101172693 – VICT3R

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

WP2 – Data management, curation and quality control (RDT)

D2.3 Data requirements and quality criteria

Lead contributor:	Frank Bringezu (frank.bringezu@merckgroup.com) (15-MKDG)
Other contributors	Stefanie Will (15-MKDG) Jan-Paul Weidenbusch (15-MKDG) Sylvia Escher (3-ITEM) Nelly Simetska (3-ITEM) Jenny Irwan (3-ITEM) Sascha Fischer (3-ITEM) Rupert Kellner (3-ITEM)
Due date:	28/02/2025
Delivery date:	06/06/2025
Dissemination level:	Public

Co-funded by the European Union

Contents

Document History

Version	Date	Description
V0.1	28 Feb 2025	First internal draft
V0.2	05 Mar 2025	Second internal draft
V0.3	07 Mar 2025	Third internal draft
V0.4	11 Mar 2025	Version for formal review
V0.5	14 Mar 2025	Version after formal review reviewer 1
V0.6	27 Mar 2025	Version after formal review reviewer 2
V0.7	28 Mar 2025	Version for Consortium Review
V1.0	6 Jun 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)

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 document summarises the work performed in Work Package 2 (WP2) of the VICT3R project, focusing on data collection, curation, and quality control to support the development of the VICT3R database. Initial contributions of control group data in SEND format from industry partners were integrated into a development database. The raw data underwent analysis and standardization to create a well-curated version suitable for project purposes. A data gap analysis identified minimum requirements for metadata and primary results. The standardization process included data collection from the VICT3R data lake, integration of datasets, and application of data cleaning and validation protocols. The efforts aimed to enhance the overall quality of data, facilitating reliable analysis and decision-making.

1. Introduction

This deliverable is focused on the results obtained from the initial data processing of in-kind contributions of control group data in SEND format provided by industry. These data have been integrated with data from the VICOG database developed during the pilot phase of the project. A standardization and data curation were applied to ensure that metadata is accurate and consistent. This included study metadata collected in the trial summary domain as well as animal metadata collected in the demographics domain of SEND. The standardization should improve the overall quality of the data and facilitates reliable analysis and decision-making.

2. Methods

The standardization workflow applied involves several key steps to ensure consistency and accuracy in data management. The following steps have been applied to ensure high quality, consistency and relevance of the data collected.

- Data Collection: Gather data from the VICT3R data lake (cf. Deliverable 2.1 Data Lake for secure data exchange)
- Data Integration: Merge standardized datasets to create the shared import files for the VICT3R DB (cf. D5.1 grit42 platform for the VICT3R DB)
- Data Standardization: Convert data into a common format or structure, including normalization of units, date formats, and application of standardized terminology.
- Data Cleaning: Identify and correct errors, inconsistencies, and duplicates within the datasets collected
- Validation: Ensure the standardized data meets predefined quality standards and is ready for use in the project work.

For the workflows applied, we employed a semiautomated analysis using R with standard libraries and additional libraries as appropriate [1,2].

3. Results

3.1 Data Collection

The VICT3R data lake was used to collect all data provided by the project partners. After downloading the data, a local repository was compiled and used for further steps of the data curation workflow. The life cycle was adjusted in the data lake to indicate that processing of the data is ongoing.

3.2 Data Integration

The principal workflow applied for data integration is outlined in Figure 1. For data standardization, the individual SEND files provided for the studies were compared with the SEND standard. Any additional columns not included in the SEND standard were removed. If additional columns are defined in SENDIG 3.1 these have been added. In this step individual studies from sponsors were consolidated into a sponsor-specific repository. After applying the SENDIG 3.1 filter, all data were incorporated into the Development database for WP2, which includes the various SEND domains as available.

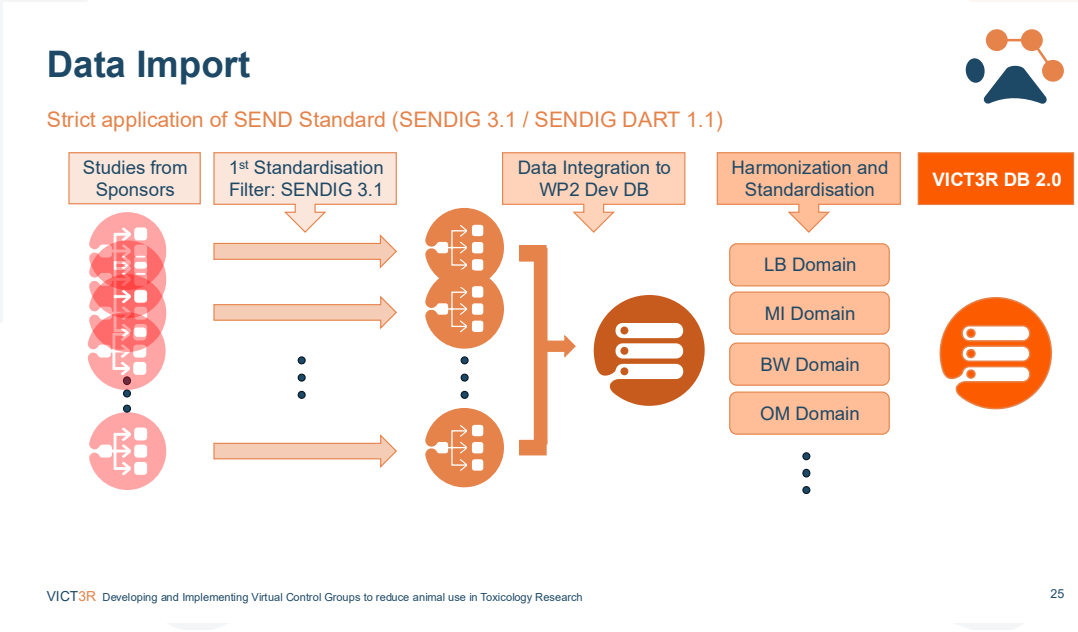

Figure 1 Data Integration Workflow applied in VICT3R WP2

Table 1 presents the in-kind data contributions received from various sponsors, taking into account the data collected during the Pilot phase of the project and the data acquired through the VICT3R data lake up to the fourth month of the project. A total of 5.881 studies have been incorporated into the development database.

Company	No of Studies
ABBVIE	20
AMGEN	20
ASTRAZENECA	10
BAYER AG	444
BMS	40
BOEHRINGER	13
IPSEN	3
JANSSEN	93
MERCK	323
NOVARTIS	1142
NOVONORDISK	13
PFIZER	29
ROCHE	2268
SANOFI	1463

Table 1 In kind data contribution by VICT3R Sponsor

Metadata enables data sharing and integration across various systems and platforms. However, it must be standardized to allow VICT3R systems to comprehend, search for, and filter specific information based on established criteria. This standardization should ensure that the same type of data is identified across different sponsors, assign common terminology, perform necessary translations, and apply appropriate calculations to achieve uniform units, values, and terminologies.

Domain Coverage

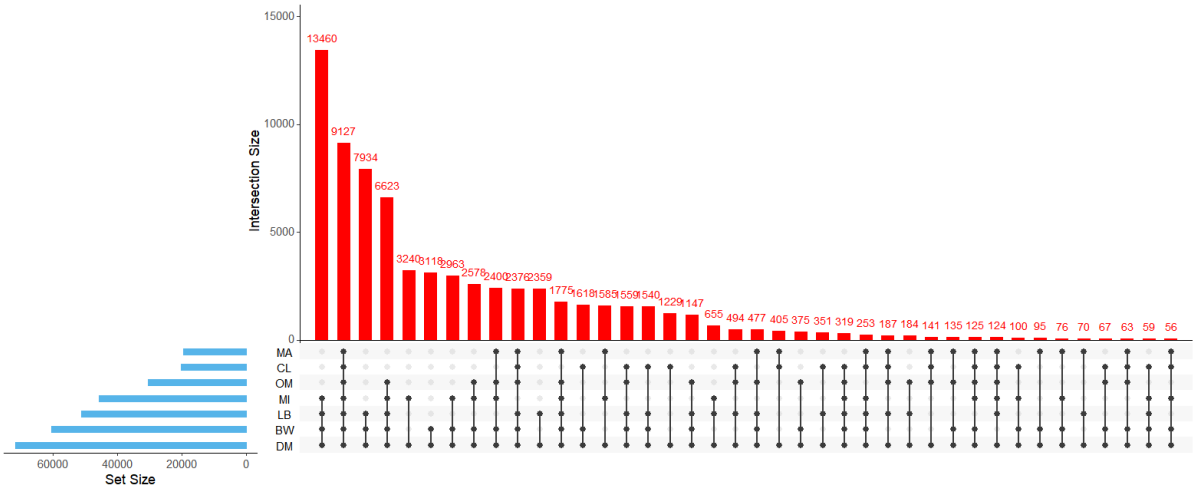

Figure 2 Upset plot of the domain coverage for the RAT dataset in the VICT3R DB version 2.0

After integration of all data provided the domain coverage was analyzed using upset plots as shown in Figure 2. The database contains records for more than 70k rats. As shown in Figure 2, the coverage is highly heterogeneous and complete coverage of the domains indicated is provided for 9127 animals.

