Paper RE07

Building an integrated pre-clinical and clinical data platform to enable rapid translational data review

Brenda Finney, Ali Dootson Instem, Stone, UK

ABSTRACT

With the FDA announcement that it is no longer a requirement to test on animals prior to human trials; there is potential for a paradigm shift in the conduct of regulatory investigations. Using integrated databases containing public clinical and legacy preclinical data, and the opportunity to add anonymized clinical trial data we show how visualization and modelling tools could aid in planning testing programs. This presentation will look at examples on how a simple search in the integrated database can pull results from multiple preclinical and clinical studies from several disparate databases. We will show how event counts within the MedDRA system organ class (SOC) terms can overlap with pre-clinical SOC results across the databases and how the nature of these potentially translationally relevant findings was investigated by reviewing the results. Attendees can expect to learn how findings can be seen across humans, monkeys, rats and dogs and how this could impact earlier decision making, and reduce the need for as much animal testing. Finally, we will learn about how to integrate such data with a unifying ontology for terminology to enable an optimized review of the data.

INTRODUCTION

In part due to the steady increase in the quantity, quality and applicability of alternative methods using modeling or *in vitro* systems, the FDA has removed the requirement for animal testing[1] from its regulations. There is now space to rely on "nonclinical tests" which according to the act "means a test conducted *in vitro*, *in silico*, or *in chemico*, or a non-human *in vivo* test that occurs before or during the clinical trial phase of the investigation of the safety and effectiveness of a drug, and may include animal tests, or non-animal or human biology-based test methods, such as cell-based assays, microphysiological systems, or bioprinted or computer models"[1].

This is a monumental change in the outlook of an agency which over many years has helped shape the procession of research and development to approval for medical use into a structured and routine process moving from preclinical animal studies to clinical human trials. One area for improvement is the "valley of death" when moving from basic science to human studies[2]. There are many reasons for this high mortality in compound progression including:

- lack of communication of negative results and other granular data,
- time and resources required from discovery to licensing,
- clinical outcomes not reflective of preclinical outcomes[3].

Translational science attempts to bridge this valley and aid in the transfer of knowledge between historically siloed areas of scientific research. As such, it is well placed to support this regulatory shift. Indeed this broadening of acceptability for nonclinical tests presents an opportunity for the development of translational science tools to incorporate data from a wider array of sources to support an application as well as an opportunity to robustly address the underlying issue of translatability.

Supporting and motivating (at least in part) this shift to include additional methodologies is the 3Rs (Replacement, Reduction, and Refinement) for animal models in research. It is a general goal to support these Rs by the addition of different testing strategies and *in silico* modeling of potential outcomes. Part of this is using the data collected from animal models, including both legacy and ongoing research, as part of the training sets for models as well as to aid longitudinal data review by experts. With the paradigm shifting to produce less data from whole animal models, the data from those models becomes more valuable for analysis on a wider scale.

One of the main goals of testing and modelling is the prediction of clinical safety, and activities to support this are not the responsibility of any single department or stage of the testing process. In addition, the monitoring of the effects associated with the compound never really ends. Not only due to long-term modeling, the content type, scale and database availability for compound data is increasing at a rapid rate across multiple areas of research and development. Scientists now have access to more data than ever before, but it is located in disparate websites, repositories and different formats and the correlation of this information can be a labor intensive, computing intensive process. Bringing databases together into a unified platform for searching and review can save a significant amount of time for the scientist.

A resource which allows the pooling of these valuable preclinical animal results across multiple companies, where this data would traditionally be kept in-house, while protecting company Intellectual Property (IP) would enable this data to serve as a more robust resource for evaluation and modelling as well as analysis for translation of effects

across species and compound families. In a wider sense, having a system that can serve across preclinical and clinical research activities can help companies streamline their research activities as well as save money with a single subscription rather than multiple fees for different systems in different departments. When the system described in this paper was initially developed by the eTRANSAFE consortium, the goal was to produce a framework for reproducible knowledge sharing with linkage of clinical and preclinical data with input from pharmaceutical, academic and regulatory partners[4]. To ensure that the data could be used in a coherent manner, tools for ontology linkage and visualization were also incorporated into the system. These tools are critical to facilitate data analysis by toxicologists and data scientists alike for making decisions which could have the benefit of reducing animal usage, improving trial patient experience or impact the planning for portions of or whole testing programs.

INTEGRATED DATABASES AND DATA SOURCES

The types of data sources included currently in the Centrus platform include both public clinical and legacy preclinical data, specific data sources are shown in Figure 1. Due to the approach to data handling and cloud-based, containerized architecture of the system, the addition of data types, sources and applications is feasible without having to redesign the entire structure. One data type which would add a great deal of value to the analysis that can be conducted in the system is anonymized clinical trial data. Given the burgeoning activity in the area of clinical trial transparency and anonymization services, these data are becoming more widely produced and amenable to data sharing systems such as Centrus. Instem is prioritizing this capability as part of the development roadmap while activities to commercialize the system are ongoing.

