101172693 – VICT3R

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

WP3 – Extension to studies other than systemic toxicity

D3.2 Academic landscape of animal-based studies

Lead contributor:	Olga Valverde (1-UPF)
Other contributors:	Javier Orihuel (1-UPF), Juan Manuel Ramirez Anguita (1-UPF), Javier Corvi (7-MBIS), Laura I. Furlong (7-MBIS), Janet Pinero (7-MBIS), Anna Simonova (7-MBIS), Julia Matyjasiak (13-VUB), Jian Jiang (13-VUB), Marco Dilger (18-BASF SE)
Due date:	31/08/2025
Delivery date:	28/11/2025
Dissemination level:	Public

Contents

Document History

Version	Date	Description
V0.1	31/10/2025	First Version for Formal Review
V0.2	21/11/2025	Version for Consortium Review
V1.0	28/11/2025	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

With the initial objective of providing overview of the types, frequency and diversity of animal-based study designs in academia, this report combines insights from three key approaches: **(1)** The **examination of animal-based research** using statistics database on the use of animals for scientific purposes (**ALURES database**), focusing on the types and frequency of studies relevant to academic landscape of animal-based studies, specifically basic and translational research. **(2)** An in-depth analysis of **Non-Technical Summaries (NTS)** - which describe how and why animals are used in specific research proposals - was conducted using **Natural Language Processing (NLP)**, focusing on studies in the most animal-intensive area of basic research. This analysis highlights the frequency of elements critical for **Virtual Control Groups (VCGs)** and **Historical Databases (HDBs)**, including Experimental Frameworks, Manipulations/Procedures, Tests/Apparatus and Data Acquisition Methods. **(3)** A dedicated **academic survey** conducted to target researchers and personnel with experience in *in vivo* **basic and translational research** and assess current practices, methodological variability, historical data availability, and potential for data reuse. Additionally, **(4)** this report also presents an **overview of the challenges associated with animal data reuse in academia**, discussing the utility of the **SEND standard**, providing a summary of other data and metadata sharing standards, and illustrating a case study of successful animal data reuse in an academic setting. This case study is particularly relevant for conceptualizing potential pathways for the implementation of VCGs in academia.

The statistics on animal-based research (ALURES statistical database) relevant to the academic landscape - covering basic and translational research - showed that **rodents** (especially mice) in nervous system studies accounted for the most significant animal use in basic research. NLP analysis of Nervous System NTS records indicated that **Neuropsychiatric and Neurodegenerative Disease Models, along with Pharmacological, Toxicity, and Pain-Related Studies**, constituted the predominant experimental paradigms. **Pharmacological and Pharmacogenetic Interventions** represented the main manipulations, while behavioral assessments and apparatus for **Anxiety, Stress, and Pain** were the most frequent tests. Data acquisition was dominated by **Histological Analyses and *in vivo* Imaging methods**. The survey on basic and translational research highlighted a preference for **pain, neurodegenerative, and addiction paradigms as the most positively valorized for VCG** approaches. Remarkably, VCGs are not perceived as an immediate necessity in academic research, but there is a strong willingness to share data and interest in building platforms like VICT3R to support data reuse and improve research efficiency.

In summary, these findings indicate that **pain, as well as Neuropsychiatric and Neurodegenerative Disease Models**, are the most animal-intensive and positively valorized areas for VCG approaches. While **challenges in standardization and data quality remain**, there is a viable path for adopting Virtual Control Groups, driven by inherent interest in enhancing reproducibility, efficiency, ethical animal use, and FAIR practices, which can be implemented through domain-specific, ontology-driven, and interoperable frameworks.

1. Introduction

Animal Virtual Control Groups (VCGs) are a concept derived from the more advanced area of virtual patients in clinical trials. It was proposed by Steger-Hartmann et al. in 2020 and was later established as a proof of principle based on a regulatory four-week study using Wistar rats. The proof of principle was provided through the work of Gurjanov et al. (2024), which drew subjects from historical animal datasets to reanalyze four-week rat oral toxicity study findings. The VICT3R consortium provides, for the first time, an opportunity to **implement VCGs beyond highly standardized regulatory studies**.

In contrast to regulatory studies, which are generally more standardized and structured, **academic research encompasses a broader diversity of experimental designs and objectives**. Consequently, adapting VCG and related approaches to the academic context requires first identifying specific niches - experimental designs, manipulations, tests and data types - that are performed with sufficient consistency and frequency to support focused development and practical application of VCGs.

Building upon our prior work (see VICT3R Deliverable 3.1.) analyzing regulatory studies and animal use reported in ALURES (the EU Animal Use Reporting System), **this report integrates complementary approaches to map current academic practices**. Academic research is predominantly basic - aimed at understanding fundamental biological processes - and translational focused on bridging these findings toward clinical applications which, according to ALURES, together account for the majority of animal use in the EU. Accordingly, we began by analyzing ALURES data to explore the use of animals across species and study types within basic and translational research domains.

This was followed by an in-depth examination of Non-Technical Summaries (NTS) using Natural Language Processing (NLP) techniques. This approach enabled the extraction of structured metadata, offering deeper insight into specific experimental frameworks, procedures, and data types most relevant to VCG applications. Although NLP-guided analysis of all NTS is technically feasible, this initial effort focused primarily on nervous system studies, due to their prevalence in basic research and their potential relevance for VCG development.

In parallel, a targeted survey was developed for personnel involved directly or indirectly in in vivo basic and translational research, aiming to identify additional potential niches beyond those highlighted in the ALURES data and the initial focus on nervous system studies. The survey was specifically designed to gather input directly from researchers working with, or involved indirectly in, in vivo animal research within basic and translational research settings. It collected detailed information on model compatibility, data availability, existing repositories, and researchers' perceptions of the usefulness of historical datasets and VCG approaches for the models and tests relevant to their fields of research. The survey was first implemented among academic partners and subsequently extended to an international audience, distributed to scientific and 3Rs organizations, with the aim of mapping animal models currently used in academic and translational research that may align with the VCG concept.

In addition to interpreting and analyzing the obtained results, **we aimed to enrich the report with a snapshot of the current state of animal data sharing, interoperability, and reuse in academia**, presented in the discussion section.

Together, **these complementary approaches provide a comprehensive understanding of animal model use and data practices in academia**, laying a solid foundation to advance VCG strategies beyond regulatory and industry settings, with the goal of promoting animal welfare, data reuse, and experimental standardization.

This report also provides **recommendations on the feasibility of extending the VCG concept and advance related tools to the academic environment**. For instance, ongoing efforts to apply and accommodate SEND (or SEND-like) formats to cover new data types and academic experimental designs that may help ensure the construction of Historical Databases and the consistent collection of high-quality metadata.

In summary, we report:

1. The exploration of the **ALURES statistical database** to map animal use across species, numbers of animals, and research purposes, with emphasis on academic-relevant categories such as basic and translational research.
2. The analysis of **ALURES NTS files using NLP tools** to identify the frequency of experimental paradigms, tests, and measurements in animal-intensive research areas, particularly neuroscience.
3. The implementation of an **academic expert survey on animal models**, aimed at mapping in vivo models used in basic and translational research and assessing their potential alignment with VCG methodologies, including data availability, existing databases, and researchers' perceptions of historical data reuse.
4. An in-depth **discussion on the feasibility and current work for extending the VICT3R VCG approach to academic studies**, including considerations for SEND-like reporting and the generation of historical datasets in basic and translational research areas.

The results and insights presented here will serve as a foundation for several key developments in the implementation of VCG in studies other than repeat dose toxicity.

2. Mapping In Vivo Basic and Translational: Evidence from the ALURES Database

The **ALURES database** contains detailed information on the use of animals in scientific settings, ensuring transparency and harmonized reporting **across the EU**. Since the adoption of Directive 2010/63/EU, EU Member States (MS) are legally required to submit annual statistics on the use of animals for scientific purposes. This is done through the publication of annual statistics per MS, and NTS of each *in vivo* project. These data improve transparency regarding animal use in the EU and contribute to a better understanding of why and how animal research is conducted. In this context, ALURES was established to ensure compliance with the Directive 2010/63/EU and to centralize the reporting and accessibility of such information.

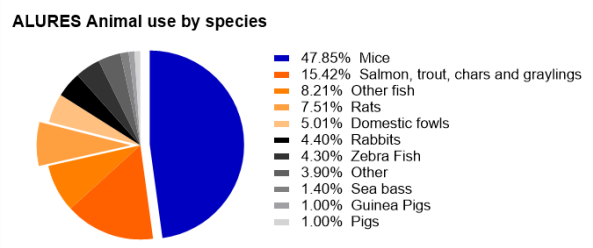


Figure 1. Information on the species of animals used in scientific projects registered in the ALURES database in 2022.

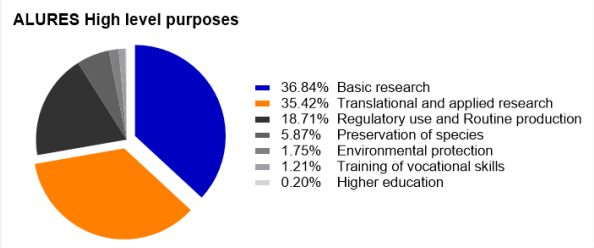


Figure 2. Types of "High-Level Purposes," registered in the ALURES database in 2022.

There are three main sections in the ALURES statistical database: **Section 1 covers the number, species, and origin of animals** used in research, testing, routine production, and education and training. Figure 1 shows the species of animals used in scientific projects registered in ALURES during 2022. **Section 2 presents the total number of uses** - both initial and subsequent - along with the purpose of use (e.g., specific research areas, types of testing), the severity of the procedures (mild, moderate, or severe), and the applicable legislative requirements. Figure 2 shows the frequency of the distinct high-level purpose categories associated with the ALURES entries during 2022. **Section 3 focuses on the use of genetically altered animals** in scientific research.

As shown in Figure 1, **mice represent by far the most used species in scientific projects across Europe**. Figure 2 illustrates that up to 71% of all animal use falls within **Basic and Translational research**, the majority of which is conducted in academic settings. Together, these data highlight both the central role of mice as experimental models and the predominance of academia-driven basic and translational studies in shaping animal use across the EU. This distribution is broadly transposable to non-EU members, where mice also dominate experimental use and most animals are employed in basic and translational research, although exact percentages may vary depending on national regulations, reporting systems, and the balance between academic and industrial research.

2.1. Animal Use in Basic and Translational Research

Based on the total uses of basic and translational research in 2022, **Neuroscience related studies represent the most significant use of selected animals** alongside oncology and immune system projects (Figure 3).

Though oncology creates the largest node altogether in Figure 3, the largest stream from Basic research category (middle section) to research field (right section), i.e. Nervous System, feeds into the Neuroscience node.

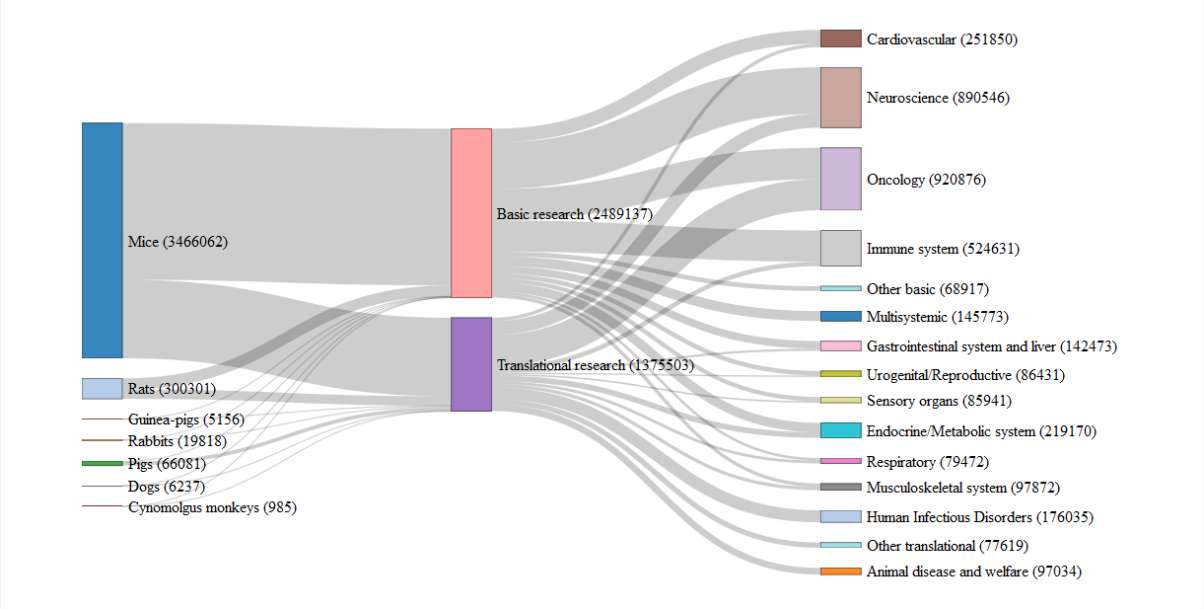


Figure 3. Number of animals used for the selected species (mice, rats, guinea-pigs, rabbits, pigs, dogs and cynomolgus monkeys) in the last reported year in ALURES, 2022.

For the selected animals, consistently from the first reported year of 2015 to the last reported year 2022, this category has the highest percentage of animal use per all basic research categories. This **supports the work in VICT3R regarding establishing VCGs in experiments of neuroscience research**. However, the high use in other research areas will require scientific collaborations and expertise of *in vivo* scientists in those fields to establish experiments where both animal use intensity and feasibility are balanced.

3. Natural Language Processing guided analysis of Nervous System Non-Technical Summaries

The ALURES NTS database provides publicly accessible information on scientific projects that involve the use of living animals. The data is derived from project summaries submitted by EU Member States for all projects authorized after January 1, 2021. Each NTS record contains valuable details on study design and experimental procedures, but this information is presented as free text across multiple sections. This free-text format, combined with the availability of records in 22 different languages, creates substantial challenges for data extraction and analysis. In some cases, even a single NTS record contains sections written in different languages. For example, the objectives section may appear in English, while other sections are written in another language. Figure 4 illustrates this complexity by showing an excerpt from an NTS record that includes references to study types and experimental methods.

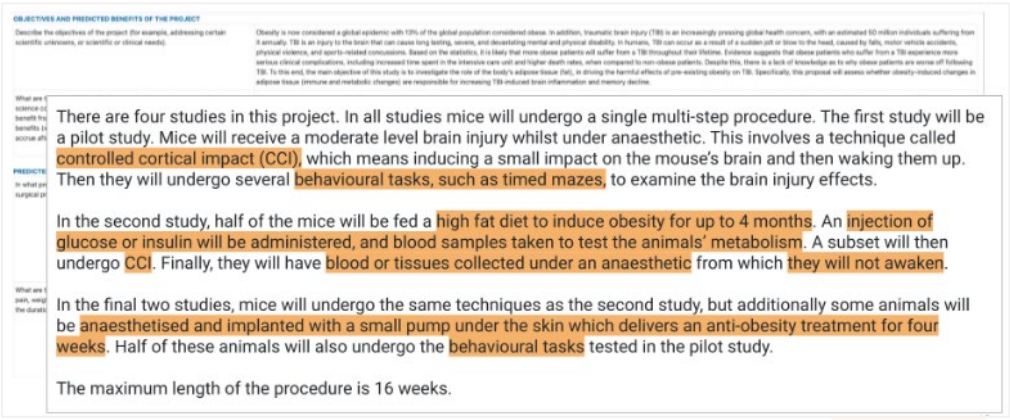


Figure 4. Non-Technical Summary phrases containing information about study type and methods.

In animal research, a wide range of experimental procedures and methodologies can be applied, varying by specific area, type of model, and study objectives, among other factors. Accurately identifying these requires substantial domain expertise. The challenge is amplified by the absence of a controlled vocabulary or comprehensive ontology for this domain. Without standardized terminology or a well-defined framework, processing the information accurately and efficiently becomes highly complex.

3.1. Natural Language Processing Approach Overview

To overcome these challenges, a potential solution involves leveraging **Large Language Models (LLMs) for data extraction from NTS files**. This approach could enable more effective and precise extraction and analysis of the data, regardless of domain, language or format. Given the information on animal use previously gathered from the ALURES statistical database we selected a dataset from the ALURES NTS web portal using the following filters:

- **Species**: Mice (*Mus musculus*)
- **EU Submission**: Yes
- **Purpose of the project**: Basic Research – Nervous System

This resulted in **6,155 NTS records**, collected on February 28, 2025. German was the most frequently used language, accounting for approximately 31.6% of all reports, followed by French at around 26.1%. Italian and Spanish were less common, each representing roughly 10% of the records. The remaining 22.3% of reports were distributed across a variety of other languages, highlighting the linguistic diversity present in the NTS dataset.

