (Y/N)
Yes	Animal	Origin (for NHPs)	No	TS.TSPARMCD "SPLRNAM" and "SPLRLOC"	TS	34.77 % (SPLRNAM), 30.18 % (SPLRLOC)	Yes
Yes	Animal	Species	Yes	SPECIES	DM	100%	Yes
Yes	Animal	Body weight at the study initiation	Yes	BWSTRESN (conditioned to BWDY near 1)	BW	100%	Yes
Yes	Animal	Sex	Yes	SEX	DM	100%	Yes
Yes	Animal	Strain	Yes	STRAIN	DM	88%	Yes
Yes	Animal	Age	Yes	AGE	DM	50%	Yes
Yes	Animal	Breeder/supplier	Yes	TS.TSPARMCD "SPLRLOC"	TS	30%	Yes
Yes	Animal	Substrain	Yes	SBSTRAIN	DM	0%	Yes
Yes	Animal	Genetic drift	No				No
Yes	Animal procedures	Site and time of blood collection, method	Yes	LBDTC / LBLOC	LB	87.70%(LBDTC) , 4.31% (LBLOC)	Yes
Yes	Animal procedures	Fasting: start time, stop time, duration, access to water	No	FWTESTCD ?	FW	100%	Yes
Yes	Animal procedures	Clinical signs observed in control animals	Yes	CLSTRESC	CL	99%	Yes
Yes	Animal procedures	Volume and timing of blood collection	No	LBDTC (timing)	LB	88%	Yes
Yes	Animal procedures	Test facility location	Yes	TS.TSPARMCD "TSTFLOC"	TS	18%	Yes
Yes	Animal procedures	Test facility name	Yes	TS.TSPARMCD "TSTFNAM"	TS	11%	Yes
Yes	Animal procedures	Anesthesia, restraint method and temperature	No	MANESTH / MANAESTH ?	TS	4%	Yes

Yes	Animal procedures	Handling (cannot be measured)	No				No
Yes	Animal procedures	Surgeries, catheterization, transport	No				No
Yes	Animal procedures	Veterinary intervention	No				No
Yes	Animal procedures	Quarantine, prestudy health screen practices	No				No
Yes	Study Design and Vehicle	Blood collection/necropsy day/s	Yes	LBDY AND MIDY	LB and MI	98.13 % (LBDY), 85.32 % (MIDY)	Yes
Yes	Study Design and Vehicle	Dose volume of the vehicle	Yes	EXVAMT and EXVAMTU	EX	96.12 % (EXVAMT), 96.15 % (EXVAMTU)	Yes
Yes	Study Design and Vehicle	Route, procedure, duration, and frequency of administration	Yes	EXROUTE, EXDOSFRM, EXDUR, EXMETHOD, EXDOSFRQ	EX	92.53% (EXROUTE), 77.99% (EXDOSFRM), 0.02% (EXDUR), 25.87% (EXMETHOD), 92.37% (EXDOSFRQ)	Yes
Yes	Study Design and Vehicle	Study start date	Yes	TS.TSPARMCD "STSTDTC" and "EXPSTDTC"	TS	67.95% (EXPSTDTC), 52.09% (STSTDTC)	Yes
Yes	Study Design and Vehicle	Excipients used in the vehicle control	Yes	TS.TSPARMCD "TRTV" and "TRTVCAT" / EX.EXTRTV	TS /EX	36.65% (TRTV), 34.01% (TRTVCAT), 96.50% (EXTRTV)	Yes
Yes	Study Design and Vehicle	Percentage of each excipient	Yes	EXTRTV	EX	97%	Yes
Yes	Study Design and Vehicle	Study duration	Yes	TS.TSPARMCD "DOSDUR"	TS	34%	Yes
Yes	Study Design and Vehicle	Study length	Yes	TS.TSPARMCD "SLENGTH"	TS	25%	Yes