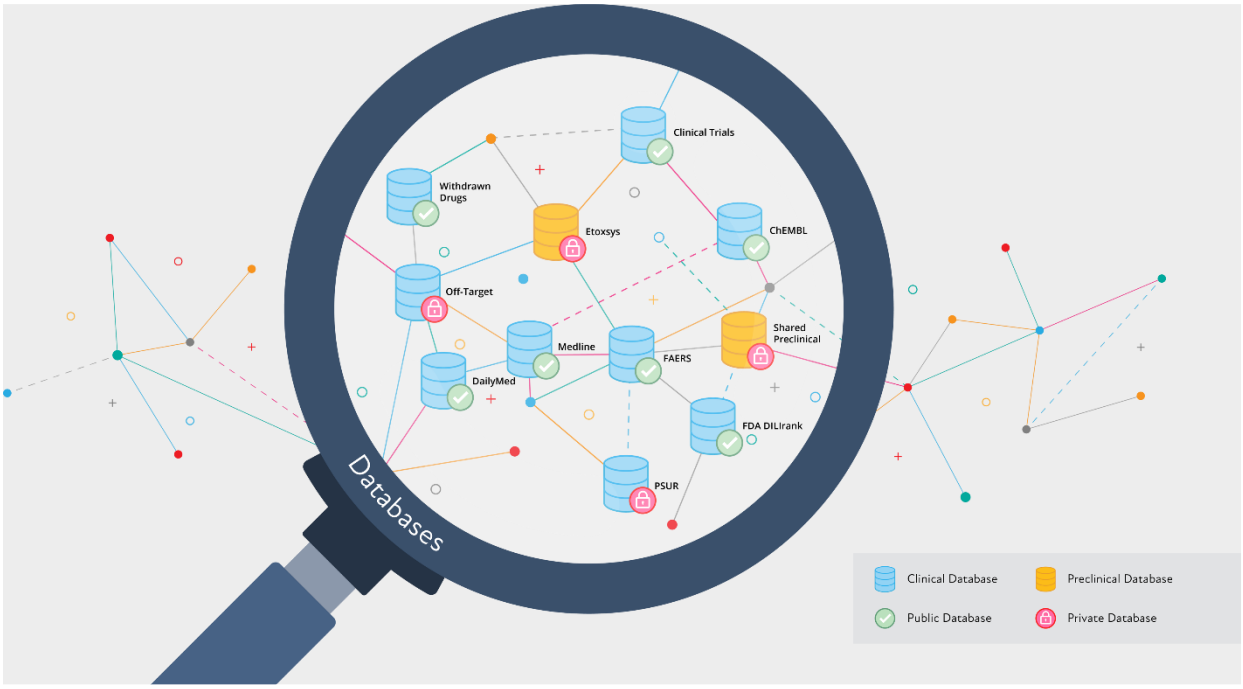


Figure 1. Databases available in the Centrus platform. Clinical databases are shown in blue and preclinical databases are shown in yellow. Publicly available databases are indicated with a green tick, and data sources unique to Centrus are indicated with a red lock.

One of the unique features of this platform is the capability to share preclinical data. Preclinical data has traditionally been held very closely within the confines of CROs and companies conducting the studies. The lack of this data in the public domain is due to several factors including; project attrition, intellectual property issues and protection of competitive advantages. This presents a challenge for teams working on tools to support risk assessment such as read-across and other *in silico* approaches[5] where large datasets can improve the accuracy of analysis. Changing the traditional reticence for sharing this type of data has been the goal of several consortia including both eTOX[5] and eTRANSAFE[6]. Companies involved in both of these initiatives have shared their data with this platform and subscribers can access this shared data for their reviews allowing a unique route for comparative analysis. To help encourage data sharing, further options were included such as partial-sharing of study data. In a partially-shared study, information on the compound structure, exact name and organization can be redacted protecting company IP,

but still allowing the study to be used for assessments of toxicological profiles or background incidences in control groups.

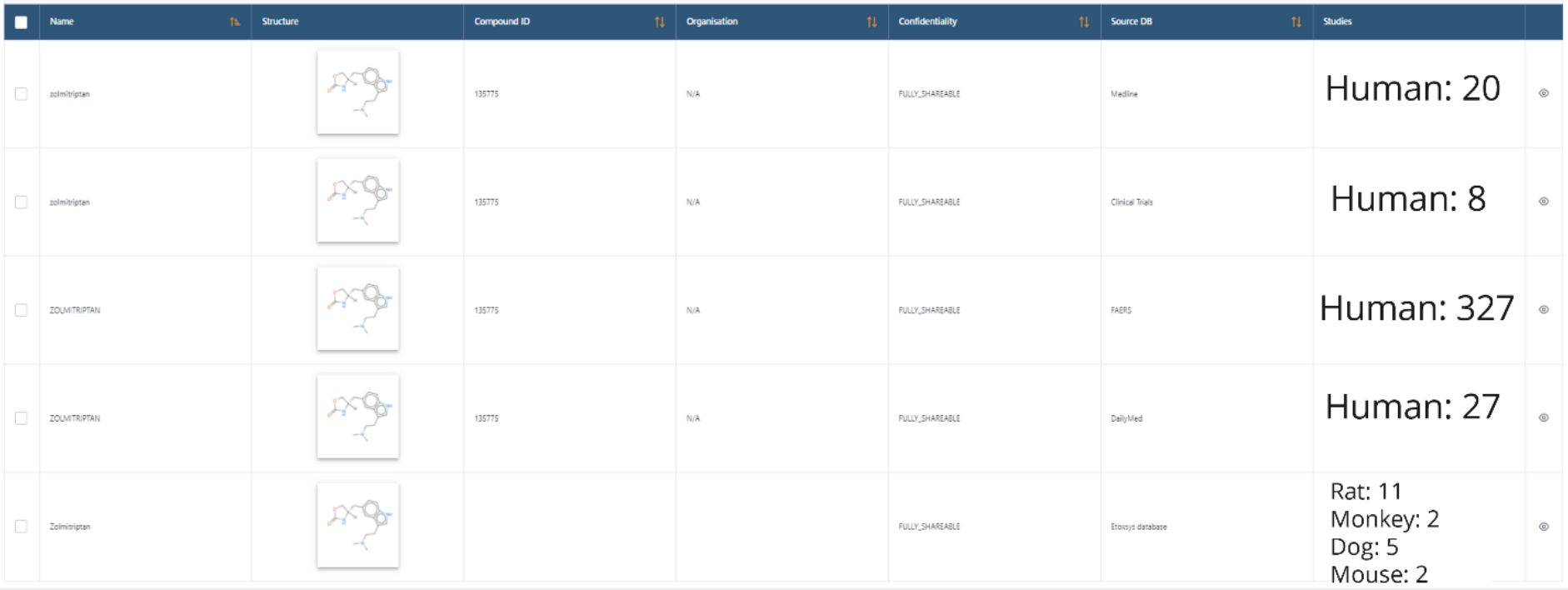
However, there are times when the external sharing of data across companies is not desirable, while it is desirable to share internally with different teams or departments. The cloud-based, containerized architecture of this system means that company-specific environments can be deployed to house non-shareable or private data while maintaining the links to the shared data pools for comparisons or analysis as needed. This capacity also means that environments can be spun up for mixed groups and projects, potentially enabling two (or more) different companies a shared space to safely house and review data related to a specific project or activity. When such a project is finished the environment can be closed down and data archived or deleted.