To analyze the methods reported in these documents, we implemented a four-stage pipeline:

i. **Stage 1 – Extraction**: LLMs are applied to extract descriptions of experimental procedures from NTS records. A multi-agent debate framework (comprising affirmative, negative, and judge agents; Liang et al., 2024) enhances accuracy across multiple languages and standardizes all outputs into English.
ii. **Stage 2 – Categorization**: The extracted procedures are classified into four categories by LLMs: Experimental Frameworks, Manipulations/Procedures, Tests/Apparatus and Data Acquisition Methods.
iii. **Stage 3 – Clusterization**: Conceptually similar procedures are merged to reduce redundancy arising from lexical variation (e.g., “euthanasia,” “sacrifice,” and “anesthetic overdose”). This step uses hierarchical agglomerative clustering based on semantic embeddings.
iv. **Stage 4 – Expert Curation**: The resulting taxonomy is presented in an interactive browser-based tree. Domain experts can refine it by adding, removing, renaming, or reorganizing nodes, ensuring both accuracy and conceptual clarity.

The outcome of this process is an **expert-curated taxonomy of experimental methods**, presented through the **ALURES NTS Experimental Methods Browser** developed **by the VICT3R consortium**. This browser-based interactive tree allows users to explore the methods systematically, navigating across categories and subcategories to examine specific procedures. By providing a structured and refined view of the data, the tool supports efficient comparison, analysis, and decision-making in animal research practices https://vict3r.medbioinformatics.com/alures/.

It is important to note that the **ALURES NTS Experimental Methods Browser** exists in two versions:

i. **An expert curation version**, which allows users to edit, add, remove, and reorganize nodes that were used during the validation process.
ii. **A public visualization version**, available at https://vict3r.medbioinformatics.com/alures/, which presents the final, expert-curated taxonomy for exploration and visualization only, without editing capabilities.

For further details on how the taxonomy was derived, see the **Methods Section**.

3.2. Results

The agent-based pipeline was applied to **6,155 NTS records**, yielding a total of **13,765 extracted terms** representing raw mentions of study designs and methods prior to categorization, clustering, and expert curation. During the clustering phase, semantically related terms were automatically grouped into clusters and subclusters to consolidate similar procedures and improve conceptual coherence.

Each extracted term was linked to the total number of NTS reports in which it appeared, allowing for frequency-based analysis of method usage. The **categorization stage used expert-defined labels**, dividing the first level of the taxonomy into four categories:

i. **Experimental Paradigms/Frameworks** (1,135 reports): This category encompasses the models, paradigms, and protocols that define the core experimental structure of a study. These frameworks establish the foundational context in which hypotheses are tested and data are generated. Examples include disease or injury models such as tumor models and experimental autoimmune encephalomyelitis (EAE).

ii. **Manipulations/Procedures** (5,519 reports): This category describes the specific interventions, treatments, or independent variables introduced to influence experimental outcomes. These may include behavioral, pharmacological, genetic, dietary, or environmental manipulations designed to cause or elicit particular responses. Examples include *high-fat diet*, *stress induction*, and *dietary manipulation*.

iii. **Tests/Apparatus** (1,139 reports): This category refers to the experimental assessments, instruments, and tools used to measure responses or outcomes within a study. These include behavioral, physiological, or biochemical tests designed to evaluate the effects of experimental manipulations. Examples include the *object recognition test* and *fear conditioning*.

iv. **Data Acquisition Methods** (4,830 reports: This category covers the modalities, instruments, and technologies used to collect experimental data. These methods capture physiological, behavioral, or environmental signals that serve as the basis for analysis. Examples include *telemetry*, *EEG*, and *actimetry*.

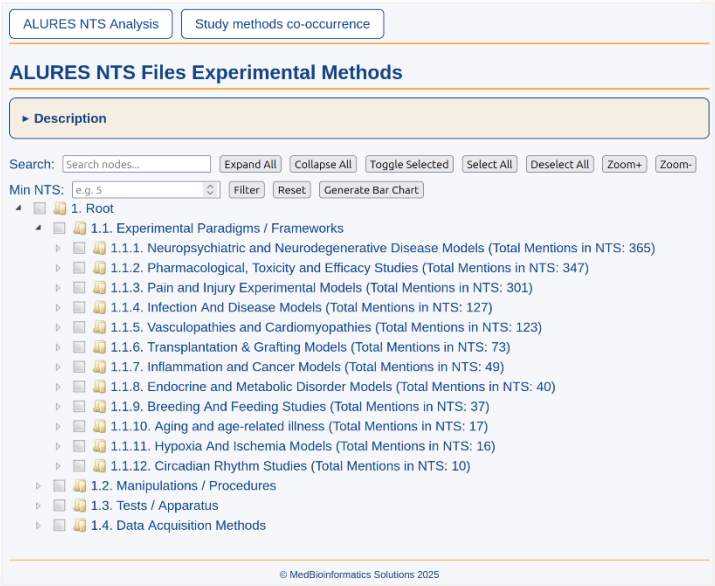


Figure 5. First level of the ALURES NTS Files Experimental Methods.

The final **expert-curated taxonomy was composed of 6,096 NTS reports and 10,077 methods**, and it is displayed in the ALURES NTS Experimental Methods Browser, available at https://vict3r.medbioinformatics.com/alures/. This browser displays the taxonomy as a hierarchical,

interactive tree with systematically numbered levels (e.g., *1.2.3.1 Method X*), allowing users to navigate between groups, clusters, and specific methods.

Figure 5 illustrates the first level of the tree, highlighting the four main experimental clusters and the total number of NTS reports associated with each. In addition, the figure provides an example of the Experimental Paradigms/Frameworks category expanded to display the subclusters it contains. This interface enables efficient exploration of curated methods, making the data directly actionable for comparative analyses, knowledge discovery, and informed decision-making in animal research. The interface is fully interactive and provides several features to enhance usability. Users can expand or collapse branches, search for terms across the hierarchy, and filter nodes based on minimum total mentions of methods in NTS. Selections can be made at any hierarchical level, with the option to enable or disable cascading selection across parent and child nodes, making it easy to isolate subsets of interest. Additional functionality includes bulk actions (select all/deselect all) and zoom controls for accessibility.

Among the four categories, the **Experimental Paradigms** stand out as particularly informative, revealing the most common foundational frameworks guiding research. The leading subcategories include Neuropsychiatric and Neurodegenerative Disease Models (365), Pharmacological, Toxicity and Efficacy Studies (347), Pain and Injury Experimental Models (301), and Infection and Disease Models (127).

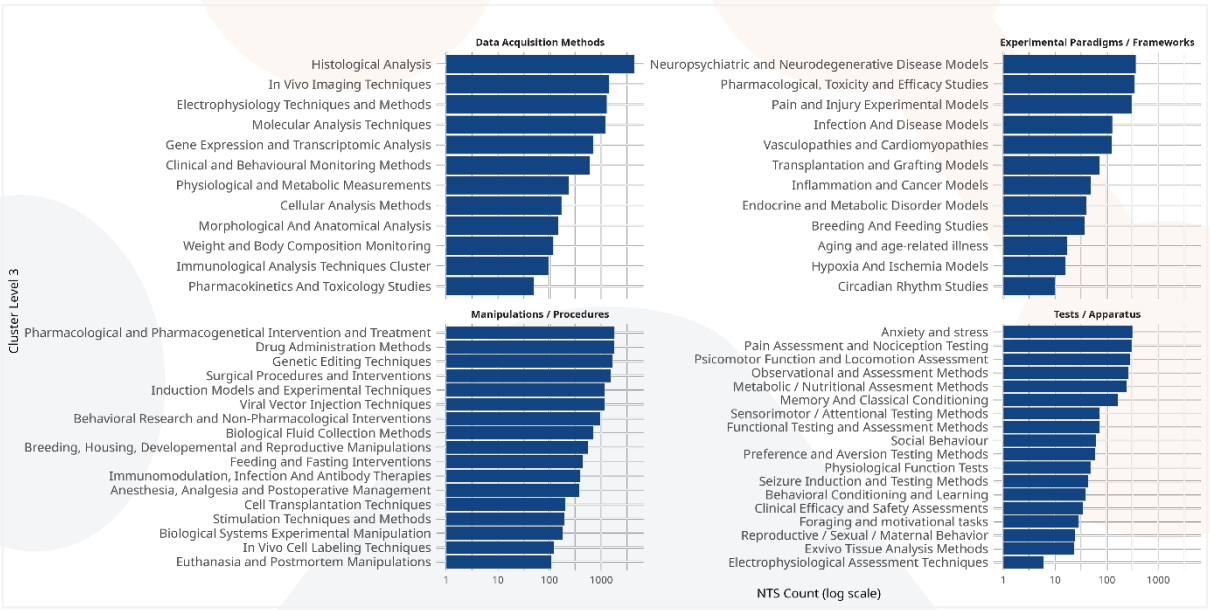


Figure 6. Bar charts illustrating the four defined categories, each displaying the distribution of their respective subclusters.

These trends illustrate how studies are structured to test specific hypotheses within diverse experimental frameworks. In the **Manipulations/Procedures category**, the most frequent subcategories are Pharmacological and Pharmacogenetic Intervention and Treatment (1808), Drug Administration Methods (1783), Genetic Editing Techniques (1645), and Surgical Procedures and Interventions (1529), among others. This reflects the variety of interventions used to modulate experimental outcomes and investigate causal relationships. For the **Tests / Apparatus category**, key

subcategories include Anxiety and Stress (318), Pain Assessment and Nociception Testing (309), and Psychomotor Function and Locomotion Assessment (280), among others. These patterns highlight the diverse array of assessments and instruments employed to measure responses and evaluate the impact of different manipulations. Finally, the **Data Acquisition Methods category** is dominated by Histological Analysis (4343), *In vivo* Imaging Techniques (1437), Electrophysiology Techniques and Methods (1278), and Molecular Analysis Techniques (1233). The prevalence of these methods underscores their essential role in capturing accurate and reliable experimental data for subsequent analysis

This visualization enables easy identification of patterns and trends across all categories, providing valuable insights that support further data analysis and interpretation. Additional related figures and resources are available in the ALURES Analysis tab on the public website. Beyond analyzing the methods and categories individually, it is also possible to investigate which study methods co-occur within the NTS reports, thereby uncovering meaningful relationships between them. To ensure that these relationships can be effectively visualized, such analyses typically focus on a defined subset of clusters or subclusters.

Building on this approach, the public site complements the Experimental Methods Browser with a set of pre-generated graphics that facilitate detailed exploration of study method co-occurrence in NTS. The Study Methods co-occurrence tab features several Sankey network plots that illustrate the relationships between the categories represented in the Experimental Methods Browser. As an example, Figure 7 presents a Sankey plot illustrating the relationships between Experimental Paradigms/Frameworks, Tests/Apparatus, and Data Acquisition Methods. This visualization highlights the sequential and interconnected nature of these elements within the dataset.

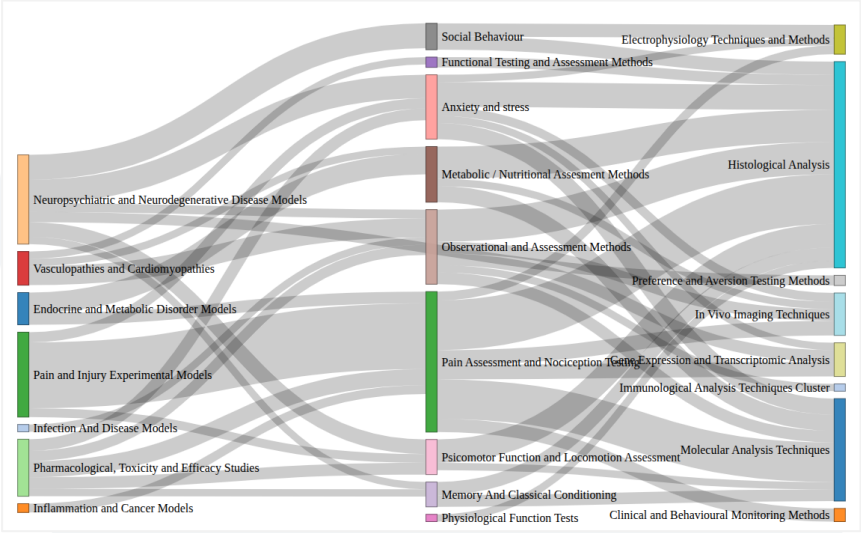


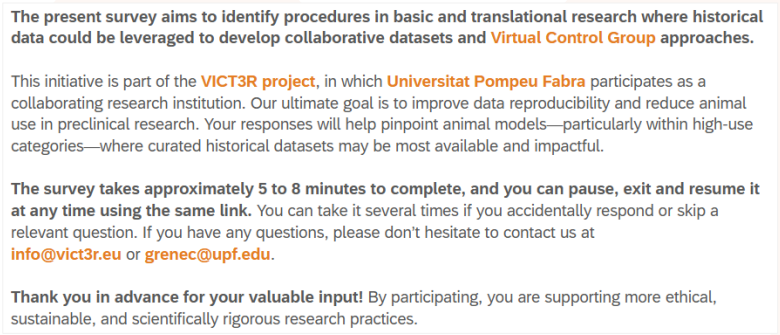
Figure 7. Sankey plot describing the cooccurrences of methods from clusters Experimental Paradigms / Frameworks – >Tests / Apparatus – > Data Acquisition Methods.

When analyzing Neuropsychiatric and Neurodegenerative Disease Models, we observe that they most frequently co-occur with the Anxiety and Stress and Social Behaviour test/apparatus clusters. There are also notable co-occurrences with Psychomotor Function and Locomotion Assessment, Memory and Classical Conditioning, and Observational and Assessment Methods. In addition, the Social Behaviour test cluster shows a strong co-occurrence with the two main Data Acquisition Methods: Histological Analysis and Electrophysiology Techniques and Methods. This brief overview illustrates the type of conclusions and insights that can be drawn from the data. For a comprehensive view, please refer to the Study Methods Co-occurrence tab at https://vict3r.medbioinformatics.com/alures/

4. Expert Survey on the Potential of Academic Models for VCG Development

To evaluate the alignment of existing academic animal models with the VCG concept, an online survey was developed targeting professionals involved in basic and translational research. The **survey aimed to capture expert insights on commonly used *in vivo* models and assess their potential and perceived relevance for VCG approaches**.

The survey was implemented using the **Qualtrics XM** platform and launched in two phases. The **initial phase** targeted researchers affiliated with **UPF and PRBB partner institutions**, focusing on identifying relevant models and internal data practices. This allows us to receive some feedback and implement minor changes. In a second phase, the survey was expanded to a broader set of national and international scientific and 3Rs organizations involved in basic and translational research, to capture a wider range of models, practices, and perspectives.


The present survey aims to identify procedures in basic and translational research where historical data could be leveraged to develop collaborative datasets and Virtual Control Group approaches.

This initiative is part of the **VICT3R project**, in which **Universitat Pompeu Fabra** participates as a collaborating research institution. Our ultimate goal is to improve data reproducibility and reduce animal use in preclinical research. Your responses will help pinpoint animal models—particularly within high-use categories—where curated historical datasets may be most available and impactful.

The survey takes approximately 5 to 8 minutes to complete, and you can pause, exit and resume it at any time using the same link. You can take it several times if you accidentally respond or skip a relevant question. If you have any questions, please don't hesitate to contact us at **info@vict3r.eu** or **grenec@upf.edu**.

Thank you in advance for your valuable input! By participating, you are supporting more ethical, sustainable, and scientifically rigorous research practices.

Figure 8. Introductory message shown to survey participants

The survey gathered both structured and open-ended responses across several **key domains**:

- **Model Characteristics**: Information on the species, strain, sex, and research application of each animal model.
- **VCG Compatibility:** Assessment of model suitability for VCG implementation based on factors such as standardization, reproducibility, and the volume of historical control data.
- **Data Availability:** Insights into the type, scope, and accessibility of legacy datasets, including whether they are stored locally or in shared repositories.
- **Existing Repositories:** Identification of platforms or infrastructures where model-related data are archived.
- **Researcher Perceptions**: Perspectives on the usefulness of historical datasets, as well as the perceived value, challenges, and feasibility of applying VCG approaches in different research contexts.