Yes	Microenvironment	Temperature	Yes	TS.TSPARMCD "ENVTEMP" and "TEMP"	TS	8.10% (ENVTEMP), 9.01% (TEMP)	Yes
Yes	Microenvironment	Time of feeding	No	FWDTC / FWDY?	FW	100%	yes
Yes	Microenvironment	Housing- single vs group	Yes	TS.TSPARMCD "HOUSEGRP"	TS	38%	Yes
Yes	Microenvironment	Cage design	Yes	TS.TSPARMCD "HOUSETYP"	TS	34%	Yes
Yes	Microenvironment	Feed- type and amount	Yes	TS.TSPARMCD "DIET"	TS	21%	Yes
Yes	Microenvironment	Humidity	Yes	TS.TSPARMCD "HUMIDT"	TS	18%	Yes
Yes	Microenvironment	Light intensity in the room, dark/light cycle period	Yes	TS.TSPARMCD "LIGHT"	TS	15%	yes
Yes	Microenvironment	Type of cage bedding	Yes	TS.TSPARMCD "BEDDING"	TS	3%	Yes
Yes	Microenvironment	Environmental Enrichment (impact on animal stress levels)	No				No
Yes	Microenvironment	Number of animals per cage (density)	No				No
Yes	Microenvironment	Cage changing frequency	No				No
Yes	Microenvironment	Ventilation at the cage level					No
Yes	Microenvironment	Cage rotations	No				No
Yes	Clinical Pathology	Instruments	Yes	LBMETHOD / LBCAT	LB	37.72 % (LBMETHOD), 1.04 % (LBCAT)	Yes
Yes	Clinical Pathology	Blood collection tubes	No	LBSPEC (for specimen collected)	LB	96%	Yes
Yes	Clinical Pathology	Units	Yes	LBSTRESU	LB	84%	Yes
Yes	Clinical Pathology	Methods	Yes	LBMETHOD	LB	38%	Yes
Yes	Clinical Pathology	Biochemical kits	No				No

Yes	Clinical Pathology	Blood sample processing/handling	No				No
Yes	Clinical Pathology	Type of anticoagulant	No				No
Yes	Clinical Pathology	Sample storage temperature and duration	No				No
Yes	Anatomic Pathology	Euthanasia and exsanguination methods	Yes	TS.TSPARMCD "MTHTRM"	TS	6%	Yes
Yes	Anatomic Pathology	Processing parameters (embedding, sectioning, staining)	No	MIMETHOD	MI	1%	Yes
Yes	Anatomic Pathology	Fixation time	No				No
No	Anatomic Pathology	Tissue/fixative ratio	No				No
No	Anatomic Pathology	Trimming procedures	No				No
No	Anatomic Pathology	Pathologist's experience	No				No
No	Others	Pre-values					Yes

Table A.2: Laboratory endpoints available in the VICT3R database pass the inclusion criteria defined in section 2. Tests are organized by specimen type, test name, total number of animals assessed, and number of studies included. This table provides a comprehensive overview of the diversity and scale of preclinical data underlying the statistical analyses. Log transformation of values when skew distribution was detected and model failures when data was insufficient are also shown.