3.3 Data Standardization

The data curation and standardization process employed the tidyverse ecosystem[1]. Tidyverse is an ecosystem of R packages designed for data science that shares a common philosophy and syntax, promoting a cohesive and efficient workflow for data manipulation, visualization, and analysis. Core packages within the tidyverse include ggplot2[2] for data visualization, dplyr[3] for data manipulation, tidyr for data tidying, readr[4] for data import, purrr for functional programming, and tibble for modern data frames. Tidyverse emphasizes clarity and simplicity, allowing users to write code that is both

expressive and easy to understand. Its design principles encourage the use of tidy data, where each variable is a column, each observation is a row, and each type of observational unit forms a table. This structured approach facilitates seamless integration across packages, making it easier for users to perform complex data analyses and create high-quality visualizations. Overall, the tidyverse has become a fundamental tool for data scientists and statisticians using R, fostering a vibrant community and extensive resources for learning and support.

Following the data integration process, the metadata for individual animals gathered in the demographic domain displays a varying degree of completeness, as illustrated in Table 2.

Name	Label	Degree of filling / [%]
STUDYID	Study Identifier	100%
USUBJID	Unique Subject Identifier	100%
RFSTDTC	Subject Reference Start Date/Time	85.54%
RFENDTC	Subject Reference End Date/Time	66.87%
RFXSTDTC	Date/Time of First Study Treatment	48.07%
RFXENDTC	Date/Time of Last Study Treatment	31.15%
BRTHDTC	Date/Time of Birth	14.31%
AGE	Age	46.53%
AGETXT	Age Range	14.56%
AGEU	Age Unit	46.53%
SEX	Sex	100%
SPECIES	Species	95.01%
STRAIN	Strain/Substrain	94.94%

Table 2 Data availability of selected Parameters from the demographics domain (DM) after data integration from the VICT3R Data Lake

Each study and each animal are annotated with unique STUDYID or USUBJID, respectively. For each animal information about SEX is available. However, other parameters show a lower degree of completeness.

Demographics Domain

For the demographics domain, the data curation focused on SEX, AGE, DTC values, SPECIES and STRAIN. Table 3 lists the standardization applied for the parameters outlined.

Name	Label	Standardization
STUDYID	Study Identifier	n.a.
USUBJID	Unique Subject Identifier	n.a.
RFSTDTC	Subject Reference Start Date/Time	PosixCT 1
RFENDTC	Subject Reference End Date/Time	PosixCT 1
RFXSTDTC	Date/Time of First Study Treatment	PosixCT 1
RFXENDTC	Date/Time of Last Study Treatment	PosixCT 1
BRTHDTC	Date/Time of Birth	PosixCT 1
AGE	Age	Numerical

1 The PosixCT format, short for POSIX Compliant Time, is a standard way to represent time in computing, specifically in Unix-like operating systems. It is based on the POSIX (Portable Operating System Interface) standard, which defines a set of operating system interfaces for software compatibility.

AGETXT	Age Range	n.a.
AGEU	Age Unit	Days
SEX	Sex	Character (M/F)
SPECIES	Species	Standardized Names
STRAIN	Strain/Substrain	Standardized Names

Table 3 Parameter specific standardization applied

Species were standardized by implementing controlled vocabularies. During curation of species names, most frequent names have been selected. In addition, information on STRAINs provided additional input for correct assignment. After curation the list of studies per curated SPECIES and VICT3R Sponsor could be obtained. Table 4 provides a tabular summary of studies per species and sponsor.

	ABBVIE	AMGEN	ASTRAZENECA	BAYER AG	BMS	BOEHRINGER	IPSEN	JANSSEN	MERCK	NOVARTIS	NOVONORDISK	PFIZER	ROCHE	SANOFI
DOG	0	0	3	62	8	5	2	11	62	0	2	0	473	269
GUINEA PIG	0	0	0	0	0	0	0	0	0	0	0	0	0	2
MINIPIG	0	0	0	0	0	0	0	2	1	0	3	0	0	3
MONKEY	0	20	3	51	8	0	0	23	35	0	1	0	302	174
MOUSE	0	0	1	31	5	2	0	10	61	0	0	0	0	140
RABBIT	0	0	0	1	0	0	0	5	2	0	0	0	0	23
RAT	20	0	3	299	19	6	1	42	162	1142	7	29	1493	852

Table 4 Studies provided by Sponsors per species in VICT3R DB 2.0

For standardization of date columns using the POSIXct format we employed the lubridate R package for the manipulation and analysis of date-time data[5]. Lubridate provides intuitive functions for converting various date formats into POSIXct objects, managing time zone differences effectively, and performing arithmetic operations on date-time values. It allows users to easily format output for presentation and handle missing date-time values without complications. As part of the Tidyverse ecosystem, lubridate integrates seamlessly with other packages, enhancing workflows in data analysis and ensuring consistent, accurate representation of date-time data across different applications and locales.

The standardized values obtained for each DTC input was stored in the respective data columns for the input columns and the input values have been changed to '*_orig' in the demographics domain.

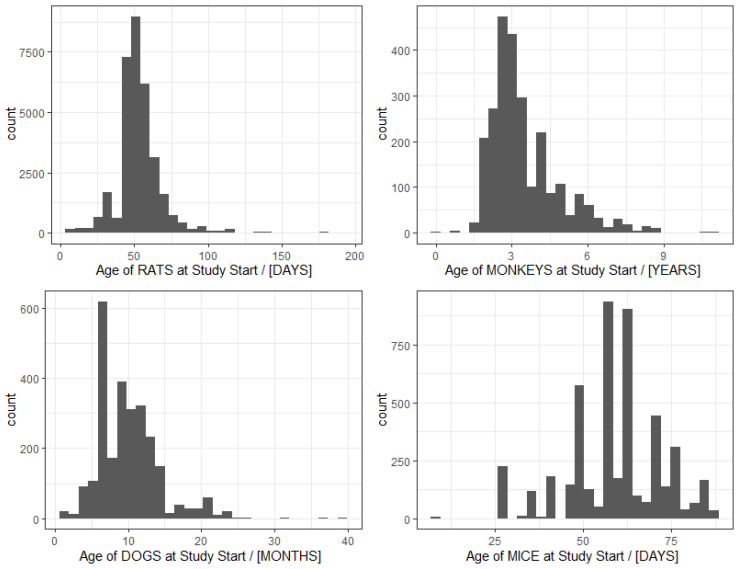

Figure 3 Curated age distribution at study start for rats, mice, dogs and monkeys as indicated. The aged was calculated based on the data available from the input data sources published in demographics and trial summary in AGE, AGEU, AGETXT, BRTHDTC, and RFSTDTC.

To standardize AGE, we utilized various data sources, including AGE input values, units from AGEU, AGETXT information, and curated DTC data for birthday and study start dates. The process followed a sequential approach, prioritizing calculated date differences based on birthday and study start when available. If those were not accessible, AGETXT, AGE input, and AGEU were curated and standardized to yield results in days, weeks, or months as indicated in AGEU. The values obtained after integration are populated as AGE_orig, AGEU_orig, AGETXT_orig.

