PreClinIE: An Annotated Corpus for Information Extraction in Preclinical Studies

Simona E. Doneva[1*], **Hanna Hubarava**[1], **Pia Haervelid**[1],
Wolfgang E. Zürrer [1], **Julia Bugajska**[1], **Bernhard Hild**[1], **David Brüschweiler**[1],
Tilia R. Ellendorff[1], **Gerold Schneider**[1], **Benjamin V. Ineichen**[1,2]
[1] University of Zurich, Zurich, Switzerland
[2] University of Bern, Bern, Switzerland
*simona.doneva@uzh.ch

Abstract

Animal research, sometimes referred to as preclinical research, plays a vital role in bridging the gap between basic science and clinical applications. However, the rapid increase in publications and the complexity of reported findings make it increasingly difficult for researchers to extract and assess relevant information. While automation through natural language processing (NLP) holds great potential for addressing this challenge, progress is hindered by the absence of high-quality, comprehensive annotated resources specific to preclinical studies. To fill this gap, we introduce PreClinIE, a fully open manually annotated dataset. The corpus consists of abstracts and methods sections from 725 publications, annotated for study rigor indicators (e.g., random allocation) and other study characteristics (e.g., species). We describe the data collection and annotation process, outlining the challenges of working with preclinical literature. By providing this resource, we aim to accelerate the development of NLP tools that enhance literature mining in preclinical research.

1 Introduction

Developing new therapies from animal models to human treatments, known as bench-to-bedside translation, has a low success rate: Only 1 in 20 therapies advances to human use (Ineichen et al., 2024). This contrasts with the extensive use of animals in research, estimated at over 50 million per year globally (Taylor and Alvarez, 2019). The factors that determine successful translation remain poorly understood (Seyhan, 2019).

A systematic assessment of information on experimental design, model and drug selection, as well as animal usage can provide insights into how animal research informs human health. The full-text, and especially the methods sections of scientific articles contain concrete, verifiable details about these aspects, which are often omitted or misrepresented in article abstracts (Li et al., 2017). These factual descriptions form the foundation of a study and are critical for evaluating its design, rigor, and to enable reproducibility (Menke et al., 2020).

However, the volume of preclinical animal studies is overwhelming, with hundreds of thousands published annually (Ineichen et al., 2023). While large-scale analysis methods exist, they primarily focus on human data or only on abstract level data (Chapman et al., 2011; Doneva et al., 2024). Animal studies, with their highly heterogeneous experimental approaches and less standardized reporting, remain largely unaddressed.

There is a critical need for computational methods to extract and integrate these data at scale, since a more detailed understanding of the drug development process could not only improve experimental animal welfare but also enhance the efficiency of human therapies. As a first step towards that goal, our study aims to create a large, manually annotated corpus of animal study publications, including abstracts and method sections. We share all resources on GitHub[1].

2 Related Work

NLP methods have been commonly applied in the preclinical domain for abstract classification tasks. For example, a recently published dataset aims to help with the identification of animal studies and alternative experimental models (Neves et al., 2023). Another application is the automated selection of relevant published articles for specific literature review questions, as well as the assessment of risk of bias items (e.g., random allocation) (Bannach-Brown et al., 2019; Wang et al., 2022b).

Information extraction from preclinical literature is an emerging, but less developed, area of research.

[1] https://github.com/Ineichen-Group/Preclinical_IE_Dataset

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 74–87

STEED, for instance, is an R-based text mining tool that uses regular expressions to automatically extract key experimental details, such as animal species, disease models, and randomization from neuroscience in vivo studies. It has been developed on 45 full-text articles and validated on 275 articles (Zurrer et al., 2024). Another approach, *Menagerie*, combines rule-based, dictionary-based, and machine learning techniques to extract six predefined animal study characteristics (Zeiss et al., 2019). This work is based on a manually curated dataset of 504 PubMed abstracts, annotated with classes such as species or animal model at the abstract level, and with gene names at the token level. Another related work targets information extraction based on the established framework of Population/Problem, Intervention, Comparator and Outcome (PICO) (Wang et al., 2022a). For this, 400 abstracts of preclinical studies have been annotated for each PICO-related mention, and the task was solved as sentence classification, followed by entity recognition. Another study proposed combining a regex-based method with a generative LLM to extract interventions from preclinical animal studies on Alzheimer's disease (Pu et al., 2024).