Another challenge is the production of data with different analysis protocols, machines, collection systems, etc. resulting in data in different formats with variations in vocabulary. The development of this system had to overcome all of these challenges. To address some of the differences inherent in the data as collected during a study, the Standard for Exchange of Nonclinical Data (SEND) format[7] was used to standardize the data prior to upload into the Preclinical Database (PCDB). In addition, the public databases were processed prior to incorporation into the system to allow use of a common data model for access and searching[10].

DATA INTEGRATION AND CROSS-DATABASE SEARCHES

Integral to this system is the capability to search multiple databases in a single search activity. Searches conducted via the Query application can proceed in a variety of ways using either compound information (any name or structure), study information (number, species), finding of interest (organ or specific event) or a combination of these factors. Strings of query characteristics can be put together to make your search as broad or as specific as is required. By allowing access to multiple databases via a single search interface a significant amount of time can be saved in conducting queries and assessing the level and content of the study data available.

Searches can also be simplified to a single aspect, as is the case with the example used in this paper – multiple databases were searched for a single term “zolmitriptan” (Figure 2) a selective 5-HT receptor agonist marketed as Zomig®. This example shows how results from both clinical and preclinical databases are displayed in the system post-search with the detail from each database and the ability to view the details of the results in the Query application. The details available include summary study information such as duration, dose route, species and organization as well as tabular views of events, findings and domain data.



Name	Structure	Compound ID	Organisation	Confidentiality	Source DB	Studies
zolmitriptan		135775	N/A	FULLY_SHAREABLE	Medline	Human: 20
zolmitriptan		135775	N/A	FULLY_SHAREABLE	Clinical Trials	Human: 8
ZOLMITRIPTAN		135775	N/A	FULLY_SHAREABLE	FAERS	Human: 327
ZOLMITRIPTAN		135775	N/A	FULLY_SHAREABLE	DailyMed	Human: 27
Zolmitriptan				FULLY_SHAREABLE	Etoxsys database	Rat: 11 Monkey: 2 Dog: 5 Mouse: 2

Figure 2. Query results from multiple databases. A single search of five separate databases using the compound name “zolmitriptan” resulted in multiple studies across five different species available from the system databases. The number of studies in the right column associated with species information are actual study number counts for the eTOX and ClinicalTrials.gov, the number of publications or labels found in Medline or DailyMed, or in the case of FAERS- a single pair of reports which reaches the threshold of at least 3 reported cases, a PRR of ≥2 and a chi-square of ≥4. Use of the “eye” icon to the right, takes the user to a page of the details of the results for the row selected.

INTEGRATING DATA WITH UNIFYING ONTOLOGIES

With the use of many different data sources as well as different species comes the challenge of bringing the information together in a shared ontology. For clarity, when speaking of ontologies we are referring to a set of concepts and categories in a subject area that shows their properties and the relationships between them. To facilitate this unification, terminologies must be mapped between the different systems to bring together equivalent terms, or link broader or narrower concepts together bridging the gap between different domains and disciplines. Data integrated with a common ontology are typically more scalable. As new data sources or elements are added, they can be mapped to the existing ontology, ensuring that the data remains consistent and usable as the data landscape evolves.

Integrating data with a common ontology provides a foundation for consistency, interoperability, and meaningful analysis. It standardizes terminology, promotes data quality, and simplifies the process of aggregating and interpreting data from various sources. This is particularly valuable in pharmacology, toxicology and safety science, where meaningful, evidence-based interpretations can be made quickly and confidently. With data following a common ontology, the process of data analysis becomes more efficient. Analysts do not need to decipher the meaning of each data element or spend time reconciling different naming conventions. They can focus on the analysis itself, drawing insights and making decisions based on a more reliable and consistent dataset. Data are not viewed in isolation but in the context of its relationships and dependencies. This contextual insight allows for more nuanced analysis and when multiple companies are working with data, using a common ontology facilitates communication and collaboration.