The survey design combined closed-ended and open-text items to enable consistent data collection while capturing qualitative insights from respondents. Dissemination was supported by institutional mailing lists, research institutions, scientific networks, and 3Rs-aligned initiatives to ensure broad engagement.

4.1. Structure overview

The survey was introduced with a **brief overview of the VCG concept** and the **goals of the VICT3R project**, explaining the purpose of the questionnaire and how the collected information would support the identification of models and data practices compatible with VCG approaches. Respondents were informed that the survey would take approximately 10 - 15 minutes to complete and that all responses would be treated confidentially.

The main questionnaire body was divided into six sections:

i. **Section 1: General Information**
This section collected basic information about the respondent's institution, research area, and professional role, to help contextualize responses and assess the diversity of participating profiles.

ii. **Section 2: Animal Use Characteristics in Your Work**
Respondents were asked to describe the types of animal models used in their current research, including species, strain, sex, and experimental context. Questions also addressed the degree of standardization and typical study designs used.

iii. **Section 3: Data Availability & Potential for Reuse in Your Work**
This section explored the extent and format of historical control data generated or stored by the respondent's group, as well as their accessibility and potential for reuse in the context of VCG approaches.

iv. **Section 4: Feasibility & Perspective in Your Work**
Focused on individual research practices, this section assessed the perceived feasibility, benefits, and limitations of implementing VCGs in the respondent's own studies.

v. **Section 5: Feasibility & Perspective in Your Field**
At a broader level, this section explored the perceived relevance and acceptability of VCG approaches across the respondent's field of research, including factors influencing wider adoption.

vi. **Section 6: Open Comments**
An open-ended space was provided for respondents to share any additional insights, concerns, or suggestions regarding VCG implementation, historical data reuse, or 3Rs implications.

At the **end of the survey**, participants were presented with a farewell message, thanking them for their time and contributions to the VICT3R initiative. Contact information was provided in case respondents wished to receive updates on the project or request further information.

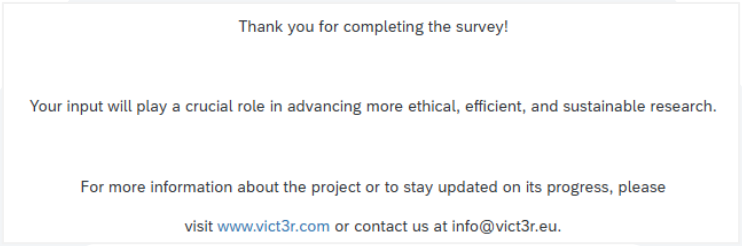


Figure 9. Closing message displayed at the survey end.

Table 1. List of Institutions Reported by Respondents (providing this information was optional).

#	Institution Name	Acronym	Count	Country	Category / Type
1	Hospital del Mar Research Institute	HMRI / IMIM	13	Spain	Research Institute
2	Universitat Pompeu Fabra	UPF	12	Spain	University
3	Vrije Universiteit Brussel	VUB	10	Belgium	University
4	Universidad Complutense de Madrid	UCM	8	Spain	University
5	Centre for Genomic Regulation	CRG	6	Spain	Research Center
6	Uiversidad Nacional de Educación a Distancia	UNED	4	Spain	University
7	Universitat de Barcelona	UB	3	Spain	University
8	University of Murcia	UM	2	Spain	University
9	Universidad de La Laguna	ULL	2	Spain	University
10	Universidad de las Islas Baleares	UIB	2	Spain	University
11	University of the Balearic Islands	UIB	2	Spain	University
12	Universitat Autònoma de Barcelona	UAB	2	Spain	University
13	CSIC	CSIC	2	Spain	Research Council
14	University of Sheffield	TUOS	1	UK	University
15	Aberdeen University	ABDN	1	UK	University
16	Indian Veterinary Research Institute, Bareilly	IVRI	1	India	Research Institute
17	Universidad Rey Juan Carlos	URJC	1	Spain	University
18	Phytoplant Research	-	1	Spain	Research Company
19	AO Foundation	AO	1	Switzerland	Foundation
20	INRLGII	INRLGII	1	Spain	Research Institute
21	IIS La Fe	IISLaFe	1	Spain	Research Institute
22	CNB-CSIC	CNB-CSIC	1	Spain	Research Institute
23	UPV/EHU	UPV/EHU	1	Spain	University
24	Instituto de Investigación Biomédica de Lleida	IRBLleida	1	Spain	Research Institute
25	Universidade de Vigo	UVigo	1	Spain	University
26	OSAKIDETZA	OSAKIDETZA	1	Spain	Health Service
27	Institut Biología Evolutiva CSIC-UPF	IBE	1	Spain	Research Institute
28	noctuRNA Therapeutics	-	1	-	Company
29	PRBB	PRBB	1	Spain	Research Center
30	IUNICS	IUNICS	1	Spain	Research Center
31	University of Lausanne	UNIL	1	Switzerland	University
32	Karolinska Institutet	KI	1	Sweden	University / Research
33	Institute of Functional Genomics	IGF	1	France	Research Institute
34	Universidad de Concepción	UdeC	1	Chile	University
35	Parc Científic de Barcelona	PCB	1	Spain	Research Institute
36	Universitat de les Illes Balears	UIB	1	Spain	University

4.2. Results

Background information in question (Q) 1 from the total of 222 respondents indicates that at least 36 unique institutions (See Table 1) have participated so far in the ongoing VICT3R survey, reflecting a **diverse and multidisciplinary research landscape**. The current sample includes public universities, biomedical and translational research centers, public health agencies, national research organizations, and private biotech companies.

The largest group of respondents (34.2%) identified as postdoctoral researchers, staff scientists, or research associates, indicating that a **substantial portion of the survey participants are directly involved in the day-to-day conduct and design of animal experiments**. Closely following, principal investigators (PIs), professors, co-PIs, and academic faculty (including lecturers and assistant professors) represented 31.6% of the sample, reflecting **strong engagement from senior decision-makers and laboratory leaders**. PhD students and doctoral candidates made up 10.5% of respondents, highlighting the **participation of early-career researchers**. Technical and operational support roles were also represented, with 11.8% of participants identifying as technicians or laboratory managers, who are essential to the **implementation and maintenance of *in vivo* protocols**. A smaller portion of the sample consisted of professionals in veterinary and animal welfare roles (5.3%), such as designated and facility veterinarians, and individuals in management or administrative positions (6.6%), including directors, coordinators, and administrative staff.

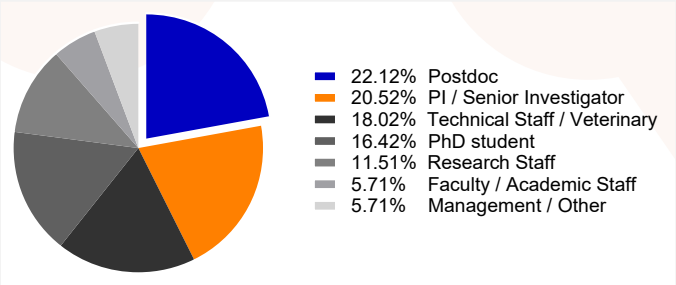


Figure 10. Main respondents' role /position.

Accordingly, 80% of respondents reported **routinely conducting *in vivo* research involving animal** models (Q2), while a smaller proportion (20%) indicated they do not. Regarding species typically used (Q3), results reveal a clear predominance of **rodent models** in current research practices. Mice are the most widely used species, reported by 88% of respondents, underscoring their central role in both basic and translational biomedical research. Rats follow at a considerable distance, being used by 36% of participants (See figure X. Please note that respondents were allowed to select more than one).

When respondents were asked to indicate their field of expertise within Basic Research (Q4), **60% identified Nervous System–related research** as their primary area of expertise. This underscores the central role of neuroscience within the academic biomedical landscape and indicates substantial potential for VCG applications in this domain. Beyond the nervous system, research on the Immune System (22%), Ethology and Animal Behavior (19%), and Oncology (15%) also featured prominently.

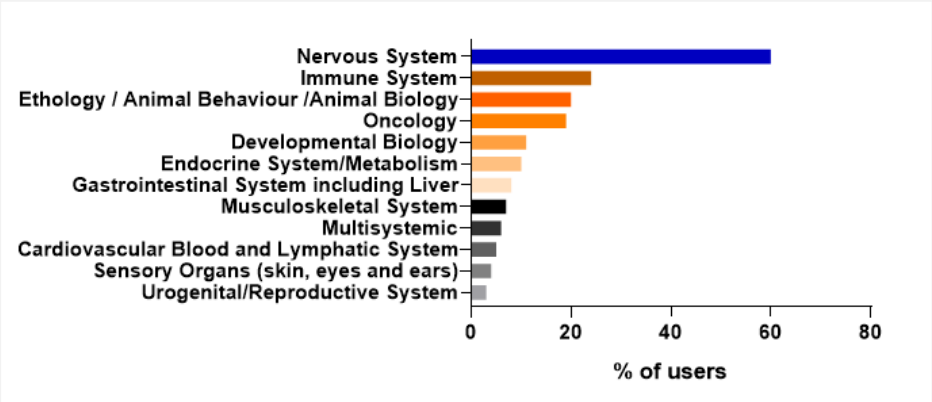


Figure 11. Fields of expertise within Basic Research.

Likewise, responses in the Category/Field of Translational Research (Q5) revealed a strong focus on **Human Nervous and Mental Disorders**, representing 43% of participants (See figure 12).

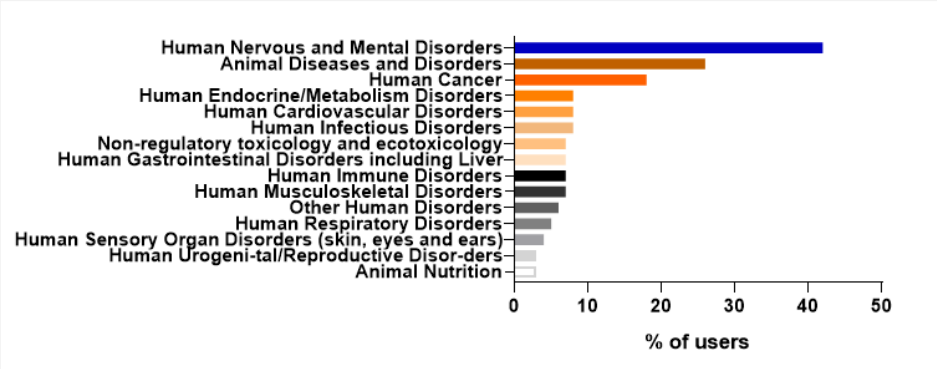


Figure 12. Fields of expertise within Translational Research.

4.2.1. Animal Use Characteristics in Respondents' Research

A total of **222** paradigms, setup or basic procedure mentions (Q6) were recorded. 68% of respondents explicitly identified using two paradigms, and 44% listed three paradigms (the maximum allowed). Aligning the experimental paradigms with ALURES categories indicates **broad thematic convergence with self-reported expertise**, particularly in areas involving nervous system research and associated disorders (Table 2).

Although survey respondents were asked to provide experimental paradigms, many responses also included manipulations, endpoints, and other aspects of experimental design. This broader range of input enabled classification of the entries into semantic clusters, similar to previous NLP analyses of the ALURES NTS. The **largest category was Experimental Frameworks**, comprising over half of all entries (90 entries, 57.7%), followed by Manipulations / Procedures (28 entries, 18.0%) and Endpoints: Tests / Apparatus (24 entries, 15.4%). Very few respondents referred to Data Acquisition Methods (3 entries, 1.9%), while 11 entries (7.1%) fell into Other / Unclassified.

Table 2. Frequencies of research purposes.

Basic Research Fields	Freq	Survey %	ALURES %
Nervous System	**51**	**48.1%**	**24.22**
Immune System	23	21.7%	16.23
Endocrine System/Metabolism	18	17.0%	5.14
Gastrointestinal System including Liver	18	17.0%	4.68
Oncology	11	10.4%	15.91
Ethology / Animal Behaviour	9	8.5%	8.37
Developmental Biology	8	7.5%	1.80
Others basic research	4	3.8%	1.65

Translational Research Fields	Freq	Survey %	ALURES %
Human Nervous and Mental Disorders	**60**	**56.6%**	**7.75**
Human Immune Disorders	18	17.0%	2.21
Human Endocrine/Metabolism Disorders	18	17.0%	2.82
Human Gastrointestinal Disorders including Liver	18	17.0%	1.07
Other Human Disorders	12	11.3%	0.84
Human Cancer	11	10.4%	15.34
Human Infectious Disorders	5	4.7%	6.41

The **Experimental Frameworks** were heavily dominated by Neuroscience Models, which account for 42.2% of this category. These paradigms cover a wide range of conditions, from neurodegenerative diseases such as Alzheimer's to psychiatric disorders including schizophrenia and addiction. The next most prominent groups are Oncology Models and Systemic & Disease Models, each representing 15.6% of the frameworks (14 entries each), highlighting a strong focus on cancer research and broader physiological conditions such as diabetes and toxicity. General / Foundational Models make up 10%, while Immunology and Inflammation and Regenerative & Surgical Models represent 8.9% and 7.8%, respectively. Within **Manipulations / Procedures**, Substance Administration is the largest sub-cluster, accounting for 35.7% of entries. This reflects the frequent use of drugs, alcohol, and other compounds to induce experimental effects. Surgical & Tissue Procedures (28.6%) and Stress & Behavioral Manipulations (25%) follow closely, highlighting the importance of invasive techniques and stress induction in experimental designs. Dietary & Gut Microbiome manipulations are less common, representing 7.1% of entries. The **analysis of endpoints and tests** reveals a near-even split between two major areas of behavioral science. Learning & Affective Behavior paradigms, which include tests for memory, anxiety, and depression, make up half of the entries. The other dominant group is Addiction & Consumption, which slightly trails and includes various self-administration and preference tests. This strong focus indicates that a majority of the reported outcomes are behavioral, centering on cognition, mood, and substance abuse. Regarding **Data Acquisition Methods** reported, the scarcity of entries made the creation of further semantic sub-clusters impractical and suggests that respondents focused more on the overarching experimental model rather than on the specific methods used to acquire the final data.

Despite the diversity of experimental paradigms reported, the **average number of animals per control group** (Q7) **was approximately 11**, a figure typical of preclinical academic research. Respondents reported repeating a given experimental paradigm (Q8) up to four times on average over a two-year period, with individual responses ranging from 0 to 10 repetitions. This indicates that while some paradigms are used only once or twice, many are reused multiple times with minor variations, highlighting their **potential for reusability and value in retrospective data integration**.

4.2.2. Data Availability & Potential for Reuse in Respondents' Research

Regarding practices of data storage and records of experimental results (Q9) a significant majority of respondents (85%) indicated that they systematically archive such data. Meanwhile, 15% reported that they partially retain records, depending on the specific project. Among (51) respondents, the most commonly used data archiving format (Q10) is spreadsheet files (e.g., Excel), reported by 84%. This is followed by paper records or lab notebooks (69%) and image/video file formats (57%), highlighting a mix of digital and analog storage practices (Figure 13). A majority of researchers retain data (Q11) for extended periods, with 45% storing it for more than 7 years and an additional 22% following conditional retention policies based on study type or funding.

When asked about **control data reuse** (Q12), only 18% of respondents reported having formally reused control data across experiments, studies, or projects. 32% reported informal reuse, typically in pilot studies, while 14% had considered reuse but had not yet done so. Notably, 37% of researchers reported never having reused control data, particularly in disease-specific genetic models and some behavioural tests, such as three-chamber or cognitive-social tasks. Formal reuse was most common in oncology (e.g., carcinogen-induced colorectal cancer), certain procedural interventions (e.g., time-restricted feeding), and basic behavioural assays (e.g., FST, cognitive tasks for dementia). Informal reuse or pilot applications spanned pain, stress, immunology models, and behavioural phenotyping, including open field, operant conditioning, and restraint stress.

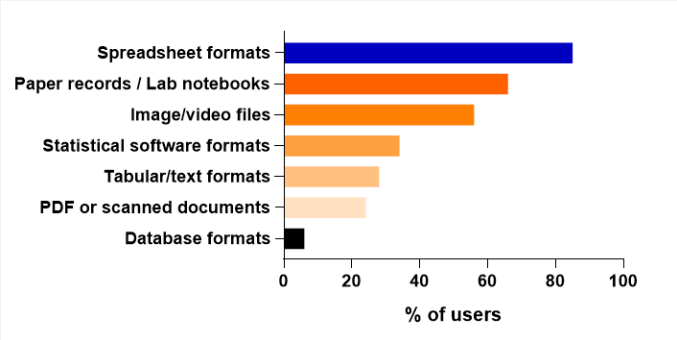


Figure 13. Most common practices of data storage

The **awareness of shared historical datasets** (Q13) varied widely across reported experimental paradigms. Overall, respondents were more likely to know of **existing datasets in their main experimental paradigm**. However, a majority remained unaware or uncertain across all paradigms: 56% in the first (26% no, 30% unsure), 80% in the second (40% no, 40% unsure), and 78% in the third (39% no, 39% unsure).