Despite recent advances, existing corpora remain limited in scope - typically focusing on narrow disease domains, containing small datasets (around 500 documents), and offering only abstract-level annotations. For example, *Menagerie* was validated solely for Parkinson's disease. In contrast, our corpus is, to our knowledge, the most comprehensive resource of its kind: it includes 725 documents from the general neuroscience domain, with manual annotations on both the abstract and the methods section, a critical source of experimental detail. This results in 1,450 annotated sections. Importantly, we used three annotation levels (document, sentence, and token), aiming to match the typical granularity of information relevant to researchers. This structure also should reflect the nature of the content: some elements, like conclusions, require sentence- or document-level annotation, whereas others, like individual drugs, can be annotated at the token level.

3 The Corpus

3.1 Data Collection

A search string for PubMed and EMBASE was designed to identify animal studies on therapeutic interventions[2]. From the retrieved references, 4,000 records were randomly selected for screening by two independent reviewers based on inclusion criteria: primary studies involving drug testing in animals.

We used the automatic fetch function of the reference management tool EndNote to retrieve PDFs, resorting to manual retrieval when necessary. We used IBM Deepsearch to convert PDFs into text[3], followed by a regular expression-based algorithm to classify paper sections such as methods and results. We included the methods sections because they typically provide more detailed descriptions of the employed methodology compared to abstracts.

3.2 Data Annotation

3.2.1 Annotation Guidelines

We define three levels of annotation. At the **document** level, one or more labels are assigned to the entire document. At the **sentence** level, we highlight the sentence where the relevant information appears (**Table 1**). Finally, at the **token** level, individual words are annotated as named entities (**Table 2**). We refined the annotation guidelines iteratively to ensure maximum clarity and optimize inter-rater agreement. The final guidelines can be accessed at Annotation Guidelines (v5), with a shortened version in **Appendix B**. Notably, spans and documents can have more than one label. For example, weight and age of animals are often reported in the same sentence, and a study can involve both mice and rats in its experiments.

3.2.2 Annotation Process

From the 4000 random references, we excluded two due to missing metadata, leaving 3,998 references. Of these, 1,018 met the inclusion criteria during the initial screening.

The annotation was conducted by five senior medical students, starting with two pilot rounds of 20 and then 50 articles annotated by all annotators to familiarize themselves with the task and to refine the guidelines. In the final annotation round, 817 articles were distributed among them, with each annotator receiving 179–181 articles with title, abstract and method sections. Of these, 20 articles were assigned to multiple annotators to calculate inter-annotator agreement (IAA). The annotators

[2]Search date: from database inception to October 09, 2023. Full search string available here: dataset search strings.

[3]IBM RPA PDF Extractor

Parameter	Label (frequency)	Krippendorff's Alpha (95% CI)
Document-level Annotation		
Animal species (A, M)	Rat (806), Mouse (531), Other (28), Rabbit (28), Monkey (20), Dog (15), Pig (10), Cat (6), Guinea Pig (6)	0.97 (0.95, 1.00)
Control (A, M)	Control-present (1135)	0.51 (0.29, 0.67)
Readout (A, M)	Physiology (400), Behaviour (938), Histology (921), Other (896), Imaging (92)	0.48 (0.39, 0.55)
Animal sex (A, M)	Not reported (717), Male (524), Female (129), Both sexes (63)	1.00 (1.00, 1.00)
Sentence-Level Annotation (Highlight)		
Study conclusions (A)	Positive (645), Neutral (22), Negative (18), Mixed (15)	0.76 (0.74, 0.78)
Animal disease model (A)	Model (649)	0.62 (0.60, 0.64)
Weight (M)	Weight (514)	0.73 (0.71, 0.75)
Age (M)	Age (476)	0.75 (0.73, 0.78)
Random allocation (A, M)	Randomization (464)	0.60 (0.56, 0.64)
Blinded outcome assessment (A, M)	Blinding (389)	0.97 (0.95, 0.98)
Animal welfare statement (A, M)	Welfare (700)	0.96 (0.95, 0.97)
Animal Research: Reporting of In Vivo Experiments Guidelines (A, M)	ARRIVE (15)	—
Sample size calculation (A, M)	Power (22)	—