In the case of this system, terminology mapping has been conducted with the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) as an intermediary. SNOMED CT has been used as it is a rich ontology containing semantic relations. On either side of SNOMED CT there are preclinical and clinical terminologies which do not have well-connected semantic relationships. For preclinical data this includes SEND Controlled Terminology (CT)[7] and Histopath terminology (HPATH)[5] among others. On the clinical side the Medical Dictionary for Regulatory Activities (MedDRA)[8] terms are used. Each of these terminologies contain different levels of terms, but by utilizing the intermediary linkages can be made starting from either type of data. The following principles have been applied to all terminologies mapped by the system while there are also domain specific characteristics for mappings:

- There should not be more than one exact one to one mapping for each concept in each direction
- Mappings from the same subject with the same mapping_group constitute an 'and' mapping. Mappings with a different mapping_group are considered 'or' mappings
- Prefer a broad mapping for concepts that do not occur in humans, details on the term should be provided in the comments[9]

By applying this mapping approach data retrieval and review are efficient. With standardized terminology and a well-defined structure, it's easier to search for and locate specific data points with the addition of context to data, which is vital for making informed decisions. Figures 3 and 4 show two different views of data from the case study search for zolmitriptan where the MedDRA System Organ Class (SOC) has been used to group findings together from all of the databases used, allowing understanding of the scale of the data. This can be presented as either the number of studies as shown in Figure 3 or number of events reported as in Figure 4. These figures also demonstrate the importance of having data visualization capability to organize and represent the data flexibly to accommodate individual analyst’s preferences or to highlight specific data features. In the system these two views are presented side by side so the reviewer can choose their preferred view or use both the heatmap and graphing feature at the same time.

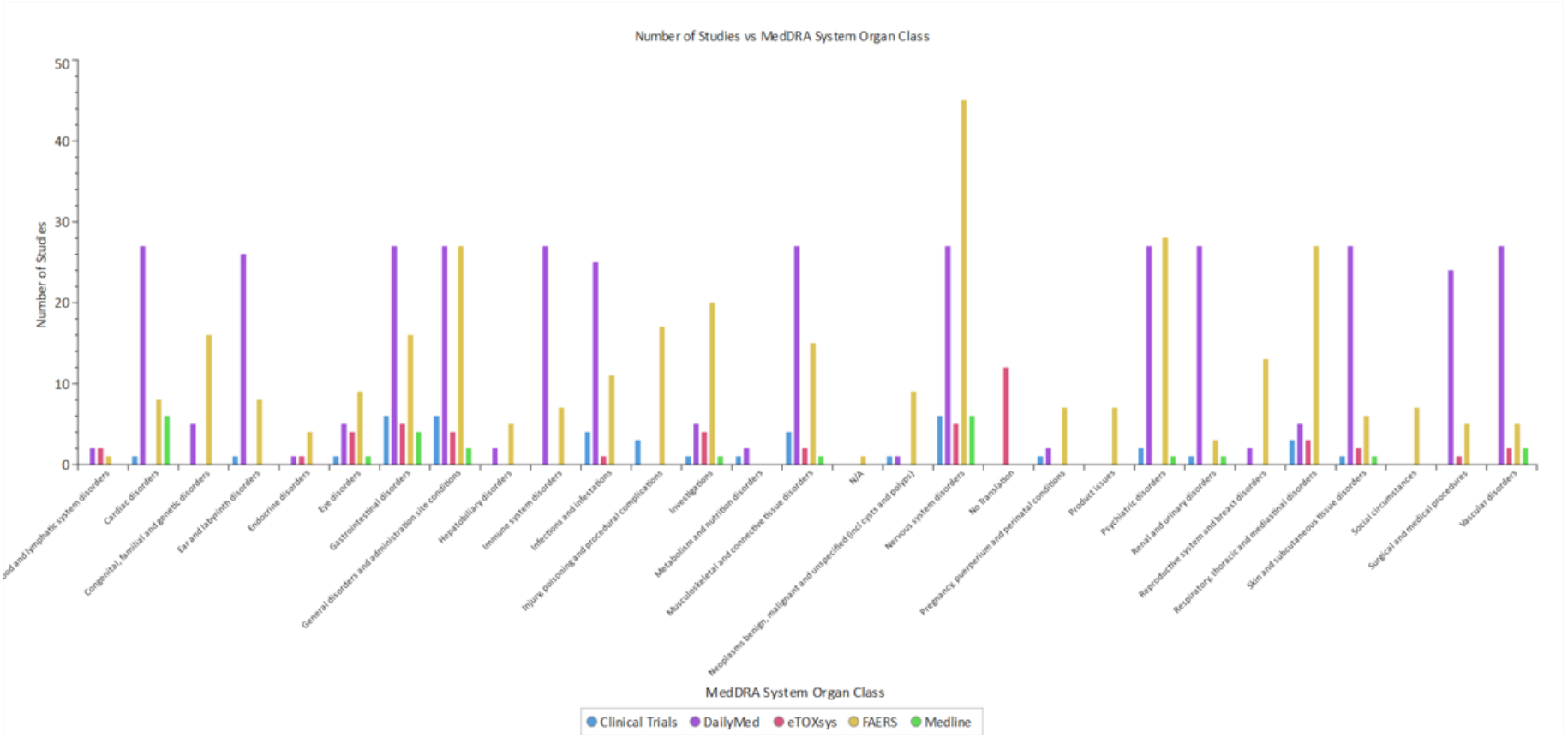
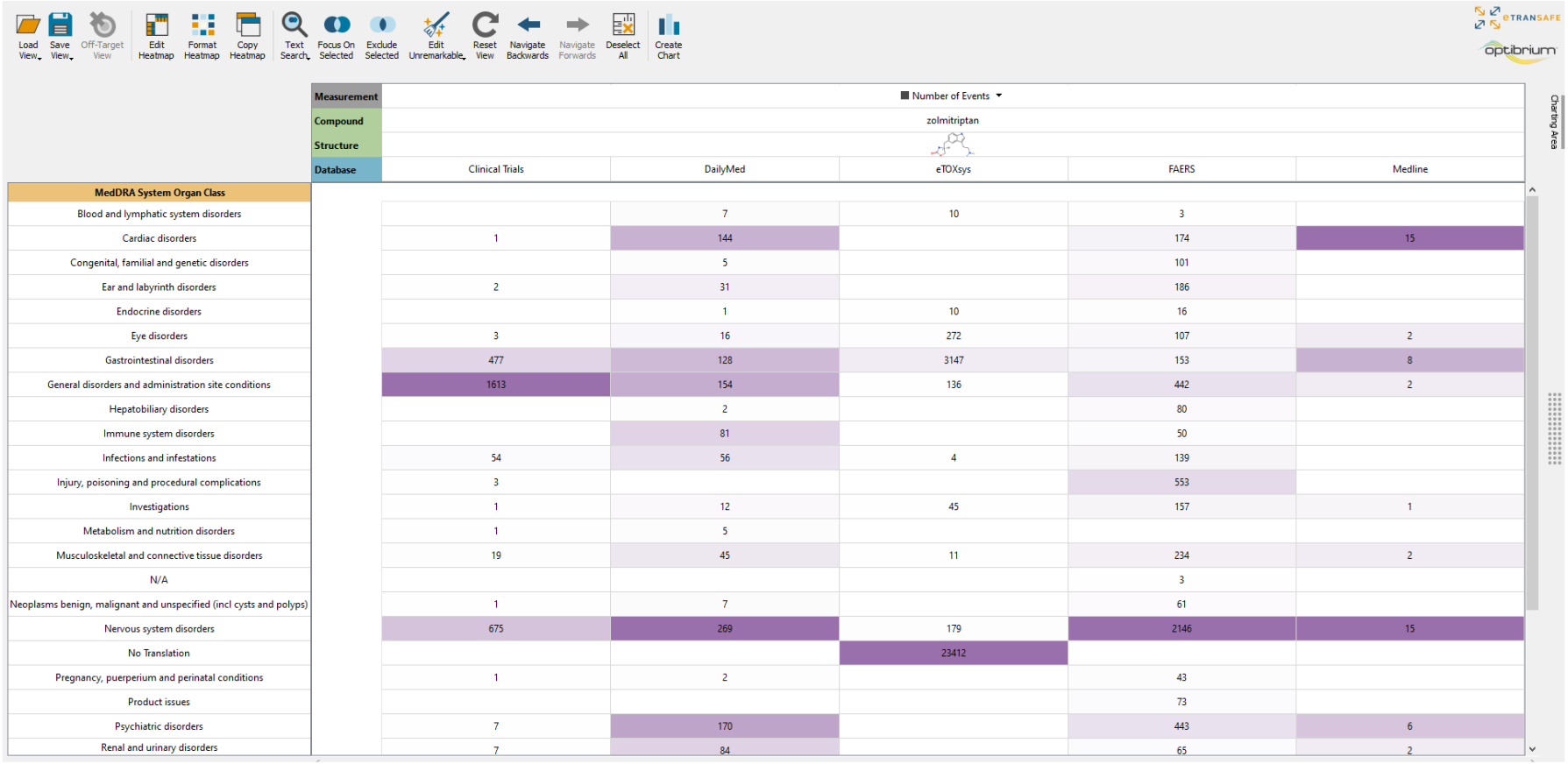