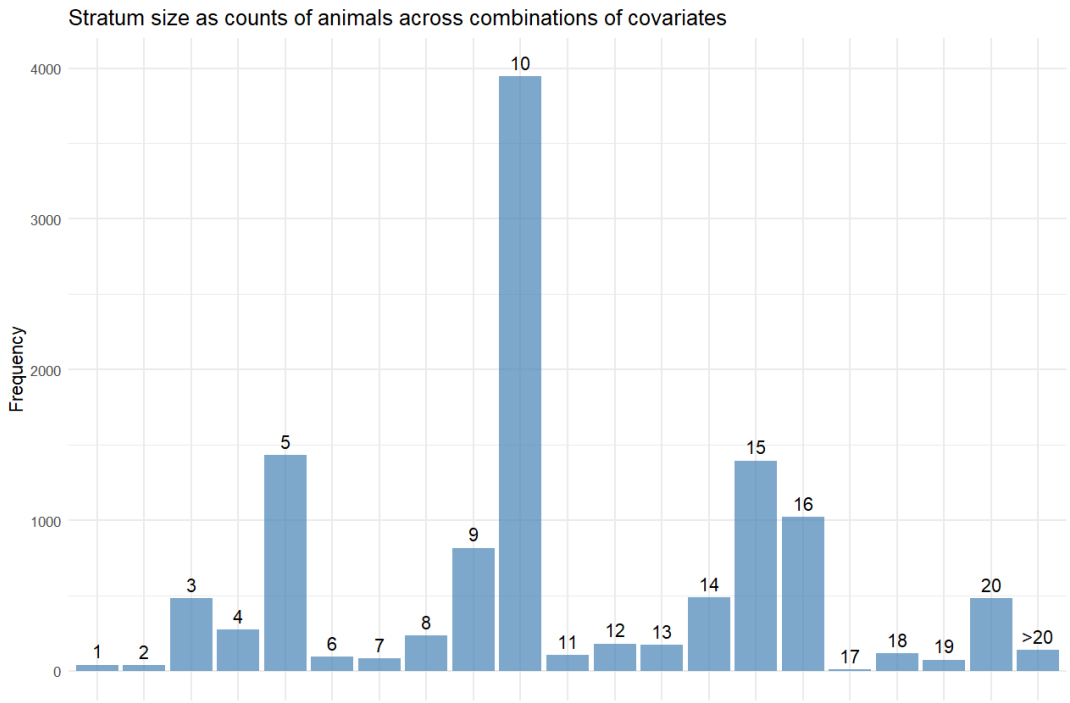
SPECIMEN	TEST	Skew (Log)	N. Animals	N. Studies	Model
PLASMA	Activated Partial Thromboplastin Time	N	1920	101	OK
PLASMA	Alanine Aminotransferase	Y	656	26	OK
PLASMA	Albumin	N	676	26	OK
PLASMA	Albumin Globulin	Y	503	21	OK
PLASMA	Alkaline Phosphatase	Y	675	26	OK
PLASMA	Aspartate Aminotransferase	Y	674	26	OK
PLASMA	Bilirubin	Y	335	16	OK
PLASMA	Calcium	Y	675	26	OK
PLASMA	Chloride	Y	677	26	OK
PLASMA	Cholesterol	Y	676	26	OK
PLASMA	Creatine Kinase	Y	139	6	OK
PLASMA	Creatinine	Y	676	26	OK
PLASMA	Fibrinogen	N	1818	95	OK
PLASMA	Gamma Glutamyl Transferase	N	187	10	OK
PLASMA	Globulin	Y	539	22	OK
PLASMA	Glucose	Y	672	26	OK
PLASMA	Glutamate Dehydrogenase	Y	373	15	OK
PLASMA	Magnesium	Y	185	5	FAIL
PLASMA	Phosphate	N	677	26	OK
PLASMA	Potassium	Y	677	26	OK
PLASMA	Protein	Y	676	26	OK
PLASMA	Prothrombin Time	Y	2085	105	OK
PLASMA	Sodium	N	677	26	OK
PLASMA	Triglycerides	Y	674	26	OK
PLASMA	Urea	N	497	18	OK
PLASMA	Urea Nitrogen	Y	109	5	OK
SERUM	Alanine Aminotransferase	Y	2008	93	OK
SERUM	Albumin	N	2092	99	OK
SERUM	Albumin Globulin	N	1442	63	OK
SERUM	Alkaline Phosphatase	Y	2093	99	OK
SERUM	Aspartate Aminotransferase	Y	2036	94	OK
SERUM	Bicarbonate	N	992	42	OK
SERUM	Bile Acid	Y	261	13	FAIL
SERUM	Bilirubin	Y	1808	97	OK