For further standardization to a common unit, AGE_orig was converted to days. The standardized results were populated in AGE and AGEU, respectively. For animals which an age rage was supplied, the age was standardized to the mid-point of the range. When age was supplied with units of months, the conversion factor to days was 30 days/month.

Input information for AGE is reported for about 46.53% of the animals in the Data Lake. After considering all data that are considered relevant for age calculation including AGETXT and DTC values, AGE values could be determined for 47.30% of the total number of animals. Figure 3 shows the curated age at study start for the different curated species as indicated.

Trial Summary Domain

The trial summary domain typically includes data on trial design, subject involved in the trial, and information on dosing, housing and environmental conditions. Standardization at this stage of the VICT3R project included route of administration, species, strain, housing conditions and type.

ROUTE

The route of administration in *in vivo* studies is crucial for ensuring the reproducibility and reliability of pharmacological and toxicological data. Therefore, it is also an important factor for statistical evaluation and further for the selection of appropriate virtual controls. Common routes of administration include oral, intravenous, intramuscular, subcutaneous, and inhalational, each presenting unique absorption profiles and systemic exposure levels. Further subcategories of these major routes have been defined and used for standardization of the information provided for each study.

The information for the route of administration is populated in the trial summary domain. The values per study are recorded using TSPARMCD = 'ROUTE' and additional information is also available from TSPARMCD = 'STITLE'.

However, after integration of all data and robust data mining for route information, we have identified 4105 studies with information on route and 1776 studies without data for ROUTE being published in TS. In addition, combination studies applying different routes have been identified. Based on this initial result, we extracted information for the records without or with mixed route from the study titles and enriched the TS domain data with this additional information on the route applied. The values for the original routes collected and subsequently enriched have been populated as TSPARMCD="ROUTE_orig" and TSPARM="Route of Administration as submitted".

Figure 4 shows the Route of administration as submitted by the sponsors in the trial summary domain after integration of all data available. No route was reported for 1886 studies. In addition, there are routes reported that cannot be standardize, e.g. '.'.

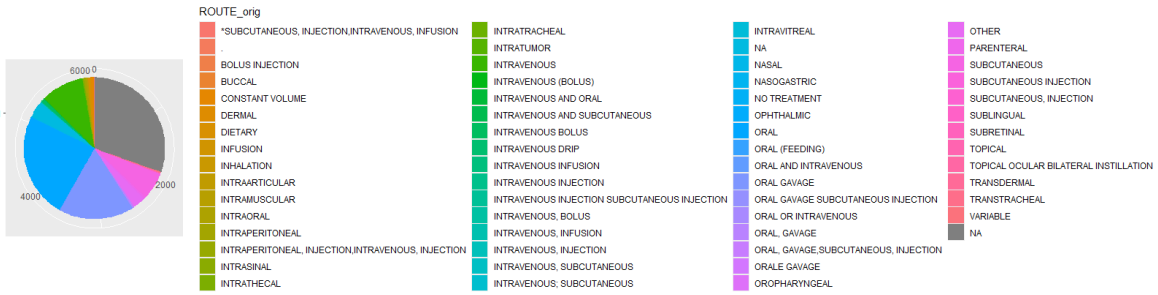

Figure 5 Route of administration as submitted by the sponsors in the trial summary domain after integration of all data available

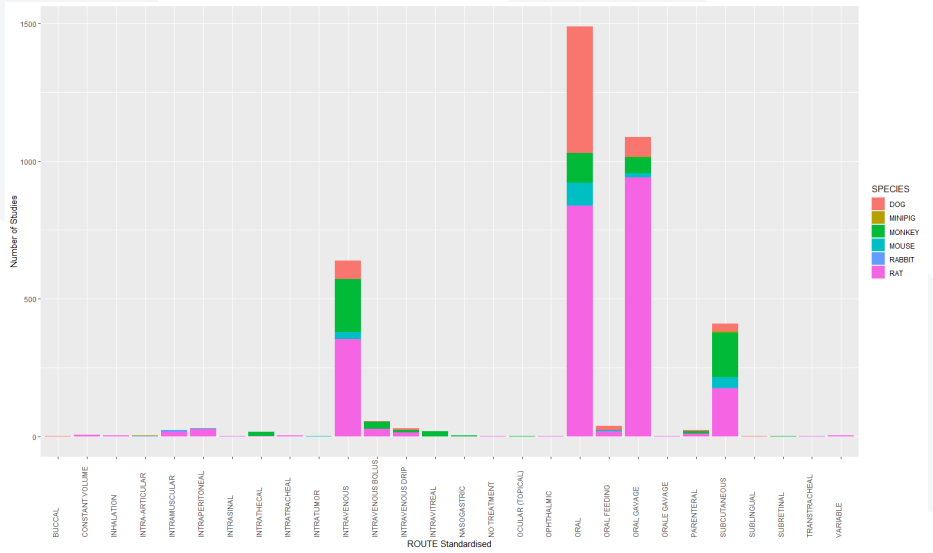

Figure 4 Curated routes of administration in VICT3R DB 2.0

A standardization of the routes published by the sponsors and enriched in the previous step of data collection was performed using standard routes published by FDA[6]. For each study, a TS record containing the result of the standardization in TSPARMCD="ROUTE" associated to TSPARM= "Standardized Route of Administration" is introduced to populate the curated data. For routes that could not be standardized, the term 'UNKNOWN' from the FDA list was applied. After curation 2261 studies with unknown routes, i.e. no or insufficient information published, have been identified. The curation efforts started from 60 different routes submitted and could reduce the number of routes to 28. As expected, the major routes are oral, subcutaneous and intravenous. Much smaller frequencies are observed for the remaining routes observed after standardization.

HOUSGRP and HOUSTYP

The type of housing can significantly influence the animals' health, behavior, and physiological responses. Controlled environments help minimize stress, which can confound study results. Housing types and grouping of animals in studies was therefore standardized to enhance consistency, reproducibility, and data quality. Standardization applied involve the following aspects: Housing Conditions, Cage Types, and Group Size. Overall, house grouping (HOUSGRP) was reported using more than 160 different terms. We have standardized and grouped these into three different categories as listed in table 5.

HOUSGRP	**Number of studies**
GROUPED	1106
GROUPED (F) / INDIVIDUALLY (M)	1
GROUPED (PRE-MATING), PAIR HOUSED (MATING), INDIVIDUALLY (AFTER MATING)	1
PAIR HOUSED	21
INDIVIDUALLY	411

Table 5 Standardization of HOUSGRP

Var1	Freq
AIR-CONDITIONED ROOM	229
CUSTOM DESIGNED PENS	35
INDIVIDUALLY VENTILATED CAGES	593
POLYCARBONATE CAGES	4546
SOLID BOTTOM CAGES	790
STAINLESS STEEL CAGES	2682
STANDARD CAGES	11527
STANDARD KENNELS	1804
STANDARD PENS	266
TYPE II MAKROLON CAGES	672
TYPE III MACOLON	20
TYPE III MAKROLON CAGES	780
TYPE IV MAKROLON CAGES	2413
TYPE IV, TYPE III	62
VENTILATED SOLID BOTTOM CAGES	846
WIRE CAGES	98

Table 6 Standardization of HOUSTYP

Using this approach could standardise 1540 studies into the categories listed in table 5. However, there were 4341 studies for which none of the standardized terms were appropriate (e.g. 'GRAMS', 'PERCT', 'GMS' …) or there was no information provided.

For HOUSTYP, more than 220 terms have been used to report the house type in SEND data provided. Table 6 lists the standardized type of housing conditions applied for the curation of HOUSTYP. Overall, the standardization has used 16 terms or categories with the number of animals as shown. The remaining, animals lack sufficient information for standardization.

BW Domain

The Body Weight (BW) domain contains information on body weight itself (BWORRES), measurement specification (BWTESTCD/BWTEST), measurement unit (BWORRESU), measurement time (BWDTC) and measurement interval (BWNOMLBL) for each individual subject (USUBJID) in a specific study (STUDYID). Other columns may provide additional information such as the reason why a measurement was not performed, the subject's fasting status, a calculation exclusion flag, or information on the (nominal) study day on which the measurement was performed.