Figure 3. Findings Grouped by MedDRA System Organ Class and Database – Graphical Presentation. The translation engine allows groupings of the findings associated with studies and reports in the databases queried based on linkage of the terminology used for the original finding and organ with the SOC category. Databases are indicated by the different colors and the number of studies with findings in the SOC are on the Y-axis.



MedDRA System Organ Class	Clinical Trials	DailyMed	eTOXsys	FAERS	Medline
Blood and lymphatic system disorders		7	10	3	
Cardiac disorders	1	144		174	15
Congenital, familial and genetic disorders		5		101	
Ear and labyrinth disorders	2	31		186	
Endocrine disorders		1	10	16	
Eye disorders	3	16	272	107	2
Gastrointestinal disorders	477	128	3147	153	8
General disorders and administration site conditions	1613	154	136	442	2
Hepatobiliary disorders		2		80	
Immune system disorders		81		50	
Infections and infestations	54	56	4	139	
Injury, poisoning and procedural complications	3			553	
Investigations	1	12	45	157	1
Metabolism and nutrition disorders	1	5			
Musculoskeletal and connective tissue disorders	19	45	11	234	2
N/A				3	
Neoplasms benign, malignant and unspecified (incl cysts and polyps)	1	7		61	
Nervous system disorders	675	269	179	2146	15
No Translation			23412		
Pregnancy, puerperium and perinatal conditions	1	2		43	
Product issues				73	
Psychiatric disorders	7	170		443	6
Renal and urinary disorders	7	84		65	2

Figure 4. Findings Grouped by MedDRA System Organ Class and Database – Heatmap Presentation. The translation engine allows groupings of the findings associated with studies and reports in the databases queried based on linkage of the terminology used for the original finding and organ with the SOC category. Databases are indicated along the top of the columns with the SOC on the Y-axis. In this presentation of the data the number of events reported in each area and database is indicated in each cell. Rows are graded by intensity of color depending on the scale of the number across all of the cells in the row, higher numbers are a darker purple.