SPECIMEN	TEST	Skew (Log)	N. Animals	N. Studies	Model
SERUM	Calcium	N	2092	99	OK
SERUM	Chloride	N	2094	99	OK
SERUM	Cholesterol	Y	2084	98	OK
SERUM	Creatine Kinase	Y	1009	43	FAIL
SERUM	Creatinine	N	2094	99	OK
SERUM	Gamma Glutamyl Transferase	Y	255	18	OK
SERUM	Globulin	Y	1543	68	OK
SERUM	Glucose	Y	2094	99	OK
SERUM	Glutamate Dehydrogenase	Y	421	24	OK
SERUM	Hemolytic Index	Y	359	12	FAIL
SERUM	Lipemic Index	N	361	12	FAIL
SERUM	Magnesium	Y	633	25	FAIL
SERUM	Phosphate	N	2120	100	OK
SERUM	Potassium	Y	2084	99	OK
SERUM	Protein	N	1575	69	OK
SERUM	Sodium	N	2092	99	OK
SERUM	Total Protein	Y	526	31	OK
SERUM	Triglycerides	Y	2084	99	OK
SERUM	Urea	Y	1045	49	OK
SERUM	Urea Nitrogen	Y	1041	50	OK
URINE	Calcium	Y	102	5	OK
URINE	Chloride	Y	268	11	OK
URINE	Creatinine	Y	791	30	OK
URINE	Glucose	Y	114	6	FAIL
URINE	Ketones Ketone Bodies	Y	155	26	OK
URINE	Leukocytes	Y	105	6	OK
URINE	Osmolality	N	177	7	OK
URINE	Potassium	Y	188	8	OK
URINE	Protein	Y	920	49	OK
URINE	Protein Creatinine	Y	194	8	FAIL
URINE	Sodium	Y	290	12	OK
URINE	Specific Gravity	N	383	13	FAIL
URINE	Urine Volume	Y	1402	53	OK
URINE	Urobilinogen	Y	270	10	FAIL
URINE	Weight	Y	190	10	FAIL
WHOLE BLOOD	Activated Partial Thromboplastin Time	N	187	14	FAIL
WHOLE BLOOD	B Lymphocytes	Y	188	10	OK
WHOLE BLOOD	Basophilic Granulocytes	Y	145	13	FAIL
WHOLE BLOOD	Basophils	Y	2428	121	OK

SPECIMEN	TEST	Skew (Log)	N. Animals	N. Studies	Model
WHOLE BLOOD	Basophils Leukocytes	Y	704	31	FAIL
WHOLE BLOOD	Basophils Total Cells	Y	405	31	OK
WHOLE BLOOD	CYTOTOXIC T Lymphocytes	N	132	8	FAIL
WHOLE BLOOD	Eosinophils	Y	2930	136	OK
WHOLE BLOOD	Eosinophils Leukocytes	Y	819	32	FAIL
WHOLE BLOOD	Eosinophils Total Cells	Y	405	31	OK
WHOLE BLOOD	Erythrocytes	N	3092	138	OK
WHOLE BLOOD	Erythrocytes Distribution Width	Y	2156	103	OK
WHOLE BLOOD	Fibrinogen	Y	247	7	FAIL
WHOLE BLOOD	Granulocytes	Y	157	13	FAIL
WHOLE BLOOD	HCT	N	2985	133	OK
WHOLE BLOOD	Hemoglobin	N	3092	138	OK
WHOLE BLOOD	Hemoglobin Distribution Width	Y	464	24	OK
WHOLE BLOOD	High Absorption Reticulocytes	Y	398	30	FAIL
WHOLE BLOOD	Large Unstained Cells	Y	2478	119	OK
WHOLE BLOOD	Large Unstained Cells Leukocytes	Y	717	28	FAIL
WHOLE BLOOD	Large Unstained Cells Total Cells	Y	400	31	OK
WHOLE BLOOD	Leukocytes	Y	2697	127	OK
WHOLE BLOOD	Low Absorption Reticulocytes	N	398	30	FAIL
WHOLE BLOOD	Lymphocytes	Y	3000	136	OK
WHOLE BLOOD	Lymphocytes Leukocytes	Y	834	32	FAIL
WHOLE BLOOD	Lymphocytes Total Cells	Y	407	31	OK
WHOLE BLOOD	MCH	N	3010	135	OK
WHOLE BLOOD	MCHC	Y	3089	138	OK
WHOLE BLOOD	MCV	N	3092	138	OK
WHOLE BLOOD	MCV Reticulocytes	N	259	18	OK
WHOLE BLOOD	Mean Platelet Volume	N	1310	69	OK
WHOLE BLOOD	Medium Absorption Reticulocytes	Y	398	30	FAIL
WHOLE BLOOD	Monocytes	Y	2984	136	OK
WHOLE BLOOD	Monocytes Leukocytes	Y	815	32	FAIL
WHOLE BLOOD	Monocytes Total Cells	Y	436	32	OK
WHOLE BLOOD	Myeloperoxidase Index	Y	132	13	FAIL
WHOLE BLOOD	Natural Killer Cells	Y	227	11	OK
WHOLE BLOOD	Neutrophils	Y	2695	118	OK
WHOLE BLOOD	Neutrophils Leukocytes	Y	834	32	FAIL
WHOLE BLOOD	Neutrophils Total Cells	Y	407	31	OK
WHOLE BLOOD	Neutrophils Segmented	Y	139	5	FAIL
WHOLE BLOOD	Platelet Distribution Width	N	556	33	OK
WHOLE BLOOD	Platelet Hematocrit	N	398	20	OK