In order to ensure an effective and comparable analysis of the measured values between different studies, we have standardized the variables BWORRES, BWORRESU and BWTEST uniformly for different species. This procedure is described below.

Unit standardization

The first step was to examine the units currently documented in the database. After evaluating which of the units were most appropriate for each species, considering their weight dimensions, we defined a unit for each species. For that, we created a database table showing the unique combinations between species and weight unit. In addition, this table contains a column in which the standardized unit of measurement is stored, and another column containing a numerical conversion factor necessary to calculate the standardized value. The results are shown in the table below (Table 7).

SPECIES	BWORRESU_orig	BWORRESU	FACTOR
DOG	g	kg	0.001
DOG	kg	kg	1
GUINEA PIG	G	g	1
MINIPIG	G	kg	0.001
MINIPIG	kg	kg	1
MONKEY	g	kg	0.001
MONKEY	kg	kg	1
MOUSE	g	g	1
OTHER (SEE PROTOCOL)	G	g	1
PIG	G	kg	0.001
PIG	kg	kg	1
RABBIT	g	g	1
RABBIT	kg	g	1000
RAT	g	g	1
RAT	GR	g	1
RAT	kg	g	1000

Table 7. Distinct units for each species and conversion into standardized values

This approach has enabled us to reduce the number of documented units of measurement, including upper and lower case, from 18 to a total of 3. In addition, a consistent procedure for calculating standardized values ensures that valid calculations of statistical parameters can be carried out.

BWTEST standardization

As mentioned earlier in this chapter, the BWTest column contains information specific to the time of weight measurement. The content of this variable is currently managed in the CDISC SEND Controlled Terminology (version 2024-09-27) under BWTEST (Body Weight Test Name) in three different ways: “Body Weight”, “Gravid Uterus Adjusted Body Weight” and “Terminal Body Weight”.

We first looked at the unique terms in this column (this time without reference to a species). There were mainly different spellings or abbreviations that led to a higher number of terms than expected. The entries originally documented in the database and the corresponding curated values are presented in the following table (Table 8):

BWTEST_orig	**BWTEST**
Body Weight	Body Weight
Terminal Body Weight	Terminal Body Weight
BODYWEIGHT	Body Weight
Term BW	Terminal Body Weight
TERMBW	Terminal Body Weight
Randomisation Bodyweight	Body Weight
Terminal Bodyweight	Terminal Body Weight
BW	Body Weight
FP Bodyweight	
Bodyweights	Body Weight
Bodyweight Ctrl/2	

Table 8. Original and standardized values for BWTEST

We left out the terms "Bodyweight Ctrl/2" and "FP Bodyweight" as we could not relate them in a meaningful way to the CDISC SEND specifications. This approach allowed us to reduce the number of BWTest terms from 10 to a total of 4.

Technical implementation unit standardization and BWTEST

To implement the standardization, we used the interfaces listed in the following graphic:

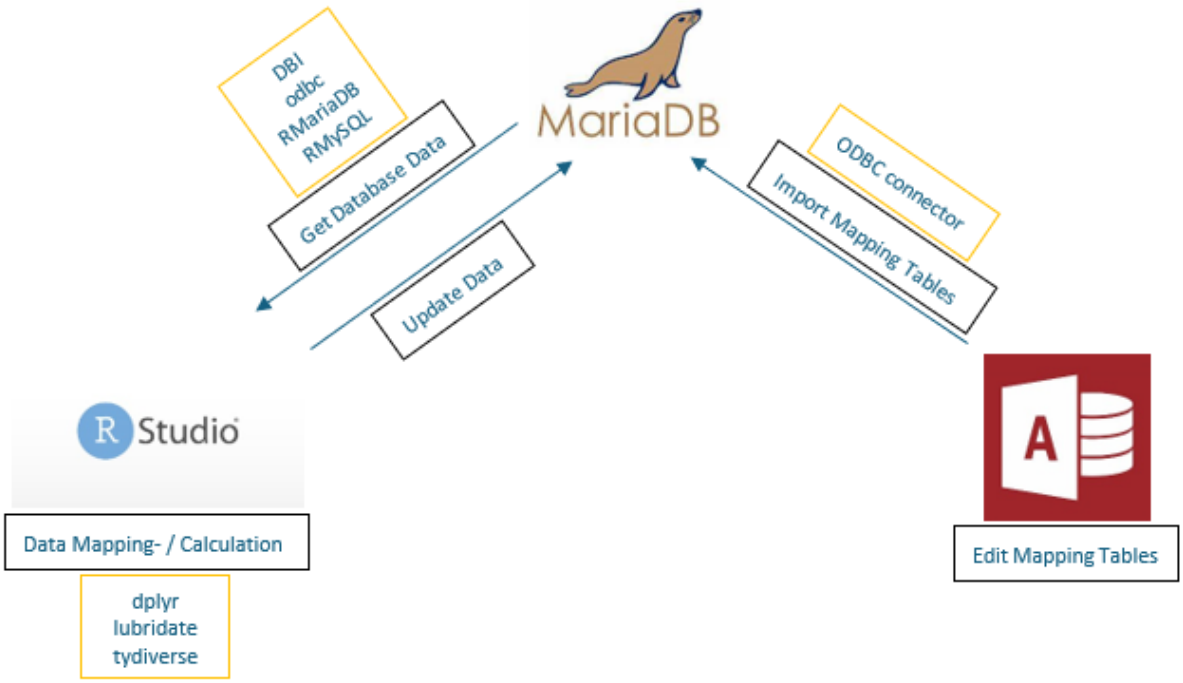

Figure 6 Standardization workflow applied in WP2

For a better understanding, the individual processing points are summarised below in an enumeration:

1. New columns were added to the MariaDB VICT3R database via RStudio (initially as placeholders) for the curated variables in the BW domain
2. The BWORRES column in the database was cloned via RStudio to keep the original values in a separate column.
3. The BW and DM domains were imported from the database into RStudio as data frames.
4. A mapping table with unique combinations (see table 7) was imported from MS Access into the database and from the database into RStudio.
5. The BW- and DM domain (SPECIES variable only) were 'merged' into a new data frame in RStudio using a left join (via USUBJID) ensuring that each record from BW was retained.
6. The mapping table was left joined (ensuring each record from BW was retained) to the merged data frame (by SPECIES and BWORRESU) and
7. Curated values were written/calculated and written in the curated columns of the merged data frame (see point 1).
8. A new data frame containing only the updated columns was created and uploaded to the database; and
9. The BW domain was overwritten (only updated columns) with the values from the uploaded data frame (see point 8).

The implementation of the standardization of the BWTEST variable was somewhat simpler here. Due to the small number of terms in the original data (see table 8), we did not use a mapping table and were able to implement the standardization via a programmed control structure in an R script.

LB Domain

The Laboratory Test Results (LB) domain captures laboratory data collected by the lab executing the study or data received from a central provider. The LB domain describes the clinical parameters, the specimen type, the examined organ or tissue, measurements and units as well as timing variables. In addition, the methods used to measure the specific effects are depicted.

The LB domain of the VICT3R database version 2.0 currently contains laboratory data for 4913 studies on different species. The majority of studies are tested in rats (N=3246), followed by dog studies (N=872), monkey (N=577) and mouse (N=191). Less frequently observed species are guinea pig, minipig or rabbit (Table 9).

Species	No. studies
Dog	872
Guinea pig	1
Minipig	9
Monkey	577
Mouse	191
Rabbit	17
Rat	3246

Table 9 Overview on the number of studies per species in the VICT3R database that provide laboratory data.

The LB data originate from different companies and from different systems and were therefore very heterogeneous e.g. with regard to the description of parameters, specimen and units. The observed heterogeneity of the data infers with an analysis of typical questions in the VICT3R project like the variability of LB parameter per species. In order to achieve better comparable data sets the shared data were therefore extensively harmonized and standardized. The original columns were renamed into …_orig and the standardized values were written to the respective columns so that the user can see which curation has been performed (see Table 10).