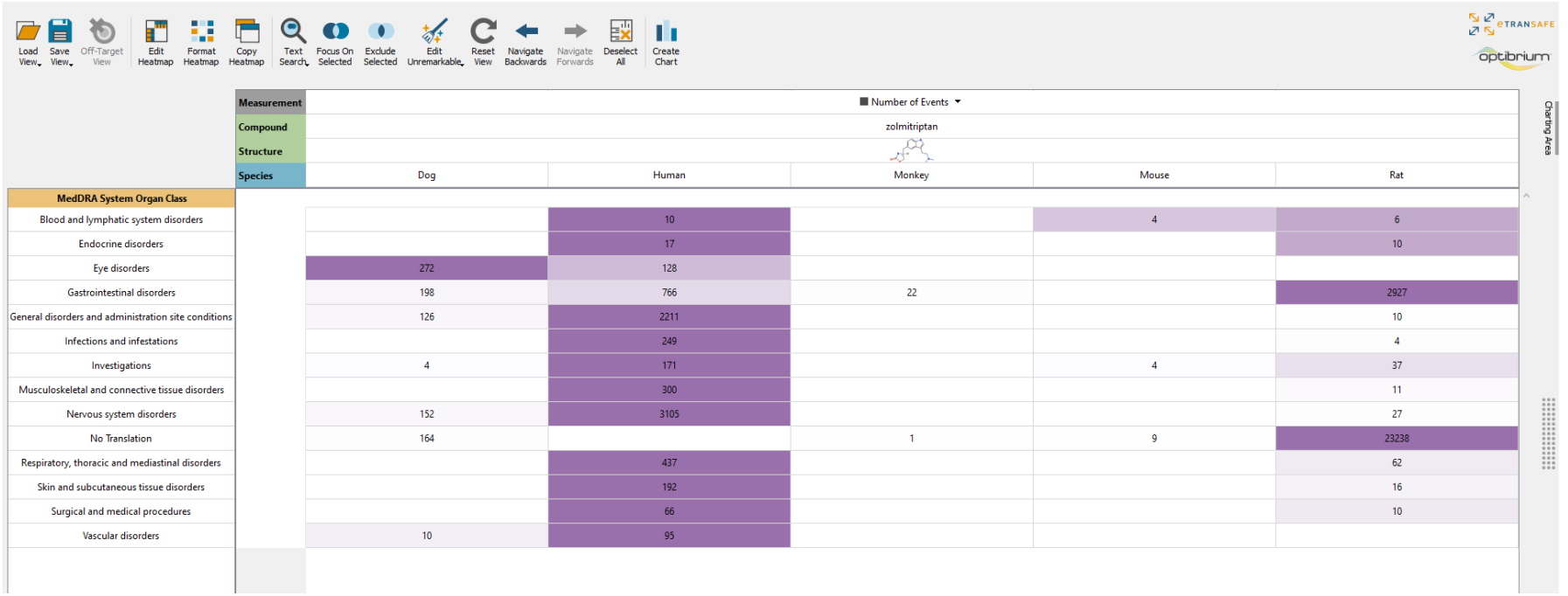
TRANSLATIONAL SAFETY ASSESSMENT ACROSS SPECIES

By using the presentation of the integrated data for several species as shown in the previous section, making use of the integrated ontological organization of the data and visualization tools, assessment of potential cross-species events can be conducted. With the flexibility to reorganize the presentation of data, the content and species can be made a focus, rather than the source, shown in Figure 5 where data has been grouped by species and database detail removed. In the view of the data presented in Figure 5, SOC classes where events were only present in humans were hidden, so that areas of potential overlap are clearly visible.

One of these areas is Eye Disorders with 128 events in humans and 272 in dogs. A deeper interrogation of this data is shown in Figure 6. Upon reviewing the original finding term used, there does not appear to be cross-over between the species as miosis is excessive constriction of the pupil and mydriasis is excessive or prolonged dilatation of the pupil. Both terms are relevant clinically, but neither appear to link to the findings listed for human patients. However, links to the original data sources are preserved which allows the scientist to validate the mapping of the terms and perform additional in-depth analysis of the study data if needed. For example, the eTOX data (used here) is available for each study as group level data while the SEND data is available down to the individual animal results. Both can be reviewed on a study-by-study or multiple study basis. For clinical data, links to ClinicalTrials.gov, DailyMed and Medline are present in the study list and opening those entries will take you to the web page relevant to the study.

Potential translationally relevant findings were observed in the Gastrointestinal disorders SOC. As shown in Figure 7, reports of salivary hypersecretion in humans, rats, dogs and monkeys. Interestingly there are also many studies that have reports of nausea in humans. Salivation has long been considered a clinical sign of nausea in many test species[11,12]. To confirm that there was not some underlying histopathological reason for this salivary hypersecretion in the animals, the microscopical data from the studies could be reviewed.

By reviewing data such as this prior to planning a trial, a company could ameliorate the potential for nausea which would likely markedly improve patient experience and retention as this is an adverse event that greatly impacts these factors for trials[12]. Considering the utility of this data review in a different way, a cross-species review could be conducted during development of a compound using a read across approach. By doing this the most relevant species could be selected for testing based on the potential toxicities or events predicted from the pooled data, or smaller more directed investigations could be conducted reducing the number of animal subjects used.



MedDRA System Organ Class	Dog	Human	Monkey	Mouse	Rat
Blood and lymphatic system disorders		10		4	6
Endocrine disorders		17			10
Eye disorders	272	128			
Gastrointestinal disorders	198	766	22		2927
General disorders and administration site conditions	126	2211			10
Infections and infestations		249			4
Investigations	4	171		4	37
Musculoskeletal and connective tissue disorders		300			11
Nervous system disorders	152	3105			27
No Translation	164		1	9	23238
Respiratory, thoracic and mediastinal disorders		437			62
Skin and subcutaneous tissue disorders		192			16
Surgical and medical procedures		66			10
Vascular disorders	10	95			

Figure 5. Findings Grouped by MedDRA System Organ Class and Species – Heatmap Presentation. The heatmap presentation was reconfigured to sort the data by species and focus on SOC where there were findings in humans and a second test species. The “No Translation” SOC entry was kept also as this is where findings without a map to SOC are held so scientific judgement can be used in the review of these and other data to determine if links are present.