Regarding the **perceived usefulness of VCGs** (Q15), responses were consistent across paradigms approximately 40–45% of respondents considered them useful, while 25–30% expressed doubts, and a similar proportion remained unsure. Most **neurological and behavioural related paradigms were categorized as "useful" for VCG applications**. These included widely used pain models (Partial Sciatic Nerve Ligation, Reserpine-Induced Myalgia), neurodegenerative disease models (Huntington's disease models), addiction and drug exposure (operant conditioning, opioid self-administration, cannabinoid exposure), some classical learning tasks and behavioural assays (fear conditioning, open field). Immunology/infectious disease paradigms were judged positively (i.e. mucosal immunity, pathogen clearance studies). Oncology was mostly considered "unsure" or "non useful". Genetic and transgenic models were evenly split across "useful", "non useful", and "unsure" categories, suggesting divergent model-dependent opinions.

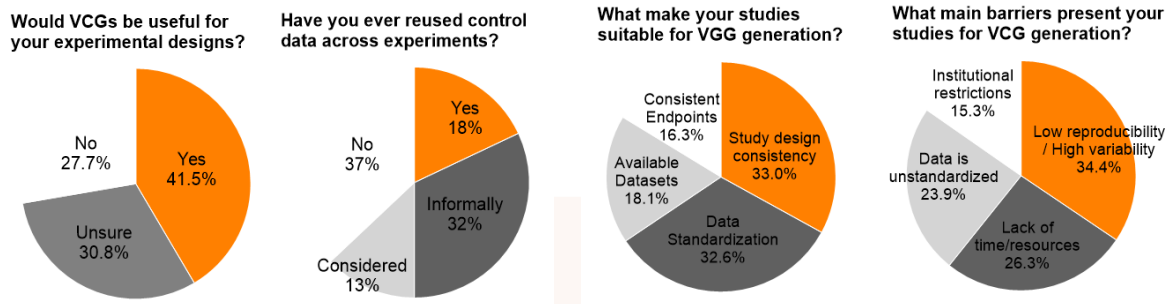


Figure 14. Respondents control data reuse practices and perceptions on Virtual Control Group application.

The **potential applications of a historical database** (**Q16**) for respondents' research revealed "**Data Exploration"** as the top priority among the options provided (see method section for the full option list). "**Study Design Optimization**" followed closely, frequently ranked second, reflecting the value placed on improving experimental planning based on historical data. Relevantly, while the "**Generation of VCGs**" was acknowledged, it tended to be ranked in the middle to lower tiers.

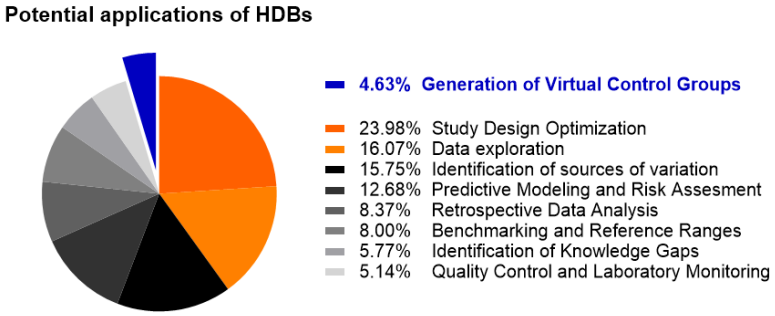


Figure 15. Respondents preferred applications of Historical Databases.

Despite this, most respondents expressed **a positive stance toward data sharing** for the development of a historical database with VCG generation capacity (Q17). While 33% indicated they would be open to sharing data outright, only 13% reported unwillingness to share data. The largest portion, 54%, were conditionally open, emphasizing the **importance of factors such as anonymization,** proper credit, or other safeguards (Q18). Only 4% of respondents anticipated additional specific conditions should be met for data sharing (Q19). When asked to specify these conditions (Q20), only one respondent provided further input, highlighting **the potential for AI and simulation platforms to**

complement historical data sharing. This respondent suggested that predictive modelling could help reduce animal use by forecasting experimental outcomes from preliminary data, which would then require only limited in vitro validation. While the use of synthetic data is already a goal within VICT3R, its success in the academic setting will depend on the availability of large, comprehensive, and highly granular datasets to ensure reliability and validity. This will be further explored in greater detail in the discussion (Sections 5.3. and 5.4).

4.2.3. Feasibility & Perspective in Respondents' Work

Among respondents, **study design consistency** emerged as the most important aspect for applying a VCG approach (Q21), selected by approximately 57% of participants. In contrast, the presence of existing datasets was the factor least frequently indicated, reinforcing earlier findings that, although many studies are standardized or consistent, shared historical datasets remain limited.

Only a small minority (2%) identified additional aspects of their studies that might make them suitable for VCG inclusion (Q22), suggesting that most researchers consider the key suitability factors to have been captured in the predefined options, or have not yet identified further considerations. When asked to elaborate (Q23), one respondent noted that all datasets from their published papers and preprints are already publicly available via the Open Science Framework (OSF), freely accessible for reuse.

The **most perceived barrier to implementing or contributing to a VCG** approach (Q24) was concerns about reproducibility and variability, cited by 80%, 68%, and 56% of respondents for their first, second, and third experimental paradigms, respectively. Notably, the overall pattern of reported barriers showed minimal variation across these paradigms, suggesting a shared perception of the key challenges regardless of research focus. Unstandardized data and lack of time or resources were also consistently highlighted, with similar proportions across paradigms, underscoring that both technical and logistical limitations are widely recognized. Institutional restrictions were less frequently reported but remain a relevant consideration for some researchers.

Two respondents identified specific additional concerns (Q25), and their comments (Q26) underscore a persistent scepticism about the feasibility of VCGs. One respondent questioned the scientific validity of relying entirely on virtual controls in behavioural neuroscience, expressing concern that grant and manuscript reviewers might not accept such methods. The other emphasized the pronounced influence of contextual variables—such as season, time of day, housing conditions, experimenter, and ambient temperature—arguing that these sources of variability pose significant challenges to reusing control data across experiments.

4.2.4. Feasibility & Perspective in Respondents' Field

When respondents were asked to identify up to three experimental paradigms that tend to show less variation in study design within their broader research field, the answers initially appeared highly diverse, even within similar fields. However, five major domains recurred: behavioural, cognitive, and addiction paradigms (26.8% of responses); neurodegenerative/CNS models (17.1%); pain and toxicity

assays (14.6%); pharmacology, molecular, and immunology approaches (14.6%); and tumor/oncology models (12.2%). Examples included tumor growth and bone disease models, cisplatin- or vincristine-induced adverse effects, streptozotocin-induced diabetes, mucosal immunology in transgenic mice, maternal separation, and middle cerebral artery occlusion in stroke. Importantly, respondents also cited a wide range of behavioral and functional assays, such as operant and food self-administration, touchscreen tasks, locomotor activity, pre-pulse inhibition, rotarod, novel object recognition, and open field tests.

Despite this apparent diversity, most paradigms cluster into a few recurring domains. Tumor growth and related oncological models were repeatedly mentioned across paradigms, whether through induced cancers, bone disease, or tumor cell injections. Pain and toxicity assays also recurred, though expressed via different implementations—sciatic nerve ligation, hot plate, von Frey, cisplatin- and vincristine-induced effects, or grip strength tests. Addiction paradigms appeared in multiple forms, including food and drug self-administration, heroin paradigms, and acute pharmacological responses. Behavioral and cognitive neuroscience tasks were frequently cited, ranging from avoidance learning and maternal separation to touchscreen-based and novel object recognition assays. Finally, neurodegenerative or psychiatric models—including Alzheimer's, stroke, Marfan syndrome, and "double hit" schizophrenia models—were consistently represented.

4.2.5. Respondents' Open Comments

Several respondents offered closing remarks (Q30) that reflected both support and cautious reflection on the VCG initiative. Some expressed strong interest, emphasizing the potential benefits for the 3Rs, improved data organization, and reduced redundancy in experimental design. Others voiced reservations, particularly regarding the validity of virtual controls in complex behavioral studies, where genetic drift, contextual variability, and experimenter effects can pose significant challenges. One respondent highlighted the value of convergence across studies rather than strict standardization. Overall, these comments underscore broad interest in the initiative while highlighting the need for nuanced implementation and ongoing dialogue with the research community.

5. Discussion

We have applied multiple complementary approaches - statistical analysis of ALURES, NLP-guided extraction of Non-Technical Summaries, and a targeted academic survey - to provide a **detailed overview of animal-based research practices in academia**. Together, these analyses highlight recurring experimental paradigms, data types, and procedural patterns that are most relevant for the development of HDBs and VCGs.

NLP analysis of Nervous System NTS records identified Neuropsychiatric and Neurodegenerative Disease Models, as well as Pharmacological, Toxicity, and Pain-Related Studies, as the predominant paradigms. The **expert survey** highlighted pain, neurodegenerative, and addiction paradigms as most VCG-relevant and showed strong willingness to share data via platforms like VICT3R. In these models, raw data likely describe Pharmacological and Pharmacogenetic manipulations and behavioral assessments for Anxiety, Stress, and Pain.

While the findings reveal both **opportunities and challenges**, including variability in reporting and limited awareness of shared datasets, they also identify concrete niches where VCG implementation is feasible. This discussion will explore these results in depth, addressing their **implications for standardization, data reuse, and broader application of VCG methodologies in academic research.**

5.1. Discussion on the NLP guided analysis of the ALURES Nervous System NTS

The integration of **LLMs into the analysis of ALURES NTS** records demonstrates a **powerful approach** to addressing the inherent complexity of multilingual, unstructured data in animal research. By automating the extraction, translation, and standardization of experimental procedures, the agent-based pipeline effectively reduces linguistic and terminological barriers that have traditionally hindered large-scale analysis. The use of multi-agent debate frameworks ensures a higher degree of reliability and consistency across languages, while semantic clustering and expert curation bridge the gap between automated processing and domain-specific understanding. This combination of artificial and human intelligence creates a scalable, interpretable, and linguistically inclusive solution for organizing research practices. The **curated taxonomy**, together with the **interactive browser**, enables comparative analyses that were previously infeasible. For instance, co-occurrence analyses and visualizations (e.g., Sankey plots) could provide a way to uncover patterns and relationships between study paradigms, apparatus, and data acquisition methods, potentially revealing new trends or methodological biases within preclinical research domains.

Analysis of the dataset revealed clear and consistent research practices relevant for VCG approaches. Neuropsychiatric and Neurodegenerative Disease Models were the most frequent paradigms, followed by pharmacological and pain-related models, reflecting a focus on translational and disease-oriented research. Manipulations included pharmacological, genetic, and surgical interventions, while tests emphasized anxiety, pain, and locomotion. Data acquisition relied on histological, imaging, electrophysiological, and molecular techniques. Co-occurrence analyses showed that study types and methods are interconnected, emphasizing the integrative nature of experimental designs, where behavioral, physiological, and molecular assessments are often combined to provide comprehensive insights.

While primarily designed for the purposes of this deliverable, the transformation of fragmented textual data into an organized knowledge resource could ultimately support researchers, policymakers, and ethics committees in better understanding the **diversity and prevalence of experimental practices across Europe**. Looking forward, the combination of LLM-driven extraction and expert-guided refinement represents a scalable model for future data harmonization in biomedical research. As LLM capabilities continue to advance, integrating them with controlled vocabularies or domain ontologies could further enhance precision and interoperability with other scientific datasets. This approach not only improves data accessibility and analytical potential but also contributes to more informed decision-making, fostering transparency, reproducibility, and ethical rigor in animal research.

Nonetheless, **NLP-guided analysis of ALURES Nervous System NTS is limited by several factors**. Incomplete reporting in some NTS reduces the ability to capture all relevant manipulations, tests, or data acquisition methods. NLP approaches primarily quantify frequency and co-occurrence patterns, but cannot assess experimental quality, reproducibility, or outcome validity. NTS also represent planned rather than completed experiments, so actual implementation may differ. Additionally, variability in terminology across laboratories can lead to misclassification. Incorporating LLM-guided extraction and standardized ontologies could help mitigate these limitations.

5.2. Discussion on the Basic and Translational Research Survey Results

The diversity obtained in the survey ensures a **representative view of academic animal research**, capturing both strategic oversight and hands-on experimental practice. The respondents' fields of expertise align with the predominant use of rodent models and the emphasis on nervous system research observed in the ALURES database, reinforcing the survey's relevance within the broader regulatory and scientific landscape and reaffirming neuroscience as a central pillar of basic and translational research, alongside significant contributions from cancer and immunology which also hold potential research practices for further development within VICT3R and similar initiatives.

Interestingly, the main experimental paradigms in which researchers work are more likely to have an associated public database compared to their second or third most frequently used paradigms. This suggests that researchers tend to rely on established shared datasets for their primary paradigm, possibly because these areas benefit from more standardized methodologies and stronger data infrastructure. In line with this, primary paradigms were reported to have higher levels of study design consistency and data standardization—features that would presumably make them better positioned for VCG generation.

Nevertheless, the majority of respondents indicated being unaware of any shared historical datasets, even in their primary research areas. This points to persistent gaps in the visibility, dissemination, and generation of databases. The existence of a database of industry study types—RDT studies in the case of VICT3R DB—can serve as a powerful catalyst for academic researchers to apply, adapt, and innovate upon these established experimental paradigms. This collaborative approach can foster both methodological advancement and data enrichment within the field.

The results on **availability and potential for reuse** indicate a general commitment to structured data practices. Spreadsheets remain the most commonly used data archiving format—and are acceptable for later transformation into standardized formats like SEND. Moreover, the fact that most respondents report storing data for more than seven years (without considering already published research and raw data) in some cases, suggests a generally favorable environment for long-term reuse and traceability.

However, the **perceived usefulness of VCGs** (Q15) reflects a broadly positive yet cautious attitude toward their implementation. While many respondents see clear potential, the consistent presence of uncertainty likely stems from the approach's novelty and concerns about methodological limitations in specific research contexts. This highlights the **need for further awareness, validation, and demonstration of VCG benefits to encourage broader acceptance and implementation in different research contexts**. As a matter of fact, respondents showed a strong interest in using historical datasets primarily to support exploratory analysis and experimental design, while more advanced or specialized applications, like VCG generation and predictive modelling, are seen as promising but currently secondary. Addressing perceived barriers related to data consistency, resource availability, and reproducibility will be essential to enable broader adoption and more sophisticated use of historical data.

The results point to a **strong potential for collaborative data sharing**, which is unsurprising given that most control and experimental data are already intended for publication. However, respondents expressed a conditional openness, noting that data sharing is most acceptable when it is **secure, appropriately credited, and regulated by clear agreements**. These preferences highlight the importance of robust frameworks that address both ethical and practical considerations, including well-defined guidelines and reliable infrastructure to support secure, transparent, and **mutually beneficial data contributions**.

Despite the diversity of experimental paradigms reported, **several recurring domains emerge as particularly promising for historical database development and VCG generation**. Behavioral, cognitive, and addiction paradigms, along with neurodegenerative/CNS models and pain assays, were frequently used and consistently rated as useful for VCG applications. Classical behavioral assays (e.g., open field, operant and food self-administration, fear conditioning) and standardized pain models (e.g., Partial Sciatic Nerve Ligation, Reserpine-induced myalgia) are especially amenable due to their recurrent use and relatively uniform protocols. In contrast, oncology and certain genetic/transgenic models showed higher variability and lower perceived VCG utility, indicating the need for additional standardization. Overall, paradigms with repeated, well-documented use in neuroscience, and more specifically behavioral pharmacology, pain, and addiction research seemed to offer the greatest potential for reducing animal use and enhancing reproducibility through HDBs and VCGs.

In summary, responses reflect both **opportunity and caution**. There is a clear set of commonly used paradigms that could support collaborative data sharing, but practical and technical constraints, particularly in highly context-sensitive models, emphasize the need for robust frameworks to ensure reproducibility and trust. The data suggest that cumulative datasets for VCGs are possible, provided that standardization and careful documentation of experimental conditions are implemented.