The curation of the LB domain consists of three steps:

Step 1: Curation of clinical parameter

The analysis of the clinical parameters showed that companies use different names or different spellings to describe the same parameter. Also typing errors were observed. To describe the parameter as accurately as possible four columns were considered:

- LBTESTCD – Lab Test or Examination Short Name (not curated)
- LBTEST – Lab Test or Examination Name (curated)

- LBCAT – Category of the Lab Test (curated)
- LBSPEC – Specimen Material Type (curated)

Standardization of terms considered as far as available the terms provided in the SEND terminology (CDISC SEND Controlled Terminology (version 2024-09-27)).

The following table provides a few examples (Table 10). It can be seen that in several cases, data are not entered into the appropriate columns, e.g. serum is specified under LBTEST, while belonging into LBSPEC. LBCAT provides data on methods, which are not expected at this place. Methods will be standardized in the next phase. LBTESTCD was not curated but often used to fill gaps, e.g. to specify the LBSPEC, as it often provides indications of measurements in urine (data not shown).

LBTESTCD	LBTEST_orig	LBCAT_orig	LBSPEC_orig	LBTEST	LBCAT	LBSPEC
	Alkaline Phosphatase		Serum	Alkaline Phosphatase	Enzyme activity	SERUM
	ALKALINE PHOSPHATASE (AP)	CLINICAL CHEMISTRY		Alkaline Phosphatase	Enzyme activity	no data
Alkaline Phosphatase	ALKALINE PHOSPHATASE	ADVIA1800 SERUM	SERUM	Alkaline Phosphatase	Enzyme activity	SERUM
Alkaline Phosphatase	ALKALINE PHOSPHATASE	BIOCHIMIE ROUTINE(PLASMA)	PLASMA	Alkaline Phosphatase	Enzyme activity	PLASMA
ALP	ALKALINE PHOSPHATASE	ADVIA1800 SERUM	PLASMA	Alkaline Phosphatase	Enzyme activity	PLASMA
ALP	ALKALINE PHOSPHATASE	AU3	SERUM	Alkaline Phosphatase	Enzyme activity	SERUM
ALP	ALKALINE PHOSPHATASE (ALK PHOS)	CLINICAL CHEMISTRY		Alkaline Phosphatase	Enzyme activity	no data
ALP	ALKALINE PHOSPHATASE (ALP)	CLINICAL CHEMISTRY		Alkaline Phosphatase	Enzyme activity	no data
AP	ALKALINE PHOSPHATASE IN SERUM	ADVIA 1650		Alkaline Phosphatase	Enzyme activity	SERUM

Table 10 Example of original and curated LB domain data.

The standardization of terms reduced the number of different terms significantly (Table 11), e.g. the number of unique terms in LBTEST decreased from 2307 to 759; while LBCAT decreased from 352 to 17 and LBSPEC from 31 to 29 unique terms. The effect is most visible when comparing the number of unique combinations of these three parameters. In the original database the combination of LBTESTCD/LBTEST/LBCAT and LBSPEC resulted in 7513 unique combinations. This number decreased to 1245 after data curation. The data curation therefore resulted in better comparable datasets, which can now be used for e.g. statistical analyses in the VICT3R project. The library of original and curated terms is provided in the database.

	LBTEST	LBCAT	LBSPEC	Combination LBTESTCD/LBTEST/LBCAT/LBSPEC
No. of original terms	2307	352	31	7513
No. of curated terms	759	17	29	1245

Table 11 Reduction of terms used for LBTEST, LBCAT and LBSPEC before and after data curation

Step 2: Harmonization of units

Units had different spellings for the same term, which needed to be standardized. The column LBORRESU, which provide the unit of the original result, was used for this purpose.

In addition to the standardization of terms also the units themselves were harmonized. Seldomly used units were converted to one commonly used unit per parameter. A factor is provided in the database (column "factor") to document the performed calculation (Table 12).

LBORRESU_orig	LBORRESU	Factor
/100 WBC	% WBC	1
per 100 WBC	% WBC	1
#/100 WBC	% WBC	1
/100WBC	% WBC	1
% WBC	% WBC	1
% of Leukocytes	% WBC	1
mg/dL	g/L	0.01
g/dL	g/L	10
mg/L	g/L	0.001
g/mL	g/L	1000
g/L	g/L	1

Table 12 Examples for unit standardization

The curation of terms and units reduced the number of unique terms by about 50%, from 261 to 127 terms.

Step 3: Calculation of SI units

A further standardization of the measured values to SI units (The International System of Units published by Bureau International des Poids et Mesures, BIPM)[7] was possible for 166 parameters. The column "Source SI" contains information on the origin of the SI unit calculation. For some parameters, we were able to standardize the values by combining similar units. These cases are marked as "summarised" in the source column. A factor is provided which was used to convert the value to the SI unit value.

For this curation step the following standardized columns were used:

- lbtest – from step 1 curated lab test
- lbcat – from step 1 curated category
- lbspec – from step 1 curated specimen
- lborres – from step 2 curated value
- lborresu – from step 2 curated unit

Table 13 provides a few examples to illustrate this curation step leading to standardized units using SI standards.

LBTEST	LBORRESU	LBORRESU curated_SI	factor SI	source SI
Cholesterol	g/L	mmol/L	2.5863	https://unitslab.com/node/58
Cholesterol	mmol/L	mmol/L	1	https://unitslab.com/node/58
Glucose	g/L	mmol/L	5.55	https://unitslab.com/node/1
Glucose	mmol/L	mmol/L	1	https://unitslab.com/node/1
Protein/Creatinine	g/mmol	g/mmol	1	summarised
Protein/Creatinine	g/mol	g/mmol	0.001	summarised
Protein/Creatinine	no units	RATIO	1	summarised
Protein/Creatinine	RATIO	RATIO	1	summarised

Table 13 Examples of units which are standardized to SI units.

Final columns measurement values LBSTRESN, LBSTRESU, LBSTRESC

In order to be consistent by all curation steps and to use the columns provided by SEND for the curated data, we have used standardized SEND columns for the measurement data. Also, here the original columns were renamed to ..._orig and the recalculated values were written to the corresponding data columns. The following column names were used for the measurement data:

- LBSTRESN – Standardized Result in Numeric Format
  - In a first step, the values were inserted from the SI calculation if available. If not, the data from the first step of standardization where no SI units could be calculated were taken in a second step. Finally, the values without standardized units were filled in if not yet filled.
- LBSTRESU – Unit of the Standardized Result
  - The same sequence as for LBSTRESN was followed here. First, the unit from the SI calculation was taken, then the unit without the SI calculation.
- LBSTRESC – Standardized Result in Character Format
  - All non-numerical values were written into this column.

Technical implementation

The technical implementation of data curation involves several steps. All terms from domains to be curated are collected and standardized as libraries in an Access database. This is currently related to BW, LB, MI and OM domains. The libraries are transferred to the MySQL database via ODBC drivers. In the next step, they are applied to the data tables and the curated data is written directly to the corresponded tables. The routine for curation

is written in PHP. Since the curation of the BW, OM and MI domains already runs in R, it is also planned to write an R script for this step. In this way, we want to ensure to stay in one technical environment.

OM Domain

The Organ Measurements (OM) domain contains information on organ measurements and relative organ weights (e.g. organ to body weight ratio). Information may be available for different organs for each subject (USUBJID) in a specific study (STUDYID). The variables OMTESTCD/OMTEST (unique identifier of the measured value according to CDISC SEND Controlled Terminology (version 2024-09-27), OMORRES (measured value or relative value), OMORRESU (unit of the measured value) and OMSPEC (defines the type of tissue, organ or fluid specimen) can be used to compare values between studies and to analyze differences between different organs. In addition, more specific information about the anatomical region of specimen (OMANTREG), specimen condition (OMSPCCND), laterality (OMLAT) and directionality (OMDIR) can be provided.

In order to ensure an effective and comparable analysis of the measurements across different studies, we have standardized the variables OMORRESU, OMTESTCD, OMTEST, OMSPEC, OMANTREG, OMLAT and OMSPCCND uniformly. This procedure is described below.

Unit standardization

The first step was to examine the different documented units (OMORRESU) in the current version of the database. By removing spelling mistakes, adapting some expressions to the CDISC SEND terminology, converting relative values to percentages and organ weights to grams, we were able to reduce the number of different units from 21 to a total of 5 (see table 14 below).