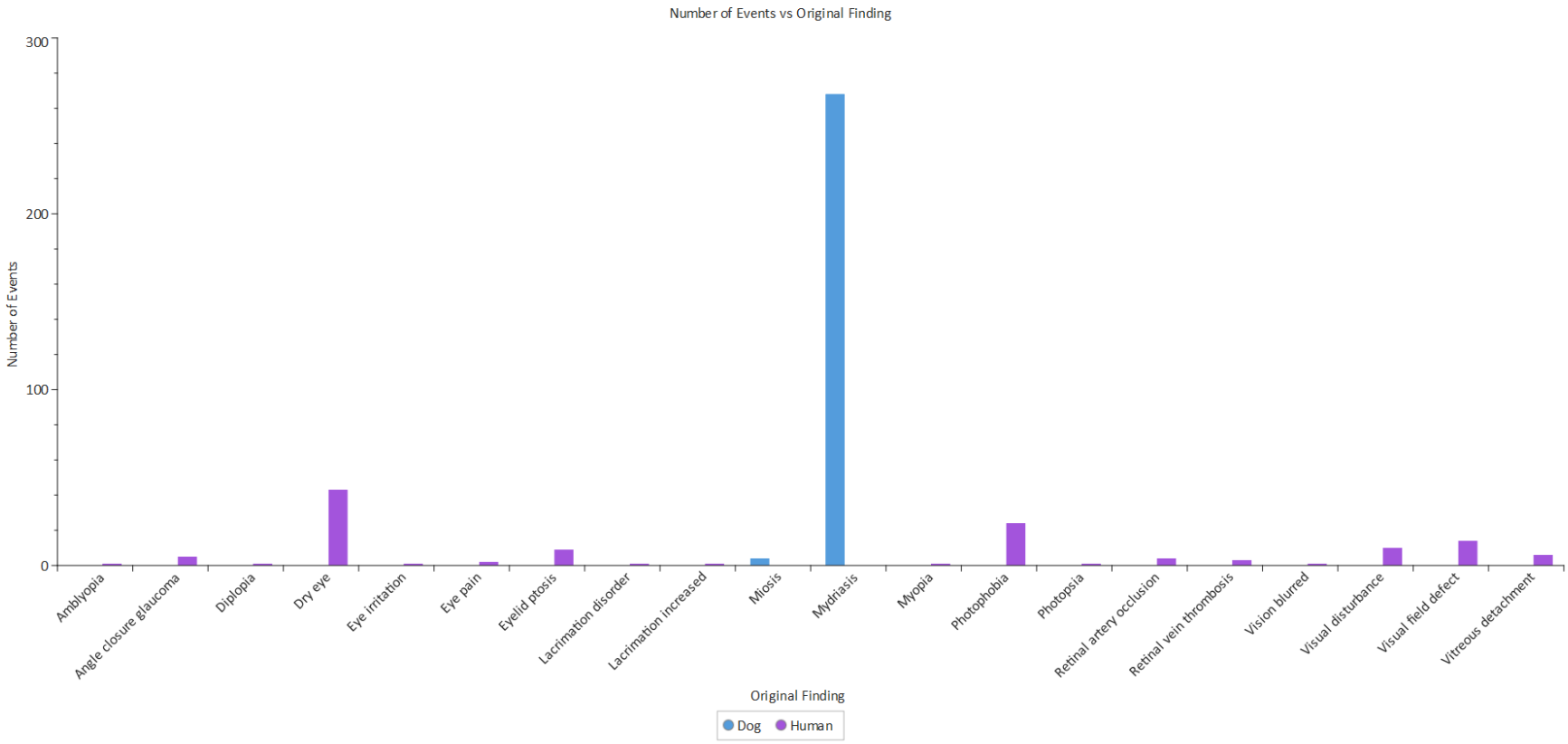


Figure 6. Investigation of potential species overlap – Eye Disorders. One SOC where there appeared to be significant overlap in the number of events reported was in Eye Disorders. A focused review of the original recorded findings within this SOC revealed that although there were findings associated with the eye in both humans and dogs these were not related.