While the survey provides valuable insights into academic animal research and the potential for HDB and VCG applications, **several limitations should be acknowledged**. The respondent pool, although drawn from diverse academic institutions, may not fully represent the diversity of research areas, potentially overemphasizing fields such as neuroscience and behavioral pharmacology while underrepresenting oncology, transgenic, or emerging experimental models. Responses relied on self-reporting, introducing potential recall bias and variability in how researchers interpreted questions about data reuse, paradigm use, and VCG utility. Additionally, the survey captured only a high-level view of experimental paradigms, without sufficient details regarding methodologies, metadata, or study-specific variables, which are critical for assessing the feasibility of control data reuse. Conditional or uncertain responses regarding data sharing and VCG usefulness further reflect the novelty of these concepts and highlight gaps in familiarity or experience among participants. Finally, selection bias may have favored respondents already interested in data sharing and retrospective control use, and the reliance on descriptive self-reported data limits independent verification of practices. Together, these factors suggest that while the survey identifies key trends and opportunities, the results should be interpreted cautiously, particularly when generalizing to the broader academic research landscape.

5.3. On the use of SEND-like reports and the generation of Historical Datasets in Academic and Translational Research

The VICT3R initiative foresees that the parameters and metadata of additional study types, including those from academic research, should be compared to the **SEND formats and database structures used in WP2 and WP5**. Consequently, the feasibility of mapping these data to common formats and potential adaptations, including terminology for parameters not covered by WP2, must be carefully investigated. In close alignment with WP2 and WP5, **modifications to the VICT3R database will be implemented to accommodate these additional data types**. In this context, a comprehensive understanding of the current landscape of animal data reporting and reuse in academia is essential, including the **implementation of FAIR principles** for data interoperability and their potential interaction with the regulatory framework for VCG, **including the potential adoption of SEND standards.**

5.3.1. Regulatory context of VCGs and the use of SEND for Virtual Control Group generation

Currently, the VICT3R database consists of many studies formatted according to the Standard for Exchange of Nonclinical Data (SEND), a standard developed by the Clinical Data Interchange Standards Consortium (CDISC). The **SEND standard has limited use in academic** pursuits as it has been specifically designed for pre-clinical or chemical safety study data. SEND formatting is required when submitting pre-clinical study data to FDA, for instance in support of Investigational New Drugs applications, although other regulatory agencies use or are piloting the implementation of the SEND format.

There is a contrast between the regulatory rigor enabled by SEND reports and their limited broader scientific utility. **Data in SEND reports are organized into domains that correspond to experimental outcomes or study metadata**, with each domain stored in a separate file. As data from different

domains are linked through study and subject identifiers, thereby enabling relational integrity, SEND reports can be mapped into a relational database schema. These domains are specifically tailored to pre-clinical study-types like single- and repeated-dose toxicity, carcinogenicity and safety pharmacology studies (e.g., regarding respiratory and cardiovascular testing). Another **set of implementation guideline** has also developed for specialized studies, including **developmental and reproductive toxicology (DART) studies** (e.g., embryo-fetal development, juvenile toxicity and fertility and postnatal studies), as well as for **studies conducted under the Animal Rule enforced by the FDA** (*i.e.*, efficacy studies of drugs where clinical testing would not be ethical/feasible). **Expansion of SEND to genotoxicity studies** (e.g., in vivo micro nucleus test and comet assay) is also underway. In SEND, each variable has strict specifications: variable named, typed (i.e. numerical, character, date, etc.), and qualifier variables which determine the details of a parameter (such as the location of a lesion or the instrument used to perform a reading). Some of these variables utilize **controlled terminology maintained and continuously expanded by CDISC to ensure semantic interoperability**. Despite the richness of these datasets, their value remains largely restricted to study sponsors, contract research organizations, and regulators, with limited public accessibility and, consequently, limited potential for reuse in academic (or industry) research.

SEND has demonstrated potential for enabling efficient cross-study data analysis, including insights into drug attrition (Kaufman et al., 2016). By organizing data around individual test animals, SEND facilitates integration and comparison across studies, helping to identify key sources of variability such as the incidence of background findings.

This standardized structure also supports the development of VCGs. The VICT3R consortium is advancing the VCG concept with the SEND standard in mind, matching historical control data in SEND format with future concurrent control groups. Nevertheless, the use of historical control data in regulatory settings does not inherently require SEND; instead, only general guidance on statistical inference is provided (Kluxen et al., 2021; Coja et al., 2022; EFSA PRR Panel, 2025).

5.3.2. FAIR principles for interoperability of academic animal data in the context of VCGs

FAIR data in academic studies

Despite SEND strengths, further harmonization is still required due to variability in industry implementation (Carfagna et al., 2020). **In academic settings,** the diversity of animal studies, together with variability in study designs, reporting practices, and the continuous evolution of new methods, poses **additional challenges for a possible uptake of SEND-based VCGs as envisioned by VICT3R** initiative. Though SEND has found limited uptake in academia, researchers are increasingly guided by the principles of FAIR data (Findable, Accessible, Interoperable, and Reusable) and the use of ontologies. **FAIR Principles do not impose specific requirements for data harmonization**, like regulatory studies with FDA-mandated SEND, but provide a framework that researchers can voluntarily adopt (Wilkinson et al., 2016). Thus, FAIRness does not mandate specific harmonization, like SEND does. Academic studies often involve diverse designs, novel endpoints, or exploratory methods that may not be readily accommodated within rigid data standards such as SEND. Therefore,

while the SEND format offers a robust and FAIR framework for harmonization, its utility in academia is constrained by its regulatory origins and lack of incentives for uptake, restrictions set by the standard, and a high barrier to entry for most academic laboratories.

FAIR principles are just intended to guide researchers to generate, codify and store their animal research data in ways that make them suitable for reuse. In addition, recoverable data can be potentially implemented in VCG creation, as long as it is properly formatted and findable. Nevertheless, data reanalysis for purposes like VCGs is not commonly practiced with FAIR-based workflows. Instead, FAIR is used more often for assisting researchers to locate external datasets (Foster et al., 2024), facilitation of meta-analyses through improved data comparability, and foster cross-disciplinary collaboration by increasing retrievability and interoperability. **General frameworks for FAIRification exist** (https://www.go-fair.org/), which may guide researchers in the broader concepts of data management. FAIR data standards can be found here: https://fairsharing.org/search?fairsharingRegistry=Standard. FAIR principles contribute to reduction of animal use by enabling better reuse of existing datasets instead of repeating studies.

Background: Linked Open Data (LOD), Resource Description Frameworks (RDF) and Ontologies

A challenge for VICT3R but also other collaborations where data is shared are differences in the meaning of shared variables which creates a major barrier to data integration. The intention of SEND was to create regulatory exchange possible. Though SEND is a tabular data format with specific domains and controlled terminology updated according to developing knowledge and regulatory needs, it is not strictly and uniformly enforced. This is why **initiatives like VICT3R must perform curation to ensure consistency**. SEND is machine-readable, since SEND data files can be parsed, but this methodology often ignores concepts like semantic interpretation and linking of datasets, explained below.

In certain FAIR data approaches, the **Linked Open Data (LOD)** paradigm offers a promising solution. LOD is the process of publishing and linking data so that it may be easily accessed through machines. By utilizing shared vocabularies, or ontologies, this can facilitate integration of historical data. Through using **RDF (Resource Description Framework)**, the foundational data model for LOD, which represents information as subject-predicate-object triples, and its query language SPARQL, researchers can retrieve data across pharmacological studies, enabling the reanalysis of drug targets, side effects, or mechanistic pathways by combining results from diverse experiments. RDF triple organization (subject-predicate-object) enables representation and semantic linking of concepts. For instance, *“Sprague Dawley rat” → is_a → “Rat” → is_a → “Rodent”*. RDF triples also enable datasets to interlink regardless of original file format which may vary among laboratories. RDF uses globally Unique Persistent Identifiers (URIs) to encode concepts, enabling to reference external databases, thus enriching data with external knowledge. This semantic linking enables synonym linkage and enhances interoperability, which can be applicable in VCG selection. This also highlights a key difference: preclinical (SEND) and clinical CDISC standards enable structured data sharing for regulatory submissions, however, true semantic interoperability across systems is supported by coding frameworks like the **Logical Observation Identifiers Names and Codes (LOINC)**, which is widely used for laboratory tests and clinical observations in patient data. LOINC provides consistent, universal

identifiers and is linked into the LOD ecosystem, ensuring that test results and measurements retain the same meaning across different contexts and systems.

Ontologies also play a key role in LOD by promoting a shared vocabulary across different study designs and are central to the FAIR data principles by making data machine interoperable and therefore reusable (Hoehndorf et al., 2015; Masuya et al., 2012). An ontology is a formal, machine-readable representation of a domain that includes a controlled vocabulary of concepts, the relationships between them, unique identifiers, and logical rules. As an example, similar terms like "heart attack" and "myocardial infarction" become the same, and with relationships it is possible to understand the study conditions thus enabling machine readability. The complexity of biomedical data requires ontological modelling. The omics and bioinformatics fields are particularly advanced with ontologies and integration since the first established ontology, Gene Ontology. The Open Biological and Biomedical Ontologies (OBO) Foundry maintained by National Center for Biotechnology Information (NCBI) keeps the list of ontologies, including those related to animals like UBERON, a cross-species anatomy ontology that integrates anatomical terms, and experiment-specific ones like the Ontology for Biomedical Investigations (OBI) regarding experimental protocols and study designs, including those involving animal subjects, Experimental Factor Ontology (EFO) for experimental variables, including organism types and conditions, or Animal Behaviour Ontology (ABO) for animal behaviours observed in animal studies. The ontologies can be mapped with tools like easy-to-use add-ons for Google suite Ontomaton (Maguire et al., 2013) or ZOOMA to map free-text to ontology terms (https://www.ebi.ac.uk/spot/zooma/). In addition, BioPortal is one of the largest repositories of biomedical ontologies helping to harness the NCBI ontologies (https://bioportal.bioontology.org/).

The challenges in the use of foundational ontologies in biomedical research are highlighted by Bernabé et al. (2023). The authors note that ontology frameworks can bring semantic rigor through structured, philosophically grounded frameworks, but their abstract concepts create a barrier in their practical use and do not align with the real needs of researchers, who focus on tangible entities like gene functions and disease phenotypes. As a result, researchers frequently prefer simpler solutions like controlled vocabularies, exchanging formal precision for ease of use. Controlled vocabularies don't capture relationships like ontologies do, hence they are less machine-usable or interoperable. The relationships enable to create knowledge networks, which is what makes FAIR data interoperable and reusable at scale.

SEND was not inherently intended for integration. It utilizes controlled terminologies, but those terms may vary in practice across data sources, hence considerable curation and standardization is necessary. If controlled terminologies are enforced, SEND variables can then be mapped to interoperable identifiers, giving them globally unique, ontology-based semantics that make the data FAIRer by supporting linking and machine interpretation. For example, SEND data can theoretically be linked to the NCI Thesaurus (NCIt) biomedical ontology, which assigns unique codes to scientific and medical terms, enhancing interoperability. Beyond this, SEND can also serve as a source for enriching human-focused ontologies, enabling them to cover animal-specific concepts and thereby support preclinical-to-clinical data interoperability, as demonstrated in radiation oncology by Giraldo et al. (2024).

FAIRification of data and metadata sharing are all important steps to be able to compare and reuse similar studies in ongoing research, such as in the context of VCG. The role of these terms in codifying and linking data which may be reutilised for animal data reuse and VCG creation is manifold. Firstly, it creates a hierarchically organised, heterogeneous and semantically linked terminology, which introduces consistency and findability. Secondly, they enable both semantic reasoning to align data and can enrich it with external database information. Thirdly, RDF and ontologies allow important metadata to be recovered which is important for traceability. Finally, these frameworks could be more machine-readable: algorithms can automatically query and filter repositories for matching controls.

Despite this, there are some critical challenges in linking biomedical data and low reuse of published schemas as each source often utilizes a unique structure which impedes data integration and reuse (Kamdar & Musen, 2021).

5.3.3. The landscape of animal data reporting and reuse in academia

Finding relevant studies for VCGs: animal study metadata sharing in practice

Harmonized and ontology-based data analysis in academia is not available for every type of study, while regulatory standardization enables much more advanced approaches like historical control data reanalysis and VCG development in VICT3R. In practice, fully public animal raw data and even metadata on the life sciences is often difficult to find for researchers, but its reusability can be unleashed via repositories like Zenodo, Figshare or other more field-specific databases. Sharing metadata can enable researchers to build collaborative networks and reuse data, if they are interested in maintaining historical controls or analyzing study variability, which is crucial during the data reproducibility crisis. Databases which follow LOD and ontologies also enable synonym linking and enriched metadata which could in the future help in the selection of a suitable control group, if metadata is exported as RDF and using ontologies, it would create the best-case scenario for data retrieval.

Location of **relevant metadata could be considered the first step in finding relevant historical data** for potential employment in a VCG. Multiple ways to report metadata have been described which are very research-specific: e.g. microarray, genomic, proteomic, brain imaging data and so on. Increasingly, collaborators attempt to identify sets of minimal metadata which should be published to judge study comparability. Specifically for animal data, a big development came with the propagation of what is now the most used metadata standard, the Animal Research: Reporting of *In Vivo* Experiments guideline (ARRIVE). The current **ARRIVE guidelines 2.0** are a checklist containing metadata on animal experiments, including non-negotiable information on study metadata as well as a recommended set on additional study design parameters. It is also possible to optionally expand this standard based on additional important experimental variables, such as Minimum Information about Animal Toxicology Experiments (MIATE), Minimum Information for Publication of Experimental Pathology Data (MINPEPA) (Scudamore et al., 2016) or more specific variables like neuroimaging data Brain Imaging Data Structure (BIDS), though the uptake of these is very variable. More recently, Moresis et al. (2024) also describe a **minimal Metadata Set (MNMS)** which develops on ARRIVE and makes animal study data more suitable for reuse. In practice, animal research data is multi-modal, for

instance pathology, laboratory analyses and imaging datasets are combined in a study – there exist ways which these can be stored in a FAIR and interoperable way, possibly combining multiple modalities, and thus, standards (Kalantari et al., 2023).

Animal data repositories

Beyond metadata, in limited cases, raw animal data is also available through open access, which can potentially be used in VCG applications. There are many **data repositories enabling data reuse and sharing**, which are classified into generalist, specialist, or institutional. Data repositories which contain significant amounts of animal data are typically specialized, or domain-specific. While generalist, or cross-discipline, repositories impose less strict (field-specific) guidelines; domain-specific databases entail stricter standards, which is optimal for data reuse in the same field. Increasingly, journals may recommend or require publication into an open source repository, though there is no single enforced standard across repositories (FASEB; re3data.org; FAIRsharing, DataCite Commons; Springer Nature). Repositories frequently combine human and animal data, which also enables translational analysis but makes it less specific to animal research.

Neurology studies are characterised by an impressive amount of data standardisation, sharing and interlinking. The International Neuroinformatics Coordinating Facility (INCF) is a world-wide organisation which maintains FAIR metadata standards and practices, and hosts a number of platforms which contain animal data, particularly in the field of neuroimaging. This includes the Neuro-Imaging Platform, the Visiome Platform, Dynamic Brain Platform and Cerebellar Platform. Further, EBRAINS Knowledge Graph (KG) contains neuroscience imaging data for a variety of experimental approaches and imaging techniques for different animal species. It uses the openMINDS metadata model, a community-based schema for representing the metadata of neuroscience datasets, models and software, and the use of URIs enhances FAIRness. OpenNeuro (formerly OpenfMRI) is a public platform for sharing neuroimaging, electroencephalograms and other types of data. It utilises Brain Imaging Data Structure (Ioanas et al., 2020) which may enable to find similar brain imaging set-ups to incorporate into concurrent study data analysis, however the animal data there is sparse. MRI datasets specifically for non-human primates can also be found in the PRIME Data Exchange platform. NeuroMorpho.org provides data from multiple animal species of digitally reconstructed neurons and glia. MouseBytes+ is also a FAIR-compliant rodent cognitive behavioural repository specific to touchscreen-based cognitive test data. Although this is not an exhaustive list of animal data repositories which may contain relevant historical control data, the application of VCGs has not been explored. This is not only due to the novelty of the concept, but also due to protocol heterogeneity, too few similar datasets to match concurrent data, variability and gaps in metadata reporting between certain platforms, and lack of field-specific VCG implementation guidelines which still hinder data reuse for virtual control applications. In a practical sense, the data in the mentioned repositories is insufficient for most practical academic applications, though as the datasets continue to grow, there may be more overlap in the future.