OMORRESU_orig	OMORRESU	FACTOR
%	%	1
g	g	1
% brain weight	%	1
% body weight	%	1
% body weight	%	1
% brain weight	%	1
% brain weigh	%	1
g/kg	%	10
G	g	1
RATIO	%	100
mg	g	0.001
cm	cm	1
kg	g	1000
GR	g	1
% body	%	1
% brain	%	1
Organ to brain weight ratio	%	100
to brain weight ratio	%	1
mm	mm	1

OMORRESU_orig	OMORRESU	FACTOR
g/100g	%	1
		1

Table 14 Unit standardization OM domain

The information on "body" or "brain" is not lost during standardization, as it is already available in the OMTESTCD or OMTEST columns.

Term standardization

To standardize the OM terms, we first looked at the unique combinations of the variables OMTESTCD, OMTEST, OMSPEC, OMANTREG, OMLAT, OMSPCCND and OMORRESU (curated version). In total there were 1432 unique combinations of these variables. The next step was to import this information into an MS Access database table, create a separate curation column for each variable, and map the OMSPEC variable to a mapping table we had already created for the MI domain. This curated column was then manually checked against the original column for plausibility and adjustments were made where necessary. Finally, the remaining missing information in OMSPEC and the other variables were standardized according to the CDISC SEND Controlled Terminology, version 2023-06-30. Details are given below.

OMTESTCD and OMTEST

In the current version of the database there were some terms for the variables OMTESTCD and OMTEST that were not listed in the CDISC SEND terminology but could be transferred to it. The following two tables provide an overview. Table 15 shows the OMTESTCD standardization table 16 gives an overview of the OMTEST standardization applied.

OMTESTCD_orig	OMTESTCD
DENSI	
EXTEN	
LENGTH	LENGTH
OB%	OWBW
OBR%	OWBR
OWBR	OWBR
OWBW	OWBW
OWHT	OWHT
RADIO	
THCKN	THCKN
WEIGHT	WEIGHT

Table 15 Standardization of OMTESTCD

The OB% and OBR% values could be transferred to the CDISC SEND terminology. There are currently no predefined terms for the DENSI, EXTEN and RADIO values, so we have not used a curated value in this case.

OMTEST_orig	OMTEST
BODY ORGAN WEIGHT	Weight
Density	
Extension of trabecular into diaphysis	
Length	Length
Organ to Body Percentage	Organ to Body Weight Ratio
Organ to Body Weight Ratio	Organ to Body Weight Ratio
Organ to Brain Percentage	Organ to Brain Weight Ratio
Organ to Brain Weight Ratio	Organ to Brain Weight Ratio
Organ to Heart Weight Ratio	Organ to Heart Weight Ratio
PATHBA - OW FRESH	Weight
PATHCA - OW FIXED	Weight
PATHCA - OW FRESH	Weight
PATHEH - OW FIXED	Weight
PATHEH - OW FRESH	Weight
Radiopacity	
Thickness	Thickness
Weight	Weight

Table 16 Standardization of OMTEST

The values Organ to Body Percentage, Organ to Brain Percentage and BODY ORGAN WEIGHT values could be transferred to the CDISC SEND terminology. For the characteristics beginning with PATH, no conclusions could be drawn from the context. These were standardized with the indication Weight, because the information Fresh and Fixed has been moved to the curated OMSPCCND column, the unit is grams, and we assume that the abbreviation OW stands for Organ Weight. There are currently no predefined terms for the Density and Radiopacity values, so we have not used a curated value in this case. As “Extension of trabecular into diaphysis” has been given in the OMTEST column, we suspect that it has been mistakenly placed there.

OMSPEC, OMANTREG, OMLAT and OMSPCCND

In the following, we describe the standardization of the remaining variables together, since a value in one column can be derived from values in the other column. Some examples are given in table 17 – 21 below to illustrate the process of standardization. The variables listed here are examples. The pattern applies to other organs.

OMSPEC_orig	OMSPEC	OMLAT
Adrenal Glands (Both) Weight	GLAND, ADRENAL	BILATERAL
Adrenal left	GLAND, ADRENAL	LEFT
Adrenal right	GLAND, ADRENAL	RIGHT
Adrenal (l)	GLAND, ADRENAL	LEFT
Adrenal(r)	GLAND, ADRENAL	RIGHT
EPIDIDYMIS (LEFT)	EPIDIDYMIS	LEFT
EPIDIDYMIS (RIGHT)	EPIDIDYMIS	RIGHT

Epididymides total	EPIDIDYMIS	BILATERAL

Table 17 Original OMSPEC to OMLAT

OMSPEC_orig	OMSPEC	OMANTREG
Lymph node, auricular	LYMPH NODE	AURICULAR
Skeletal muscle, extensor digitorum longus	MUSCLE, SKELETAL	EXTENSOR DIGITORUM LONGUS
Skeletal muscle, gastrocnemius	MUSCLE, SKELETAL	GASTROCNEMIUS
Skeletal muscle, plantaris	MUSCLE, SKELETAL	PLANTARIS
Skeletal muscle, soleus	MUSCLE, SKELETAL	SOLEUS
TESTIS, SEMINIFEROUS TUBULE	TESTIS	SEMINIFEROUS TUBULE
GANGLIA LOMBOSACRAL	GANGLION	LUMBOSACRAL
GANGLION TRIGEMINAL	GANGLION	TRIGEMINAL
TK - Brain, cortex	BRAIN	CORTEX
TK - Skin, dorsal	SKIN	DORSAL
TK - Skin, ventral	SKIN	VENTRAL
UTERUS/CERVIX/OVIDUCT	UTERUS/CERVIX	OVIDUCT
Uterus/vagina	UTERUS	VAGINA

Table 18 Original OMSPEC to OMANTREG

OMANTREG_orig	OMSPEC_orig	OMSPEC
mandib salivary	Gl	GLAND, SALIVARY, SUBMANDIBULAR
mandib saliv	Gl	GLAND, SALIVARY, SUBMANDIBULAR
thyroid	Gland	GLAND, THYROID
mand.	Sal. Gl	GLAND, SALIVARY, SUBMANDIBULAR
mandib	Saliv gl	GLAND, SALIVARY, SUBMANDIBULAR
mandibular	Salivary gland	GLAND, SALIVARY, SUBMANDIBULAR

Table 19 Original OMANTREG to OMSPEC

OMANTREG_orig	OMLAT
left	LEFT
right	RIGHT
total	BILATERAL

Table 20 Original OMANTREG to OMLAT

OMLAT_orig	OMLAT
1	
2	BILATERAL

Table 21 OMLAT standardization

Comments to results shown on tables 17-21:

Case 1: For OMLAT (table 21) the number 1 could not be curated because it did not make sense in the context of the organ systems (for all of them).

Case 2: In table 18 listing OMSPEC, organs were labelled as “weight”, “terminal body weight” or “body” (e.g. adrenal gland/terminal body weight). In these cases, the additional information could be omitted as it is already included in other columns.

Case 3: In table 18 listing OMSPEC, the term /brain was added to the organ (OMSPEC) prior to data donation to represent the ratio of the organ to the brain. However, in these cases OMTEST and OMTESTCD were set incorrectly (Organ to Body Weight Ratio / OWBW). Here we have corrected the information in the columns (e.g. OMTEST = Organ to Brain Weight Ratio | OMTESTCD = OWBR).

Case 4: In table 18 listing OMSPEC, the organs were presented with the designations "PD" or "TK" in the OMSPEC column. We couldn't see any context here and removed it in the curated column.

Case 5: In table 18 listing OMSPEC, the OMSPEC column included information on specific tests (ELISA, ImmunoTox, IPT). We have removed these in the curated column.

Case 6: The OMSPEC column contained only unit information (body weight, final body weight, final body weight, final fasting body weight) or other information that we could not assign (body, necropsy body, terminal body). In these cases, we have omitted a curated value.