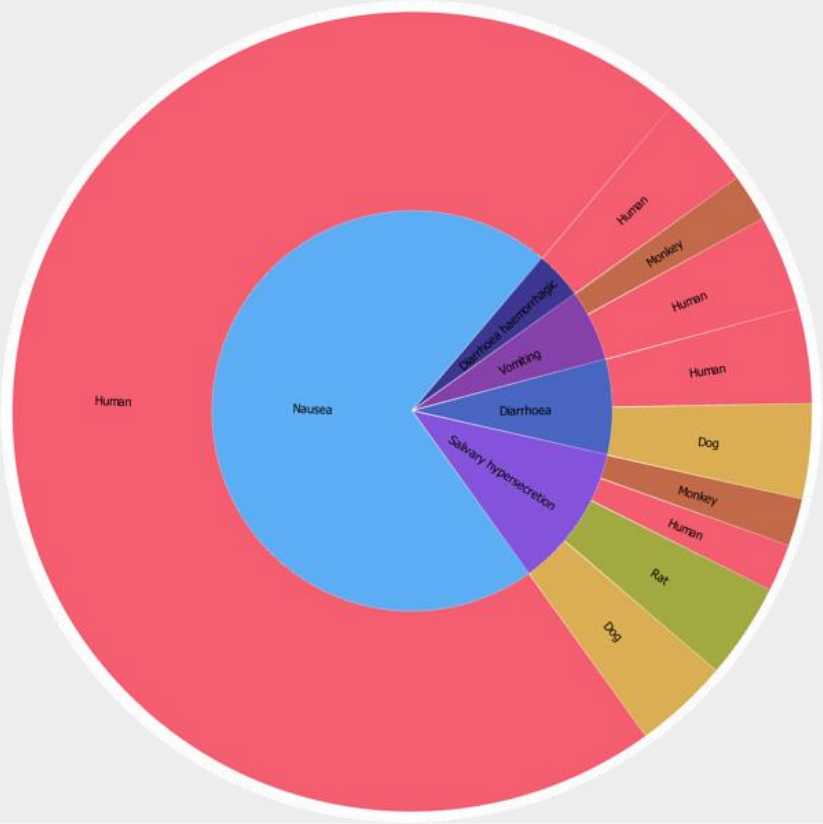


Figure 7. Investigation of potential species overlap – Gastrointestinal Findings. This sunburst plot of the gastrointestinal data was prepared with the finding at the center with the species scaled to study number on the outside ring.

FUTURE DIRECTIONS AND IMPLICATIONS

Not every aspect of data is categorized and recorded in any of the databases available within the system currently. For example, there may be preanalytical factors that not clearly indicated and could have impacted results. For example blood collection methods impacting laboratory test results. Therefore robust data QC is needed by subscribers and system owner prior to upload into the proprietary databases and critical review of the public data sources when conducting analysis in the system. However, due to the approach to data within the system, it should be possible to add unique information that is currently represented in SEND or other standard to add nuance and perspective. One area where this is currently being developed is by using applications within the platform to produce a SEND-like domain to help indicate treatment-related findings and other scientific interpretations present in study reports.

Unique compounds can add complexity to the capacity for prediction due to the lack of read-across using structural characteristics. The addition of other characteristics for the compound within the system such as intended-target or intended-use would add a different view, resulting in capacity for target-related safety assessments. But as investigation types expand, the ability to bring those data together remains a key requirement to make the best use of the this data to feed predictive and computational capabilities. By capturing domain knowledge in a structured format, we can allow for the development of AI and machine learning models that leverage this knowledge to automate tasks, discover patterns, and make predictions.

The ontology-based approach makes data discovery easier. Users can explore data using ontology concepts, relationships, and hierarchies, making it simpler to locate relevant information for their specific needs. Ultimately, ontologies reduce ambiguity in data interpretation. This is especially important when dealing with unstructured or semi-structured data, as ontologies provide a framework for extracting meaning from such data.

Therefore, having a system such as this which is malleable and scalable in terms of applications and database connections will be an asset for toxicologists and data scientists.

CONCLUSION

Ultimately, Centrus enhances the reviewer's capability and capacity by bringing data together. With the use of ontology and terminology mapping, the scientist can see groupings of findings so that potential cross-species translation could be highlighted and reviewed. Understanding the relative importance of findings such as these to the testing program could refine testing strategies, accelerate timelines and empower researchers. The modular architecture of the system allows for the standardized addition of data sources and expansion of the utility of the database. Future applications involving machine learning and predictive modelling (enabled by the use of large datasets collected by subscribers) will revolutionize the way we collect, examine and utilize preclinical and clinical information.

If you'd like to join this collaboration where we focus upon a future transformed for the benefit of everyone, please contact me to continue the conversation. We are excited to explore the possibilities of intelligent solutions empowering collaboration and life-enhancing science.

REFERENCES

1. FDA Modernization Act 2.0, September 2022
2. Seyhan AA (2019) Lost in Translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. Trans Med Comm. 4:18. doi.10.1186.s41231-019-0050-7
3. Fernandez-Moure JS (2016) Lost in Translation: The Gap in Scientific Advancements and Clinical Application. Front. Bioeng. Biotechnol. 4:43. doi: 10.3389/fbioe.2016.00043
4. Sarntivijai S, Blomberg N, Lauer KB et al. (2022) eTRANSAFE: Building a sustainable framework to share reproducible drug safety knowledge with the public domain [version 1; peer review: 2 approved] F1000Research, 11 (ELIXIR):287 https://doi.org/10.12688/f1000research.74024.1
5. Cases, M.; Briggs, K.; Steger-Hartmann, T.; Pognan, F.; Marc, P.; Kleinöder, T.; Schwab, C.H.; Pastor, M.; Wichard, J.; Sanz, F. (2014) The eTOX Data-Sharing Project to Advance *in Silico* Drug-Induced Toxicity Prediction. Int. J. Mol. Sci. *15*, 21136-21154. https://doi.org/10.3390/ijms151121136
6. Sanz F, Pognan F, Steger-Hartmann T, et al. (2023) eTRANSAFE: data science to empower translational safety assessment. Nat. Rev. Drug. Discov. Aug 22(8):605-606.
7. https://www.cdisc.org/standards/foundational/send (accessed October 10 2023)
8. https://www.meddra.org/ (accessed October 10 2023)
9. https://github.com/mi-erasmusmc/send-snomed-mappings
10. https://etransafe.eu/
11. https://www.petmd.com/dog/symptoms/nausea-dogs
12. Holmes AM, Rudd JA, Tattersall FD, Aziz Q, Andrews PL. Opportunities for the replacement of animals in the study of nausea and vomiting. Br J Pharmacol. 2009 Jul;157(6):865-80. doi: 10.1111/j.1476-5381.2009.00176.x.

ACKNOWLEDGEMENTS

Thank you to Frances Hall and Michael Wasko for their invaluable help in editing and review of this manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brenda Finney
Instem
Diamond Way
Stone ST15 0SD
Email: brenda.finney@instem.com
Web: www.instem.com