SPECIMEN	TEST	Skew (Log)	N. Animals	N. Studies	Model
WHOLE BLOOD	Platelets	N	2969	136	OK
WHOLE BLOOD	Prothrombin Time	Y	177	13	FAIL
WHOLE BLOOD	Reticulocyte Hemoglobin Content	N	478	32	FAIL
WHOLE BLOOD	Reticulocytes	N	2631	125	OK
WHOLE BLOOD	Reticulocytes Erythrocytes	Y	1281	49	OK
WHOLE BLOOD	Reticulocytes Total cells	Y	405	31	OK
WHOLE BLOOD	T Lymphocytes	Y	228	11	OK
WHOLE BLOOD	WBC Basophil Channel	Y	241	17	FAIL
WHOLE BLOOD	WBC Peroxidase Channel	Y	241	17	FAIL

Figure A.1: Frequency of animal subgroup



The following displays the frequency of animal subgroup sizes after stratifying by experimental covariates (interaction stdyear-studyid, sex, route, strain). While a group size of 10 is the most common, the stratification step also created many very small subgroups, some containing only one or two animals.

In presence of such tiny subgroups the statistical models fail to converge, and as such no results are available for such tests.

Table A.4: The 30 most frequent tests in the LB domain at LBDY=28+/-3, from animals that were only required to have information for sponsor and body weight.

LBSPEC	LBTEST	n
WHOLE BLOOD	MCHC	10859
WHOLE BLOOD	Leukocytes	10007
WHOLE BLOOD	Lymphocytes	9712
WHOLE BLOOD	Monocytes	9695
WHOLE BLOOD	Eosinophils	9512
WHOLE BLOOD	Erythrocytes	9502
WHOLE BLOOD	Hemoglobin	9469
WHOLE BLOOD	MCV	9434
WHOLE BLOOD	Platelets	9373
WHOLE BLOOD	MCH	9339
WHOLE BLOOD	HCT	9303
WHOLE BLOOD	Neutrophils	8671
WHOLE BLOOD	Basophils	8567
WHOLE BLOOD	Reticulocytes	8063
WHOLE BLOOD	Large Unstained Cells	7899
SERUM	Glucose	7896
SERUM	Sodium	7698
SERUM	Calcium	7682
SERUM	Chloride	7646
SERUM	Potassium	7618
SERUM	Aspartate Aminotransferase	7401
SERUM	Creatinine	7340
SERUM	Phosphate	7334
URINE	pH	7254
SERUM	Triglycerides	7251
SERUM	Alkaline Phosphatase	7239
SERUM	Alanine Aminotransferase	7193
SERUM	Albumin	7177
SERUM	Bilirubin	6887
SERUM	Cholesterol	6688

Table A.4: The 30 most frequent tests in the OM domain at OMDY=28+/-3, from animals that were only required to have information for sponsor and body weight.

OMSPEC	OMTEST	n
liver	Weight	4918
kidney	Weight	4887
spleen	Weight	4858
heart	Weight	4822
gland, adrenal	Weight	4793
brain	Weight	4742
thymus	Weight	4682
gland, pituitary	Weight	3430