Technical implementation standardization

The interfaces used to implement standardization have already been described in the BW domain (see Technical Implementation Unit Standardization and BWTEST). The principle is the same for the OM domain. We imported the mapping tables from our MS Access DB into the MariaDB VICT3R database for a) units and b) terms. In the next step, we added new columns to the database for the curated values in RStudio using a function, imported all the necessary tables from the database (OM, mapping tables for units and terms) into data frames, standardized the units in the first step using a left join and the terms in the second step also using a left join. Finally, we used a function in RStudio to import only the curated columns into a temporary table in the database, and another function in RStudio to update the original table with the values from the temporary table.

MI Domain

Data curation strategy

Microscopical Domain (MI) captures detailed information regarding microscopic observations. MI domain comprises data from histopathological findings, organized into distinct columns. Each of these columns represents the attributes to the study findings. Standardization is needed to ensure uniformity in the documentation of the microscopical findings. Database version 2.0. contained the original MI domain, while the curation of MI domain is still ongoing. Version 2.0.1. will be released upon completion of the curation.

Ten specific columns were curated according to CDISC SEND Controlled Terminology (version 2024-09-27) to compile a standardized catalogue of these combinations (Table 22). This catalogue will be utilized to map both the original and newly incoming dataset. By integrating all ten columns from the original dataset, a total of 91,848 unique combinations was identified. This included information from 3,844 studies (STUDYID), comprising a total of 56,138 animals (USUBJIDs). The curation of the MI domain was executed in two phases due to the complexity of the data and due to challenges of managing all possible column combinations simultaneously. The first phase of MI domain curation included variables related to specimen such as MISPEC, MIANTREG, MISPCCND, and MILAT. Subsequently, the second phase included variables reflecting the findings, MIORRES, MISTRESC, MIRESCAT, MICHRON, MIDISTR and MISEV.

Variable name	Variable description	Example
MISPEC	Specimen material type	LIVER, KIDNEY, STOMACH, etc.
MIANTREG	Region or subregion of specimen	CORTEX, MEDULLA, PERIVASCULAR etc.
MISPCCND	Specimen condition	AUTOLYSED, READABLE, NOT READABLE etc.
MILAT	Specimen laterality	BILATERAL, LOCALLY EXTENSIVE, UNILATERAL and combination of those.
MIORRES	Result	
MISTRESC	Standardized result	
MIRESCAT	Result category	BENIGN, MALIGNANT, METASTATIC, NON-NEOPLASTIC etc.
MICHRON	Chronicity	ACUTE, CHRONIC, SUBCHRONIC and combination of those.
MIDISTR	Distribution of the findings	DIFFUSE, FOCAL, MULTIFOCAL and combination of those.
MISEV	Severity of the findings	1 OF 3, 1 OF 4, 1 OF 5, 2 OF 3, 2 OF 4, 2 OF 5 etc.
ANNOTATION	0 = entries contained one specific result and splitting was not needed 1 = entries contained more than one results, and splitting was needed 5 = entries contained information about methods only 99 = open questions to be clarified with sponsors 100 = unremarkable findings and incomplete entries	

Table 22 Variables from the MI domain utilized for version 2.0.3. release.

Annotation column was added to indicate special cases during the standardization process. In some cases, MIORRES and/or MISTRESC contained multiple specific findings. For these instances, the corresponding rows were annotated with '1' to indicate that they needed to be handled further, such as splitting. If a row uniquely represented a single finding, it was considered ideal and annotated with '0'. The assigned annotation was based on the information provided in MIORRES and MISTRESC. Results such as 'UNREMARKABLE', 'NORMAL', 'NO VISIBLE LESION', 'NO CORRELATION', and 'ARTIFACT' received an annotation of '100'. This also included notes like 'NOT EXAMINED', 'NOT APPLICABLE', 'PART MISSING', and 'ABSENT/PART NOT PRESENT'. Open questions or entries requiring clarification from sponsors were indicated with '99'.

The curation of the MI domain was executed in two phases due to the complexity of the data and due to challenges of managing all possible column combinations simultaneously. The first phase of MI domain curation included variables related to specimen such as MISPEC, MIANTREG, MISPCCND, and MILAT. Subsequently, the second phase included variables reflecting the findings: MIORRES, MISTRESC, MIRESCAT, MICHRON, MIDISTR and MISEV.

Curation of MISPEC: MISPEC in their original terms was entered heterogeneously in different spellings. These were harmonized according to SPEC (Specimen) NCI Code: C77529 from CDISC.

Curation of MIANTREG: Information available in MIANTREG column were extracted and standardized. In instances where the information in MIANTREG column are not available, often they were misplaced in other columns. MISPEC and MISPCCND columns were checked for additional information in such cases.

Curation of MISPCCND: information about the specimen condition were extracted from original column, where this information was available. Analogue to the curation of MIANTREG, where the information was misplaced, MIANTREG and MISPEC were checked for misplaced MISPCCND entries.

Curation of MILAT: laterality of the examined specimen was provided in MILAT original column, and where the information was available, this was standardized according to LAT (Laterality), NCI Code: C99073. Likewise, other columns MISPEC, MIANTREG and MISPCCND were considered in case for misplaced entries.

Curation of MIORRES: MIORRES provided the information on original findings. Often, most of the information related to findings were pooled as free text in MIORRES. Therefore, it was necessary to investigate MIORRES in detail to extract and reassign the entries in the correct corresponding columns. MIORRES also contained combined effects. In such cases, annotation was assigned for future splitting strategy.

Curation of MISTRESC: MISTRESC provided the standardized results from each laboratory. Where there were entries available, these were extracted and standardized for different spellings, capitalization and typographical errors. In many cases, the standardization did not follow SEND terminologies. In such cases, the standardization of this column was conducted according to NCI Code: C120531 for non-neoplastic findings and C88025 for neoplastic findings.

Curation of MIRESCAT: this was conducted according to the provided terminologies in CDISC with NCI Code: C90017. When information was provided, this was taken curated according to spellings, typographical errors and capitalization. However, in cases where information was not available, other columns such as MIORRES and MISTRESC were considered as they contained information about the category of the results (non-neoplastic and neoplastic, as well as benign, malignant and metastatic).

Curation of MICHRON: this was conducted according to the provided terminologies in CDISC with NCI Code: C120529. When information was provided, this was taken curated according to spellings, typographical errors and capitalization. However, in cases where information was not available, other columns were searched for possible misplaced entries.

Curation of MIDISTR: this was conducted according to the provided terminologies in CDISC with NCI Code: C120530. When information was provided, this was taken curated according to spellings, typographical errors and capitalization. However, in cases where information was not available, other columns were searched for possible misplaced entries.

Curation of MISEV: this was conducted according to the provided terminologies in CDISC with NCI Code: C90000. When information was provided, this was taken curated according to spellings, typographical errors and

capitalization. However, in many cases, the grading of the severity did not follow CDISC proposed terms, i.e. MILD, MODERATE, etc. In such cases, this information was pertained.

The curation not only curated free text into SEND-controlled terminology but also ensured that information was assigned in the correct columns. Variations in spelling were harmonized, and typographical errors were corrected (Table 23). Often, results are grouped together in a single column, such as in MIORRES, without accurate assignment to the corresponding columns. In such cases, information was assigned to the right columns representing this variable (Table 24). When necessary, where specific findings might correlate with the species, the species information from DM domain needed to be queried to assign the accurate terminology for such findings. During the curation process, SEND terminology was utilized as the standard. However, not all findings are associated with a registered SEND controlled terminology. In such cases, findings were assigned to closest possible terms.