Case studies of animal data reuse

In **academia**, data interoperability reuse are approaches which require **commitment and collaboration**. As a case study, the open data commons for spinal cord injury (SCI) named ODC-SCI is a great example of this, with 170 collaborating laboratories and 409 datasets from human, minipig and predominately rodent studies shared in accordance with FAIR principles. The ODC-SCI has developed a list of minimal data elements, referred to as Community-based Data Elements (CoDEs),

which should be included in submitted studies. This is comparable to the SEND required variables like subject ID, sex, strain, but also includes domain-specific information which is crucial like spinal cord injury type and injury level. Sheoran et al. (2025) assessed the effect of CoDEs implementation on data standardisation and quality in ODC-SCI, which showed that data completeness and interoperability in the database could improve, but curation was either way required since inconsistencies in data quality. This included variations of formats, units, human errors and semantic differences, such as for injury type: some researchers referred to contusion injury in a specific spinal area while others would refer to a specific vertebra. Inconsistencies are almost inevitable when merging diverse datasets, but this work serves as an excellent model for planning data sharing and integration strategies in biomedical research. In the field of SCI, the recovery of historical data to support current knowledge has also been implemented to assess present hypotheses. The FAIRification of historical data reported by Almeida et al. (2022) included the digitization and submission to ODC-SCI of three thoracic SCI studies on with data from 1125 rats to address the question of whether perioperative blood pressure influences outcomes.

5.3.4. VCGs in academic research: perspectives and conclusions

In **conclusion**, animal data publishing using FAIR standards and using RDF and ontology frameworks could form the basis for optimal data interoperability and sharing, but the implementation of this varies across fields. Many challenges remain, like the complexity and heterogeneity of protocols. Though FAIR data and metadata standards, as well as specific tools like ontology linking are available for researchers, the barrier to entry is still quite high, and there is large variability in schemas across domains. Animal data repositories containing control animal data exist, and when data is considered sufficient, they may provide a way forward for incorporating historical control data into academic VCGs. SEND is a data format which may be pertinent for regulatory settings, but introducing VCGs in academia will require care and stakeholder engagement in specific subfields. Due to high data requirements versus availability, it still depends on the stakeholders to manage and share data to create VCGs suitable for their purpose. Academic VCGs would have to be domain-specific and how they would have to incorporate existing available ontologies.

Certain data types, like omics data are very variable based on technical conditions with batch effects and time mismatching, and therefore much less likely to be reusable in the context of VCGs, though historical control data could still be important in this context to understand baseline biological activities.

Data formatting and sharing for unknown future purposes is extremely time consuming and there is a barrier to entry for researchers who are specialised in their own fields. There is a need for educating researchers on FAIR literacy (Gonzalez Soltero et al., 2024). Information of virtual controls and pipelines to creating them should be shared to academics, so that stakeholders with an interest in this vision can have grassroots initiatives in their own field of interest.

Despite this, successful reuse of historical control data in cases like the ODE-SCI database as well as the efforts of the VICT3R consortium to establish VCGs in the context of certain neurobehavioural studies reflects a growing momentum toward animal data reuse and underscores the expanding potential applications of VCGs in future studies.

5.4. Current Efforts and Future Directions in SENDification of Academic Data

Current efforts in SENDification of academic data focus on bridging the gap between unstructured or semi-structured research outputs and regulatory-compliant formats, highlighting challenges in identifying, standardizing, and mapping study variables and ensuring metadata completeness. Current experimental workflows combine expert-guided data processing, synthetic dataset testing, and emerging AI tools, such as LLMs, to support the systematic standardization and mapping of study variables.

5.4.1. Advancements and opinions about the generation of SEND reports from academic works using the NV domain

The generation of **SEND-compliant datasets from academic data sources remains a challenging task** that has not been widely explored, as data collection in academic settings is typically not conducted with regulatory standards in mind.

Usually, source data does not include all the required and expected variables across the different SEND domains. Moreover, datasets from different institutions - or even from the same institution - may be organized differently, requiring the implementation of a tailored **Extract, Transform, Load (ETL) tool** for the SENDification of each specific dataset format

Another crucial aspect to consider is **the lack of metadata** describing the meaning and relationships of the variables present in the source data. In some cases, this metadata is incomplete, which hinders the SENDification of variables that are not unambiguously defined. Ultimately, the design and implementation of each ETL tool must be conducted by a team that includes both domain experts familiar with the source data and specialists in SEND specifications. Alternatively, **LLMs can be considered a potential tool** to support the SENDification process. Preliminary attempts in the context of the VICT3R consortium have shown promising results. However, the lack of contextual information in tabular data remains a major challenge. To be effective, LLMs must be guided by a comprehensive set of prompts that describe the key aspects of the standardization process for a given source format. Moreover, an iterative validation phase of the resulting mappings would still be required.

Nervous system study reports have been identified as a relevant data source for the development of the VCG within the VICT3R consortium. To date, no official NV domain is included in the SEND Implementation Guide (SENDig). Although the use of alternative domains within the SENDig may serve as a viable solution for mapping this type of data into the SEND format, a preliminary, non-official version of the NV domain has been used for beta testing the SENDification of nervous system datasets.

A first SENDification attempt was performed on synthetic datasets derived from Open Field tests conducted at UPF. The team in charge of the design of the ETL tool consisted of one expert in the dataset and one expert in ETL implementations, both of them with sufficient knowledge of the SEND specification. Two SEND domains, demographics (DM), and NV were the initial targets in this exercise. Statistics on the number of SEND variables mapped per domain with respect the total number of variables split per variable core type are shown in Table 3.

Table 3. SEND variables mapped on synthetic datasets.

DOMAIN	Variable core type					
	Required		Expected		Permisible	
	Total in SENDig	mapped	Total in SENDig	mapped	Total in SENDig	mapped
DM	7	7	3	2	10	5
NV	5	5	11	11	25	3

The number of Required and Expected variables in the DM and NV SEND domains mapped from this synthetic dataset is high, with 9 out of 10 and 16 out of 16 variables covered, respectively. Nevertheless, the coverage of Permissible variables is significantly lower, particularly in the NV domain. This synthetic dataset was created as an initial attempt to map academic study results to the SEND format. Its structure and content were iteratively refined until most of the core Required and Expected SEND variables were covered.

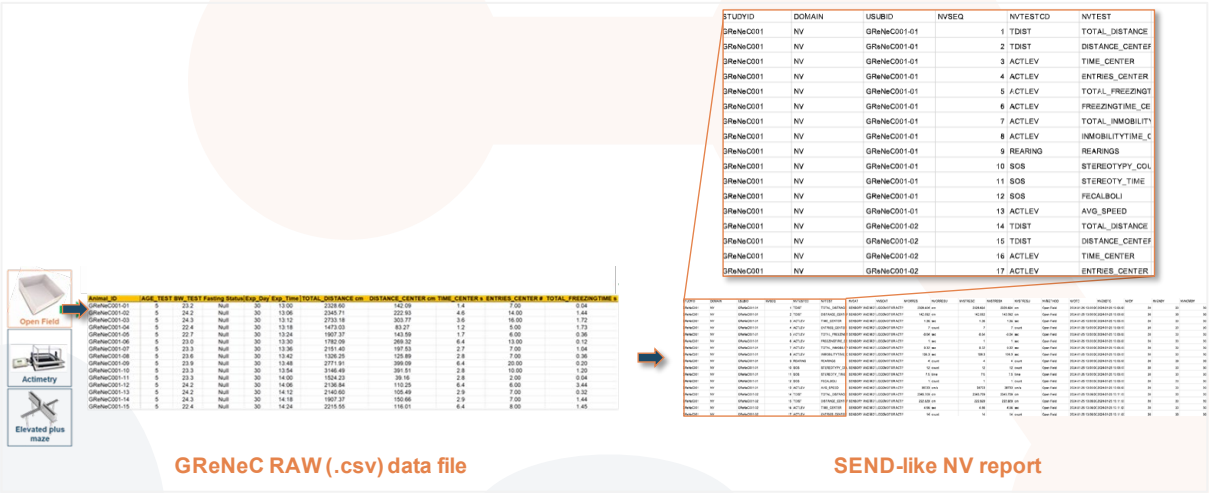


Figure 16. Illustrative process of SEND-like reports from academic Nervous System data files.

For real datasets, the inclusion of missing Required, Expected, and especially Permissible variables may require considerable effort, when feasible. In some cases, this information can be added to the datasets, but in general, it cannot be found or derived if it was not originally collected. In any case, implementing programmatic mapping routines requires an exhaustive description of the source dataset or extensive feedback if source metadata is not complete.

Currently, some datasets containing real data from academic Open Field studies have been shared to assess the feasibility of SENDification. After a preliminary review, the lack of metadata appears to be a significant obstacle that can only be overcome with feedback from the dataset providers.

Relevantly, the inclusion and mapping of variables relevant for key domains related to experimental design, such as Trial Summary (TS) and Trial Set (TX), is regarded as generally more challenging. These are rarely captured in a standardized or tabular format within real-world datasets that would allow

for direct conversion to the SEND format, unlike more structured endpoint-related variables (e.g., body weight). While they can be readily integrated into synthetic datasets by individuals with sufficient knowledge, these variables are often described in free text or illustrated through figures and non-standardized diagrams in academic contexts.

5.4.2. Advancements and opinions on NLP tools and the extraction of metadata, protocol taxonomy, and so on from published literature.

An experimental NLP tool is currently being developed to extract structured information from toxicological texts such as study plans and reports. This tool leverages LLMs to identify and organize relevant information within the TS (Trial Summary), TX (Trial Set), and DM (Demographics) domains. The approach aims **to facilitate the systematic capture of metadata and methodological details** from complex textual sources, supporting more efficient data organization, comparison, and reuse within toxicological research.

Building on this foundation, we are exploring **ways to extend and adapt this approach to the broader body of academic literature**. The long-term objective is to enable the automated extraction of key elements such as experimental designs, research objectives, and contextual metadata from scientific publications. As part of this effort, we are currently evaluating which types of published literature would provide the most relevant and representative test cases for this new approach. By doing so, we aim to enhance the discoverability, transparency, and interoperability of research data, ultimately contributing to the creation of more connected and machine-readable scientific ecosystems.

5.4.3. Advancements and opinions on data curation platforms for academic studies.

Following the previous section, which focuses on the extraction of structured information from toxicological and academic texts, the next step in this experimental workflow involves **the SENDification process**. After information extraction, **the tool performs a SENDification step, aligning the extracted content with the SEND terminology and structural conventions**. This enables the transformation of unstructured toxicological or experimental data into a format that is more interoperable and consistent with established regulatory data models.

It is important to note that **this workflow is independent of the standard curation and harmonization procedures developed within WP2 and remains outside the current consortium workflow**. Instead, it serves as an experimental initiative aimed at exploring the potential of emerging artificial intelligence technologies for data curation. The system **combines large language models (LLMs) with Retrieval-Augmented Generation (RAG) techniques** to enhance contextual understanding and improve the accuracy of aligning extracted information with controlled vocabularies and domain-specific structures.

At present, the tool is in an experimental evaluation phase, during which multiple LLM configurations and prompting strategies are being tested to assess their performance in terms of accuracy, scalability,

and adaptability across diverse study types. The insights gained from this phase will guide potential future integrations of AI-driven processes into academic data curation platforms, with the aim of complementing and strengthening existing manual and semi-automated curation workflows.

6. Conclusions

Key Findings from Academic Animal-Based Research

Analysis of academic animal-based research shows that certain study types are more frequent and standardized, making them particularly relevant for VCG development. **Nervous System studies dominate basic research, with Neurodegenerative Disease Models, Pharmacological, Toxicity, and Efficacy Studies, and Pain and Injury Experimental Models** representing the most implemented paradigms in the most animal-intensive areas of academia.

Pharmacological, genetic, and surgical interventions illustrate the breadth of manipulations typically employed, while behavioral assessments of anxiety, pain, and locomotion remain central. These are supported by histological, imaging, electrophysiological, and molecular analyses, reflecting the integration of behavioral, physiological, and molecular dimensions in experimental design.

Perceived Usefulness and Standardization for VCGs

Expert survey results indicate that **VCGs are perceived as moderately useful across academic research**, with 40–45% of respondents rating them positively. Neurological and behavioural paradigms, including pain, neurodegenerative, addiction, and classical learning tasks, emerged as the most consistently suitable for VCG applications. In contrast, oncology and some genetic models showed greater uncertainty or divergence.

Most responses clustered around recurring domains: behavioral/cognitive/addiction paradigms, neurodegenerative/CNS models, pain and toxicity assays, pharmacology/molecular/immunology approaches, and tumor/oncology models. Despite variability in manipulations and secondary endpoints, this suggests that core study types provide a feasible foundation for VCG implementation.

Diversity, Barriers, and Opportunities

While academic study designs exhibit diversity, they also cluster around common and relatively standardized paradigms, particularly in neuroscience. Well-established experimental frameworks with robust data infrastructures and archiving practices facilitate data reuse and VCG generation. Conversely, less standardized approaches present challenges in terms of data consistency, resource availability, and reproducibility. Importantly, these more variable areas may benefit most from historical datasets and interoperable platforms, enabling not only reuse but also better understanding of experimental variability. Practical implementation is needed to assess the full potential of VCGs in academia.

Next Steps for VCG Development: Selection of Study/Data Types

VCG development should prioritize **study types with the highest potential for standardization** and reusability, such as **Nervous System studies, Neuropsychiatric and Neurodegenerative Disease Models, and pain or pharmacological paradigms**. Efforts should focus on datasets with sufficient metadata and structured variables while addressing gaps in Required, Expected, and Permissible SEND variables. Iterative refinement and validation of synthetic and real datasets will be critical to ensure reliable VCG implementation.

Role of the VICT3R Project

VICT3R can **accelerate VCG development by leveraging NLP and LLM-based tools** to extract **structured metadata, protocol taxonomy, and experimental design elements** from published literature in the identified research niches. Combined with **experimental SENDification workflows and data curation platforms**, these tools can standardize and harmonize heterogeneous datasets, improve interoperability, and support integration of historical control data. Experimental evaluation of **AI-assisted mapping strategies** will guide future automation, scalability, and practical implementation of VCGs in academic research.

7. Methods

7.1. Large Language Models for ALURES NTS Analysis

Traditional Natural Language Processing (NLP) techniques have predominantly relied on supervised learning, where algorithms are trained on large, annotated datasets to perform specific linguistic tasks such as named entity recognition, part-of-speech tagging, or syntactic parsing. This approach demands substantial manual effort in creating and maintaining domain-specific corpora, making it resource-intensive and often inflexible when adapting to new domains or evolving research questions. The introduction of transformer-based models, particularly those in the BERT family, marked a significant advancement. These models enabled transfer learning in NLP, allowing direct application to downstream tasks or fine-tuning on smaller, domain-specific datasets, thereby reducing the need for extensive manual annotation while improving task performance. The development and wide-scale availability of Large Language Models (LLMs) further revolutionized the field. LLMs are trained on massive, diverse corpora and can be used off-the-shelf for a broad range of applications, often requiring nothing more than prompt engineering or retrieval-augmented generation (RAG) to perform effectively. This flexibility allows LLMs to be deployed rapidly without the need for task-specific training, making them especially valuable for complex or time-sensitive projects. However, these models come with limitations, such as potential biases in training data, a lack of transparency in the datasets used for proprietary models, and the risk of hallucinating information. To mitigate these challenges and enhance capability, agentic systems have emerged—multi-agent frameworks that coordinate LLMs to tackle complex tasks. These systems can decompose sophisticated scientific queries into manageable subtasks, execute them sequentially, and synthesize comprehensive answers. They also offer the potential for continual learning, evolving over time to better handle the dynamic nature of scientific literature. Together, LLMs and agentic systems represent a transformative shift in the ability to extract, process, and understand complex information from text at scale.

In this context, we have developed a solution utilizing LLMs to extract study types and methods from NTS records. By leveraging the power of these models, our solution can efficiently analyze the diverse linguistic formats and structures found in the NTS records, enabling more accurate and consistent extraction of relevant data. This approach not only improves the efficiency of data extraction but also facilitates better understanding and comparison of scientific projects across different domains.

The extraction phase

For the extraction of study types and methods from NTS files, our system specifically utilizes information from several key columns. These include “Objectives of the project”, which provides the overarching aims of the study; “In what procedures will the animals typically be used”, detailing the experimental procedures; “Expected impacts/adverse effects on the animals” and “Expected Harms”, which describe the anticipated consequences for the animals; and “Reasons for the planned fate of the animals after the procedure”, indicating the intended post-experimental handling. By focusing on these fields, the extraction process captures the critical elements necessary to understand and categorize animal-based studies accurately.

To extract the methods that were used in animal-based studies, we employed a multi-agent debate system. This system consists of three specialized agents implemented using Google's Gemini-1.5-flash model:

- **Affirmative Agent**: Proposes experimental methods based on text analysis
- **Negative Agent**: Provides alternative perspectives and critical evaluation
- **Judge Agent**: Evaluates arguments and determines the final method identified

Figure 4 illustrates the multi-agent debate system architecture for the extraction of methods and procedures.

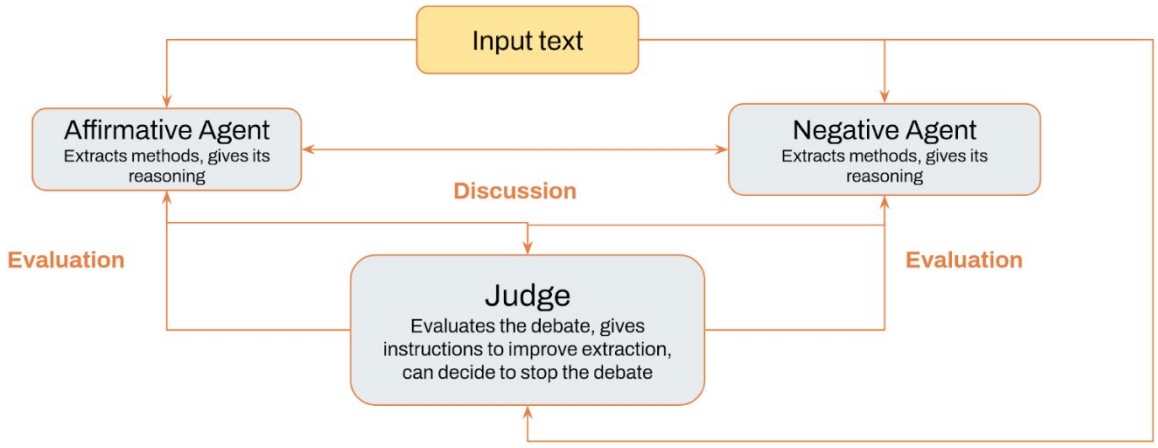


Figure 17. The multi-agent debate system architecture is used to extract relevant information from the NTS files.

The debate follows a structured format:

Round 1:

- An Affirmative agent proposes initial methods.
- A Negative agent provides independent analysis.
- The Judge evaluates both perspectives.

Subsequent Rounds:

- Agents refine their analyses based on judge feedback. They discuss each other's arguments.

- Process continues until the Judge decides to stop the debate or a maximum number of rounds is reached.

Agent prompts are structured to enforce a consistent output format that consists of a list of experimental methods separated by semicolon. The judge is responsible for the final answer and the answer language. Since the dataset has 22 languages, the prompt included specific instructions about the language required for the final answer (English).

Example of debating agent's prompt:

> *You are the debater. You participate in the Debate whose goal is to identify the experimental method used on the animals from this text. Please provide your answer as a list of methods and your reasoning under the field 'Reasoning'. Output format: experimental method 1: reasoning; experimental method 2: reasoning; experimental method 3: reasoning;... Final answer: [experimental method 1; experimental method 2; experimental method 3; ...]. You don't have to agree with your opponent.*

Example of Judge's prompt:

> *You are the judge of the Debate whose goal is to identify the experimental method used on the animals from this text. Decide which one of the debaters gave the better and more correct answer. Ensure that the experimental method(s) is returned in English. If debaters return answers in other languages, translate them into English. Provide your answer as a list of methods and your reasoning under the field 'Reasoning'. If the debaters return similar answers or there are no more relevant experimental methods to extract, please, stop the debate.*

We defined a maximum of 3 rounds of debate with the option for the Judge to stop the debate earlier if he believes the answer is already completed or the agents have agreed in their answers. Computational efficiency is further enhanced through parallel processing for concurrent debate sessions.

Categorization

The first step in organizing the extracted terms is to categorize them into five distinct categories defined by experts: 1. Experimental Paradigms/Frameworks, which includes models, paradigms, or protocols that define the core structure of the study (e.g., "tumor models" or "experimental autoimmune encephalomyelitis"), excluding specific manipulations or measurements; 2. Manipulations / Procedures, which encompasses specific interventions or independent variables introduced to influence outcomes (such as "high fat diet," "stress induction," or "dietary manipulation"); 3. Tests / Apparatus, which refers to experimental assessments or tools used to measure responses (e.g., "object recognition test" or "fear conditioning"); and 4. Data Acquisition Methods, which covers the modalities or technologies used to collect data (such as "telemetry," "EEG," or "actimetry").

This categorization was performed using the OpenAI GPT-4.1 model through prompt engineering, guiding the model to classify each term accurately according to the expert-defined categories. This custom, expert-driven categorization is essential because it allows a first level of organization for the extracted methods.

The Clustering phase

Building on the categorized structure obtained in the previous step, the clustering phase further organizes terms within each expert-defined category. This ensures that semantic grouping respects

the initial classification, avoiding cross-category conflation and preserving domain-specific distinctions.

In this context, clustering refers to grouping related terms into coherent hierarchies based on semantic similarity, producing structured representations that support analysis and interpretation. This step addresses two key objectives:

1. Aggregation and prioritization: Organizing extracted terms into analyzable hierarchies makes it easier to identify the most relevant methods within each category of animal studies.
2. Normalization of semantic redundancy: Lexical variation and paraphrasing in source texts—e.g., "euthanasia" versus "sacrifice"—are consolidated into unified concepts, reflecting the true conceptual structure rather than superficial wording differences.

Clustering Methodology

- **Embedding Generation**: Each method description within a category is converted into a high-dimensional vector using OpenAI's text-embedding-3-large model. This encodes semantic meaning, ensuring conceptually similar terms are positioned closely in vector space.
- **Distance Matrix Construction**: Pairwise cosine similarities are computed between all embeddings within the category, generating a distance matrix that reflects semantic proximity.
- **Hierarchical Agglomerative Clustering**: Terms are progressively merged into clusters using average linkage, with clustering performed separately within each category to maintain alignment with the expert-defined classification.
- **Hierarchical Levels:** Clustering is applied across multiple hierarchical levels—clusters, subclusters, and sub-subclusters—allowing both fine-grained and broader thematic structures to emerge. Several similarity thresholds were tested at each level, and expert review selected the final thresholds: 0.6 for first-level clusters, 0.5 for second-level clusters, and 0.4 for third-level clusters.
- **Label Generation:** Each cluster and subcluster is assigned to concise, human-readable label generated by a lightweight LLM (gpt-4o-mini). Labels are synthesized from the most representative terms within the cluster, bridging algorithmic structure with human interpretability.

By performing clustering separately for each prior category, the method ensures that the resulting hierarchies remain semantically coherent within their domain context. This integrated approach, categorization followed by within-category clustering, enhances both the interpretability and the analytical utility of the extracted methods.

Expert Curation and Interactive Tree

Following categorization and clustering, the resulting hierarchies are refined through manual expert curation to produce the final, coherent taxonomy of experimental methods. Expert curation is a critical step, ensuring that the automatically generated clusters accurately reflect the conceptual structure of the data and maintain domain-specific validity.

This process is carried out using the ALURES NTS Experimental Methods Browser, an interactive, browser-based tool implemented with the jsTree JavaScript library. The tree displays the taxonomy as a hierarchical structure, including groups, clusters, subclusters, sub-subclusters, and individual methods. Within this interface, experts review cluster contents, adjust hierarchical relationships, merge or split clusters where appropriate, and ensure that all labels are meaningful and representative of the underlying methods. Manual validation guarantees that subtle semantic nuances, ambiguous terms, or context-dependent methods are correctly interpreted, something that automated clustering alone cannot fully achieve.

7.2. Expert Academic Survey

The survey was conducted using the **Qualtrics XM platform**, which enabled diverse question types, including multiple choice, matrix tables, sliders, rank order, and open-text responses, and supported complex logic such as conditional branching. Qualtrics ensured secure data collection, real-time monitoring, and easy export of anonymized data for analysis. This flexible platform helped engage a broad international audience and obtain comprehensive, high-quality responses.

The questionnaire included **30 questions** in various formats to capture detailed information. These formats comprised open-ended text, single-text entries, Likert-scale matrix tables, sliders, multiple-choice (single or multiple selections), and rank-order questions. Informative text and graphics provided context without requiring responses. This mixed-format design balanced thorough data collection with respondent engagement and clarity. The questionnaire was structured into six thematic sections to systematically address key dimensions of this topic.

A full report of the questions and their rationale follows, providing detailed insight into the survey design and objectives.

Section 1: General Information

The first section of the survey collects **essential background details from each respondent**. In question 1 (Q1) we asked for their name (optional), main institution, department, primary research group, and professional role. The name field is optional to respect respondents' privacy and encourage honest, open participation. Allowing anonymity helps reduce potential bias or hesitation when providing candid feedback, particularly on sensitive topics such as data sharing and research practices. This background information helps contextualize survey responses by identifying the organizational and research environments in which participants operate. Understanding these aspects enables more accurate interpretation of the data, ensures representation across diverse institutions and disciplines, and supports tailored analysis of perspectives based on respondents' roles and affiliations within the scientific community.

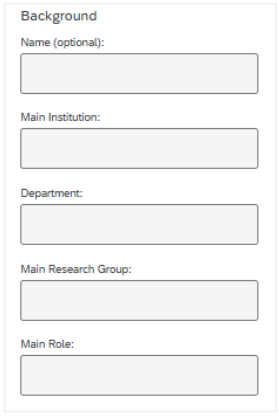


Figure 18. Question 1: Background information

Q2 asks whether respondents routinely conduct in vivo animal research. Allowing a "No" response acknowledges that some participants may be involved in related scientific fields—such as data analysis, computational modeling, or translational research—that do not directly involve animal experiments but are still relevant to the VCG concept and 3Rs objectives.

Q4 and Q5 were designed to collect structured information on the types of in vivo studies conducted across basic research (Q4) and translational research (Q5). Both questions provided a comprehensive list of study type categories and allowed for multiple selections to reflect the complexity and overlap in research activities. Responses provided an overall view of the expertise and research focus of the participant sample, helping to contextualize subsequent responses related to feasibility, data availability, and VCG alignment.

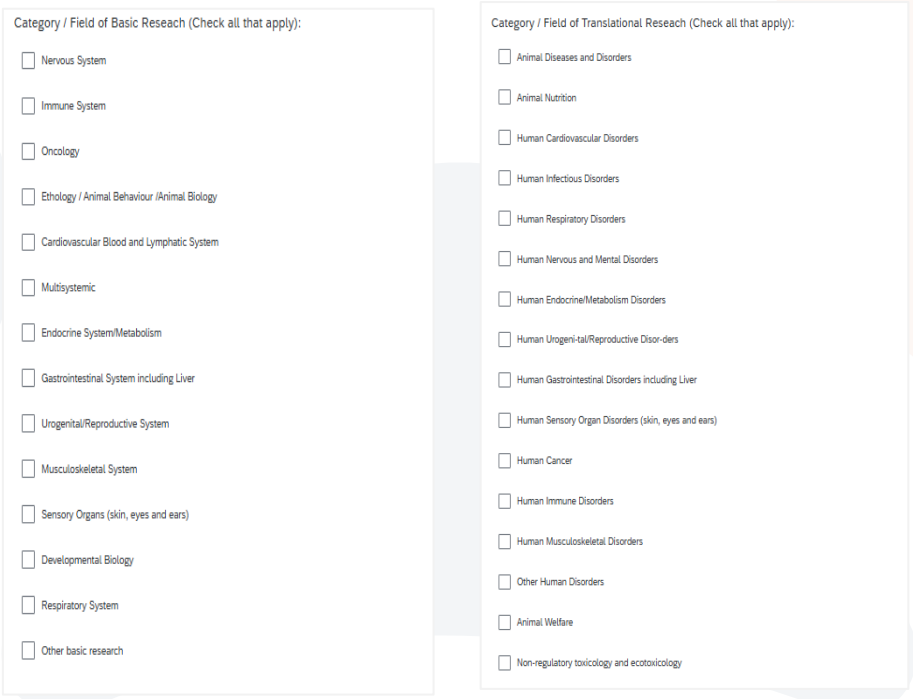


Figure 19. Questions 4 and 5: types of in vivo studies conducted

Section 2: Animal Use Characteristics in Your Work

Q6 is one of the most critical components of the survey, as it asks respondents **to specify the experimental paradigms they most frequently employ in their research**. Rather than focusing on broad study categories, this question targets specific experimental setups or basic procedures that are routinely used.

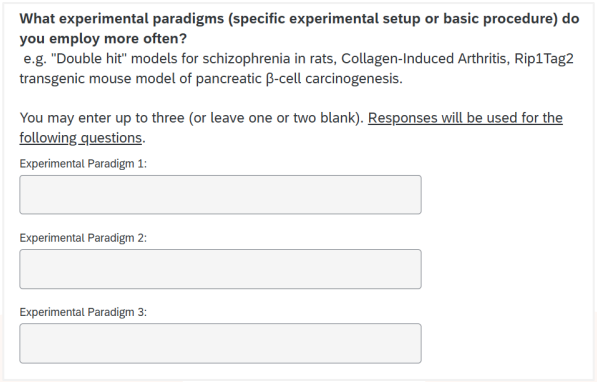


Figure 20. Question 6: Experimental paradigms most frequently employed

Respondents were asked to list up to three paradigms, allowing for focused and relevant input while avoiding excessive variability across responses. The information was collected in free-text format to capture discipline-specific detail. While enumerating all individual paradigms across respondents is not feasible due to the diversity and specificity of practices, this level of granularity is essential for identifying concrete opportunities where VCG strategies could be piloted or scaled.

In addition, it provides a valuable foundation for cross-referencing with existing historical datasets and repositories, supporting an evidence-based assessment of VCG feasibility at the level of specific experimental paradigms.

Q7 [“On average, how many animals are used per control group in each experiment?”] and Q8 [“On average, how many times do you repeat the experimental paradigm (with minor variations) over a two-year period?”] were designed to gather quantitative insights into typical experimental practices and the potential for generating reusable control data.

Section 3: Data Availability & Potential for Reuse in Your Work

The goal of this section was to assess both **the habitual practices around data retention and documentation, and the awareness or use of shared resources** that could support historical data reuse.

Q9 asked whether respondents keep detailed records of both experimental and control group data (either: “Yes, systematically archived”, “Partially, depending on the project” or “No”), providing insight into the availability of structured information suitable for retrospective analysis. Subsequently, Q10 explored the primary formats in which research data are archived (e.g., spreadsheets, lab notebooks, digital repositories), helping to gauge the level of digitalization and standardization. And Q11

addressed the retention period of research data, i.e., how far back researchers can typically trace their records, which is crucial for evaluating the potential lifespan of historical control data.

To help identify existing informal practices that align with the VCG approach, Q12 asked respondents whether they had ever reused control data across different experiments, studies, or projects. Response options included: "Yes", "Informally or for pilot studies", "Not yet, but I have considered it", and "No". This question aimed to capture both actual reuse practices and openness to the concept, providing valuable context on how closely current workflows may already resemble a VCG-compatible model.

Q13 then assessed whether respondents were aware of any existing shared historical datasets relevant to their line of research. This question provided insight into the data-sharing culture and the current infrastructure supporting access to historical data—factors that are essential for evaluating the feasibility of a VICT3R-led initiative to aggregate or build upon such datasets.

To ensure that all respondents had a shared understanding of the VCG concept before being asked to evaluate its applicability, Q14 was included as a brief, non-response item that introduced and explained the rationale, purpose, and potential benefits of VCGs. This served to establish a common informational baseline, regardless of participants' prior familiarity with the concept.