MISPEC	MIANTREG	MISPCCND	MILAT	MISPEC_ curated	MIANTREG _ curated	MISPCCND _ curated	MILAT_ curated	ANNOTATION
ADRENAL CORTEX			BILATERAL	GLAND, ADRENAL	CORTEX		BILATERAL	0
Adrenal Gland		capsule		GLAND, ADRENAL	CAPSULAR			0
Adrenal Gland		cortex		GLAND, ADRENAL	CORTEX			0
Adrenal gland	capsule		UNILATERAL	GLAND, ADRENAL	CAPSULAR		UNILATERAL	0
ADRENAL GLAND	CORTEX			GLAND, ADRENAL	CORTEX			0
ADRENAL GLAND(S)		Autolytic, readable	UNILATERAL	GLAND, ADRENAL		AUTOLYSED/ READABLE	UNILATERAL	0
GLAND, ADRENAL	NEDULLA		BILATERAL	GLAND, ADRENAL	MEDULLA		BILATERAL	0

Table 23 Examples for harmonization of free text MISPEC, MIANTREG, MISPCCND and MILAT into SEND controlled terminology

MIORRES	MISTRESC	MIRESCAT	MICHRON	MIDISTR	MISEV	MIORRES_ curated	MISTRESC_ curated	MIRESCAT_ curated	MICHRON_ curated	MIDISTR_ curated	MISEV_ curated	ANNOTATION
Hemorrhage; sinusoid; acute; multifocal	Hemorrhage; sinusoid; acute; multifocal					HEMORRHAGE, SINUSOIDAL	HEMORRHAGE	NON-NEOPLASTIC	ACUTE	MULTIFOCAL		0
Atrophy with inflammation, acinar cell, focal, mild	ATROPHY					ATROPHY, ACINAR CELL/ INFLAMMATION, ACINAR CELL	ATROPHY/ INFLAMMATION	NON-NEOPLASTIC		FOCAL	MILD	1

ADENOCARCINOMA PARS DISTALIS	Adenocarcinoma	Malignant				ADENOCARCINOMA, PARS DISTALIS, MALIGNANT	ADENOCARCINOMA, PARS DISTALIS, MALIGNANT	MALIGNANT				0
Adenoma cortical tubular cell focal	Tubular Adenoma	Benign		Focal	2 OF 5	ADENOMA, TUBULAR CELL, BENIGN	ADENOMA, TUBULAR CELL, BENIGN	BENIGN		FOCAL	2 OF 5	0

Table 24 Standardization of terms for MIORRES and MISTRSC including modifiers

Current status of data curation

The first phase of standardization included MISPEC, MIANTREG, MISPCCND, and MILAT. The integration of these columns generated a total of 5,024 combinations. To date, the curated version of these columns resulted in a total of 2,300 combinations. The second phase of the curation strategy included MIORRES, MISTRESC, MICHRON, MIDISTR, MISEV and MIRESCAT. The integration of these columns resulted in a total of 64,966 combinations. Currently, the curation is ongoing, and the status of these columns has resulted in a total of 18,767 combinations (Figure 7).

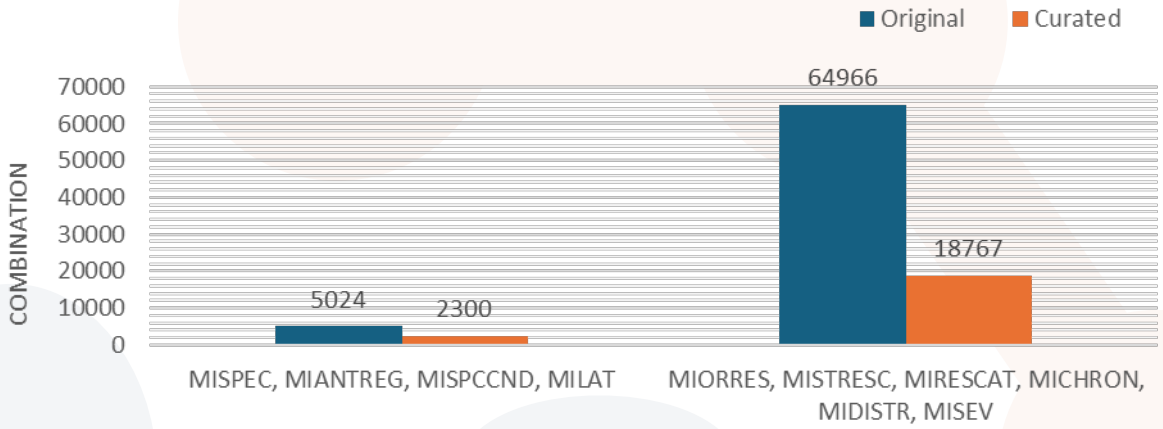

Figure 7 Status of MI domain standardization. Blue highlighted bar indicated the combination of the mentioned columns of the original data, orange highlighted bar of the currently curated data combinations.

Overall, the curated data will provide harmonized and SEND-controlled entries for the upcoming analysis. These curated columns will facilitate the integration of those, thereby enhancing the comprehensiveness and depth of the analysis.

4. Discussion

The standardization of data in the VICT3R project is crucial for ensuring consistency and accuracy across various studies. Key steps in the workflow included data collection from the VICT3R data lake, where contributions from multiple sponsors were compiled. The integration process compared individual SEND files against the SEND standard, removing extraneous columns and ensuring compliance with standardized formats.

Data cleaning involved identifying and correcting errors, inconsistencies, and duplicates within the datasets. Validation ensured that the standardized data met predefined quality standards, making it ready for project work. The project also focused on standardizing metadata, particularly in the trial summary and demographics domains, to improve data sharing and integration across systems.

The standardization of housing types and grouping variables was particularly emphasized due to their significant impact on animal health, behavior, and study outcomes. By categorizing housing conditions and grouping animals consistently, the project aimed to enhance data quality and reproducibility.

The laboratory domain standardization within the VICT3R project is pivotal for harmonizing the diverse laboratory data collected from various studies. This domain captures critical laboratory parameters, specimen types, examined organs or tissues, measurements, units, and timing variables. Given the heterogeneity in data reporting practices among different organizations, the standardization process aimed to create a consistent framework for analysis.

Key steps in this standardization included the curation of clinical parameters. The analysis revealed that different companies used varying names and spellings for the same laboratory tests, leading to inconsistencies. To address this, the project utilized controlled terminology from the CDISC SEND Controlled Terminology to standardize the naming conventions across parameters such as LBTEST (Lab Test or Examination Name), LBCAT (Category of the Lab Test), and LBSPEC (Specimen Material Type). This effort significantly reduced the number of unique terms, enhancing the comparability of laboratory datasets and facilitating robust statistical analyses.

Additionally, the standardization process involved harmonizing units of measurement. Different studies reported results using various units, which could complicate comparisons. By converting seldom-used units into a commonly accepted unit per parameter, the project aimed to streamline data interpretation and ensure consistency across studies. The inclusion of a conversion factor in the database further documented these transformations, providing transparency and clarity for users.

Weight measurements are critical in toxicology studies, influencing the assessment of drug efficacy and safety. The standardization of weight measurements in the VICT3R project focused on ensuring that body weight data (BW) were reported uniformly across different species and studies. This process began with a thorough examination of the units documented in the database, where various units were used inconsistently.

To standardize weight measurements, the project defined a common unit for each species, creating a mapping table that documented the unique combinations of species and weight units. This table facilitated the conversion of recorded weights into standardized values, reducing the complexity of data handling from 18 different documented units to just three. This simplification is crucial for performing valid statistical analyses and comparisons across studies.

Furthermore, the standardization process extended to the BWTEST variable, which captures the timing of weight measurements. By analyzing the original terms and applying a curated set of standardized terms, the project significantly reduced the number of unique terms used, ensuring clarity and consistency in reporting.

Overall, the standardization of both the laboratory domain and weight measurements plays a vital role in enhancing the reliability and comparability of data within the VICT3R project. These efforts not only support

regulatory compliance but also contribute to the overarching goal of reducing animal use in research by facilitating the development of virtual control groups and improving the quality of scientific data.

5. Conclusion

The document highlights the importance of rigorous data management practices in the context of toxicology research and the development of virtual control groups, ultimately contributing to the reduction of animal use in research.

6. References

[1] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686

[2] Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4

[3] dplyr version 1.1.4: Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://github.com/tidyverse/dplyr

[4] Wickham H, Hester J, Bryan J (2024). *readr: Read Rectangular Text Data*. R package version 2.1.5, https://github.com/tidyverse/readr, https://readr.tidyverse.org.

[5] Grolemund G., Wickham H. (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25.

[6] Route of Administration, CDER Data Element Number. C-DRG-00301, Data Element Name. Route of Administration. Data Element OID: 2.16.840.1.113883.3.26.1.1.1 Data Element NCI Concept ID: C38114 https://www.fda.gov/drugs/data-standards-manual-monographs/route-administration

[7] https://www.bipm.org/en/