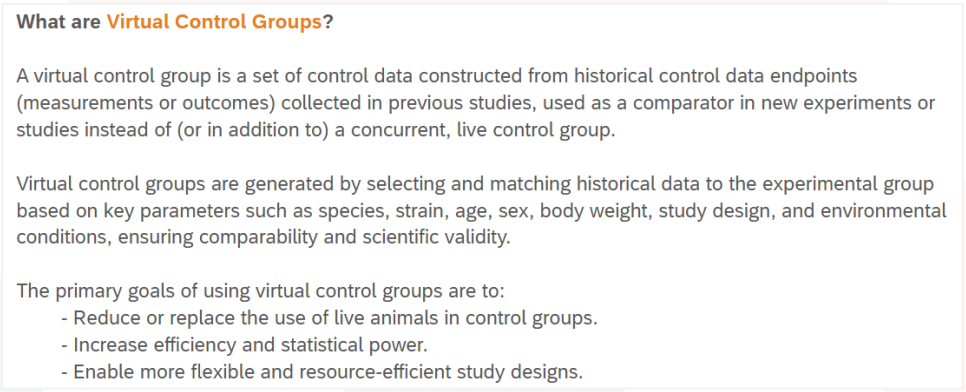


Figure 21. Information on the rationale, purpose, and potential benefits of VCGs.

This explanatory item outlined the idea of reusing historical control data, where appropriate, as a partial or full substitute for concurrent control groups in in vivo research, aiming to improve data efficiency and reduce animal use in line with 3Rs principles.

Immediately following this, Q5 asked respondents: "Do you think Virtual Control Groups would be useful for your main experimental paradigms?" This question served as an initial attitudinal checkpoint, capturing researchers' perceptions of VCGs' potential relevance and utility in their own work. It helped identify areas of openness or skepticism and contextualized responses in subsequent sections regarding feasibility, perceived barriers, and data needs.

Beyond the feasibility of VCG implementation, the survey explored the broader perceived value of curated historical datasets in research workflows. Q16 was used to explore these perspectives by

asking respondents, through a rank-order format, to prioritize various potential applications. These included benefits such as benchmarking, meta-analysis, methodological refinement, retrospective quality control, and hypothesis generation.

The aim was to identify which secondary uses of historical data are most valued by the scientific community. Understanding these broader applications serves multiple purposes: it helps justify investments in data curation and infrastructure by demonstrating their multifaceted value; it supports strategic communication about VCG-related initiatives to diverse stakeholders; and it informs the future development of the VICT3R project, particularly in relation to data governance, platform design, and user engagement strategies aligned with real-world research needs.

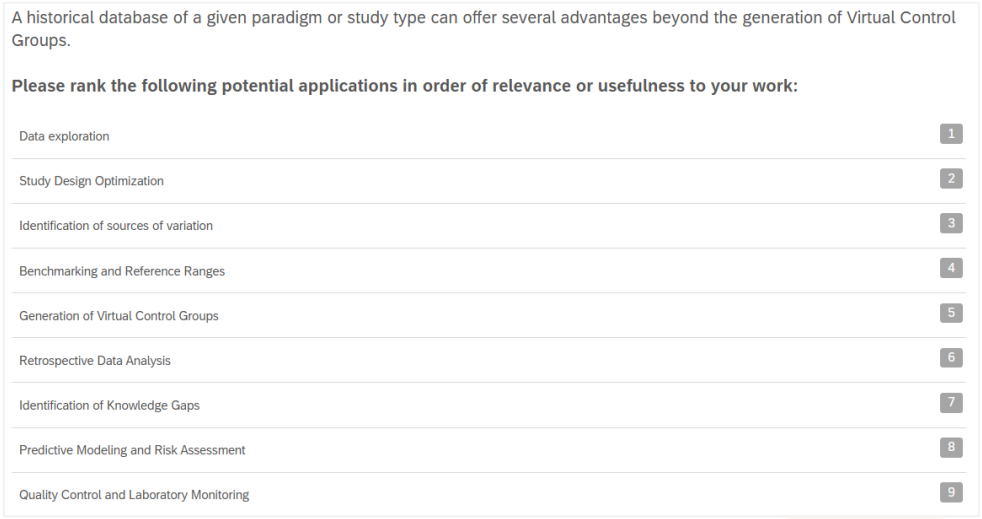


Figure 23. Question 16: Potential applications (shown in random order).

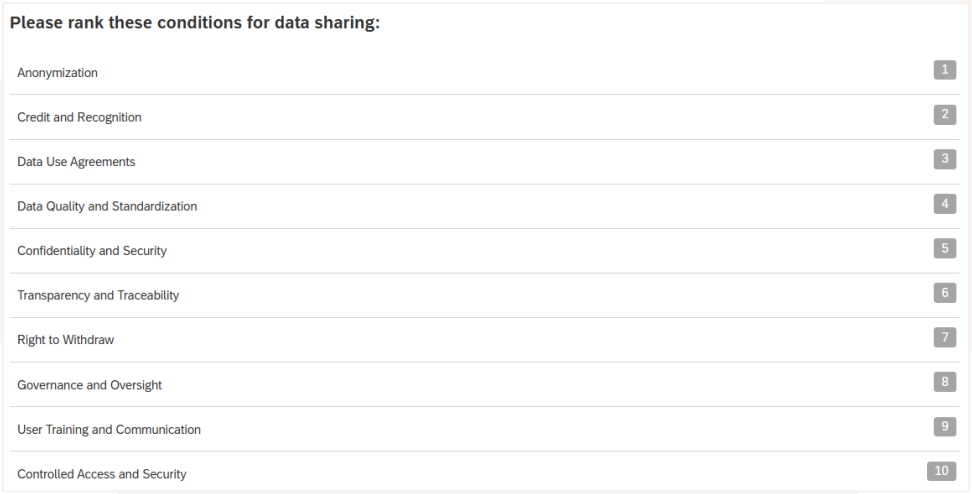


Figure 24. Question 18: Data-sharing requirements (items shown in random order).

To gauge the practical feasibility of creating a shared historical database for VCG development, the survey assessed **researchers' willingness to contribute past experimental data**, including under what conditions. This was the purpose of Question 17, which asked whether participants would share legacy data, offering response options such as Yes, No, or Possible, with conditions (e.g., anonymization,

credit, institutional approval). The question aimed to identify both support and key barriers to data sharing.

To further explore the conditions under which researchers would consider sharing their data, Q18 asked respondents to rank predefined data-sharing requirements, such as authorship recognition, institutional permissions, or data anonymization, in order of importance.

The use of predefined options aimed to reduce respondent burden and increase completion rates, while still capturing key priorities and concerns related to data contribution for VCG-compatible repositories.

Q19 and Q20 were included to capture any additional, unanticipated conditions for data sharing that were not covered by the predefined options in Q18. While Q18 presented common requirements, these open-ended questions allowed respondents to specify other concerns or criteria they consider important. This approach ensured a more comprehensive understanding of potential barriers or safeguards valued by the research community, providing valuable input to shape flexible and inclusive data governance policies.

Section 4: Feasibility & Perspective in Your Work

This section explored respondents' views **on the practical applicability of VCG** approaches within their own research and the potential barriers to implementation.

Q21 asked respondents to identify which aspects of their studies might be suitable for applying a Virtual Control Group (VCG) approach. Participants could select multiple options or leave the question blank if none apply. The options included factors that support VCG feasibility such as study design consistency, endpoint consistency, data standardization, and the existence of relevant datasets. This question aimed to pinpoint practical enablers within respondents' research that could facilitate successful VCG implementation.

To capture any additional perspectives beyond the predefined options, Q22 asked whether respondents foresaw other aspects of their studies that might support VCG inclusion, followed by Q23, which provided a free-text field for elaboration if "Yes" was selected.

Q24 queried respondents on perceived barriers to implementing or contributing data for VCGs, listing common challenges such as unstandardized data, institutional restrictions, limited time/resources, and concerns about reproducibility or variability. To identify further obstacles not covered by the checklist, Q25 asked if any other barriers might render their studies unsuitable, with Q26 offering a text field for detailed input if applicable.

Section 5: Feasibility & Perspective in Your Field

Q27, asked respondents **to identify up to three experimental paradigms that tend to show less variation in study design** within their broader research field.

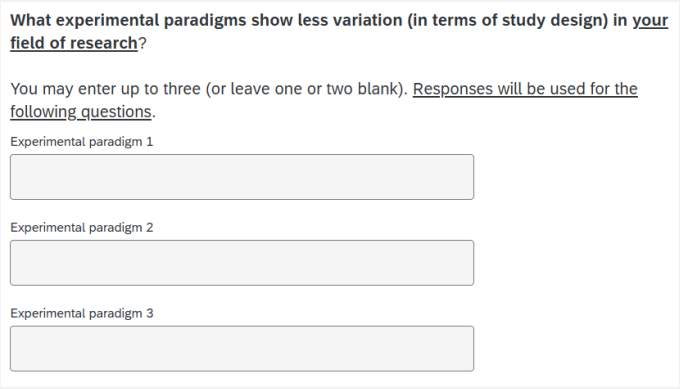


Figure 25. Question 27: Experimental paradigms with less variation within respondent research field.

As in Q6, the question allowed for up to three entries, giving respondents flexibility to provide focused and meaningful input. Importantly, the aim was not limited to capturing respondents' personal or direct expertise, but to leverage their professional knowledge of common practices and levels of standardization across their field.

Identifying paradigms that are more uniform in design and implementation is key to evaluating where VCG approaches may be more readily applied, as these offer greater potential for reliable data reuse and cross-study comparisons.

The entries in Q27 were also used to customize follow-up questions (Q28 and Q29), enabling a targeted assessment of VCG feasibility, acceptance, and potential limitations in relation to the specific experimental paradigms identified by each respondent. Q28 asked respondents to reflect on what reproducibility concerns might affect those paradigms, offering structured response options such as study design variability, endpoint consistency, and data standardization. This question aimed to identify potential barriers to reliable historical data reuse and VCG implementation within each model type. Q29 followed by asking whether respondents were aware of any historical data repositories relevant to the paradigms they listed. This served to assess the existing data infrastructure in each field and the practical feasibility of developing or applying VCG strategies.

Section 6: Open Comments

The final section of the survey included Q30, **an open-ended prompt inviting respondents to share any additional comments**, questions, or expressions of interest in further collaboration related to the initiative. This section provided space for participants to elaborate on earlier responses, raise issues not covered elsewhere in the questionnaire, or signal their willingness to engage in follow-up discussions. It served both as a qualitative feedback channel and as a way to identify potential contributors or stakeholders for future phases of the project.

Survey Data Analysis

The collected survey data were analyzed using the built-in tools of the Qualtrics XM platform. These tools facilitated data cleaning, visualization, and export for further statistical analysis, enabling a comprehensive evaluation of response patterns, trends, and key insights across thematic sections.

The analysis primarily focused on frequency distributions to assess the prevalence of specific responses across questions. Additional descriptive statistics, such as measures of central tendency and variability, were used where applicable. For scaled questions (e.g., Likert items and sliders), mean scores and response patterns were examined to identify trends and consensus. Rank-order data were analyzed to determine priority rankings among options. Open-text responses were qualitatively reviewed to extract key themes and insights.

8. References

1. Kaufman, L., Gore, K., & Zandee, J. C. (2016). Data Standardization, Pharmaceutical Drug Development, and the 3Rs. ILAR journal, 57(2), 109–119. https://doi.org/10.1093/ilar/ilw030
2. Carfagna, M. A., Anderson, J., Eley, C., Fukushima, T., Horvath, J., Houser, W., Larsen, B., Page, T., Russo, D., Sloan, C., Snyder, K., Thompson, R., Ullmann, G., & Whittaker, M. (2021). Leveraging the Value of CDISC SEND Data Sets for Cross-Study Analysis: Incidence of Microscopic Findings in Control Animals. Chemical research in toxicology, 34(2), 483–494. https://doi.org/10.1021/acs.chemrestox.0c00317
3. Masuya H. (2012). Roles and applications of biomedical ontologies in experimental animal science. Experimental animals, 61(4), 365–373. https://doi.org/10.1538/expanim.61.365
4. Foster, C., Wignall, J., Kovach, S., Choksi, N., Allen, D., Trgovcich, J., Rochester, J. R., Ceger, P., Daniel, A., Hamm, J., Truax, J., Blake, B., McIntyre, B., Sutherland, V., Stout, M. D., & Kleinstreuer, N. (2024). Standardizing Extracted Data Using Automated Application of Controlled Vocabularies. Environmental health perspectives, 132(2), 27006. https://doi.org/10.1289/EHP13215
5. Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: a functional perspective. Briefings in bioinformatics, 16(6), 1069–1080. https://doi.org/10.1093/bib/bbv011
6. Maguire, E., González-Beltrán, A., Whetzel, P. L., Sansone, S. A., & Rocca-Serra, P. (2013). OntoMaton: a bioportal powered ontology widget for Google Spreadsheets. *Bioinformatics (Oxford, England)*, *29*(4), 525–527. https://doi.org/10.1093/bioinformatics/bts718
7. Bernabé, C.H., Queralt-Rosinach, N., Silva Souza, V.E. *et al.* The use of foundational ontologies in biomedical research. *J Biomed Semant* **14**, 21 (2023). https://doi.org/10.1186/s13326-023-00300-z
8. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18
9. Kamdar, M.R., Musen, M.A. An empirical meta-analysis of the life sciences linked open data on the web. *Sci Data* **8**, 24 (2021). https://doi.org/10.1038/s41597-021-00797-y
10. Gonzalez Soltero, R., Pino García, D., Bellido, A. *et al.* FAIR data management: a framework for fostering data literacy in biomedical sciences education. *BMC Med Res Methodol* **24**, 284 (2024). https://doi.org/10.1186/s12874-024-02404-1
11. Scudamore, C. L., Soilleux, E. J., Karp, N. A., Smith, K., Poulsom, R., Herrington, C. S., Day, M. J., Brayton, C. F., Bolon, B., Whitelaw, B., White, E. S., Everitt, J. I., & Arends, M. J. (2016). Recommendations for minimum information for publication of experimental pathology data: MINPEPA guidelines. *The Journal of pathology*, *238*(2), 359–367. https://doi.org/10.1002/path.4642

12. Kalantari, A., Szczepanik, M., Heunis, S., Mönch, C., Hanke, M., Wachtler, T., & Aswendt, M. (2023). How to establish and maintain a multimodal animal research dataset using DataLad. *Scientific data*, *10*(1), 357. https://doi.org/10.1038/s41597-023-02242-8
13. FASEB DataWorks! Help Desk. (n.d.). *Finding a repository*. Retrieved October 17, 2025, from https://dataworks.faseb.org/helpdesk/kb/finding-a-repository
14. re3data.org – Registry of Research Data Repositories. (n.d.). Retrieved October 17, 2025, from https://www.re3data.org/
15. FAIRsharing. (n.d.). *FAIRsharing: Collections, standards, and policies for data and metadata*. Retrieved October 17, 2025, from https://fairsharing.org/
16. DataCite Commons. (n.d.). *Repositories*. Retrieved October 17, 2025, from https://commons.datacite.org/repositories
17. Springer Nature. (n.d.). *Biological sciences repositories — research data policy*. Retrieved October 17, 2025, from https://www.springernature.com/gp/authors/research-data-policy/biological-sciences-repositories/12327160
18. Ioanas, H. I., Marks, M., Garin, C. M., Dhenain, M., Yanik, M. F., & Rudin, M. (2020). An Automated Open-Source Workflow for Standards-Compliant Integration of Small Animal Magnetic Resonance Imaging Data. *Frontiers in neuroinformatics*, *14*, 5. https://doi.org/10.3389/fninf.2020.00005
19. Sheoran, A., Fond, K. A., Davis, L. M., Huie, J. R., Vavrek, R., Axtman, P. J., Lemmon, V., Bixby, J. L., Visser, U., Gensel, J. C., Fouad, K., Ferguson, A. R., Grethe, J. S., Bandrowski, A., Martone, M. E., & Torres-Espin, A. (2025). Data reporting quality and semantic interoperability increase with community-based data elements (CoDEs). Analysis of the open data commons for spinal cord injury (ODC-SCI). *Experimental neurology*, *385*, 115100. https://doi.org/10.1016/j.expneurol.2024.115100
20. Almeida, C. A., Torres-Espin, A., Huie, J. R., Sun, D., Noble-Haeusslein, L. J., Young, W., Beattie, M. S., Bresnahan, J. C., Nielson, J. L., & Ferguson, A. R. (2022). Excavating FAIR Data: the Case of the Multicenter Animal Spinal Cord Injury Study (MASCIS), Blood Pressure, and Neuro-Recovery. *Neuroinformatics*, *20*(1), 39–52. https://doi.org/10.1007/s12021-021-09512-z
21. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., & Tu, Z. (2024). *Encouraging divergent thinking in large language models through multi-agent debate.* In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17889–17904). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.992