Table 1: Overview of document-level and sentence-level annotation categories. The "Label (frequency)" column lists the available labels for each category along with their frequency in the final complete annotated dataset. For document-level annotations, the frequency represents the number of documents (abstracts or methods) assigned to each label. For sentence-level annotations, it indicates the number of unique sentences associated with each label. The last column provides the Krippendorff's Alpha inter-annotator-agreement score for that label on the subset of the corpus annotated by all annotators (15 articles). The rows with a missing score correspond to the labels not sufficiently represented in the subset. Abbreviations: A, abstract; M, methods.

Entity Type	Entity #	Unique #	Avg Char Count	Examples	Krippendorff's Alpha (95% CI)
Therapy-drug (A, M)	10348	2437	17.7 ± 18.1	beta-lactam antibiotic, ZM241385	0.73 (0.70, 0.75)
Therapy-other (A, M)	5216	1728	20.3 ± 15.8	auditory habilitation, treadmill training	0.59 (0.57, 0.61)
Disease (A)	3790	958	19.8 ± 11.0	minimal seizures, chronic paraplegia	0.79 (0.76, 0.81)
Strain (A, M)	1196	159	10.5 ± 18.7	Sprague Dawley, Fisher 344	0.84 (0.79, 0.88)
Animals-number (A, M)	342	144	5.5 ± 4.8	Eighty-five, 128	0.78 (0.50, 0.93)

Table 2: Overview of token-level annotations with total entity counts, unique instances counts, average character number, and annotation examples. The last column provides the Krippendorff's Alpha inter-annotator-agreement score as measured on the subset of the corpus annotated by all annotators (15 articles). Abbreviations: A, abstract; M, methods.

were allowed to exclude articles from annotation if they did not fit the inclusion criteria[4].

The annotators used a custom recipe developed for the browser-based tool Prodigy to perform the manual annotation (Montani and Honnibal, 2017). An annotation task example is shown in **Supplementary Figure 2**.

To compile the final dataset and perform an error analysis, the 20 multiple-assigned articles were reviewed, with conflicts adjudicated through discussion. The final dataset consists of 725 unique articles, and corresponds to 1450 abstract and method sections.

[4]For example, some initially included papers were related to a diagnostic procedure rather than an intervention.

3.2.3 Inter-Annotator Agreement

Five of the 20 common documents were excluded by one or more annotators as not meeting the inclusion criteria. To ensure the IAA score is measured among all annotators, we removed the excluded articles from the agreement calculations. This left us with 15 unique articles, each with an abstract and method section (30 annotated documents).

We report the IAA among the five annotators using Cohen's Kappa for pairwise agreement calculation and Krippendorff's Alpha for the calculation of agreement among all annotators (Cohen, 1960; Hughes, 2021).

3.3 Results

3.3.1 Corpus Overview

Our final annotated corpus consists of the abstracts and methods sections from 725 published neuroscience articles, primarily dated between 2010 and 2020. The most frequently represented journals include *European Journal of Pharmacology* and *PLoS ONE* (**Supplementary Figure 1**).

Based on the **document level** annotations, the corpus predominantly comprises studies involving mice and rats (**Table 1**). Additionally, there is a marked bias toward male animals and reporting of animal sex information in the methods section. Furthermore, there were often multiple selected options for the experimental readouts (**Supplementary Figure 3**).

At the **sentence level**, the annotations reveal that the majority of conclusion statements within the corpus present positive findings. However, adherence to reporting best-practice appears limited, with relatively little to no mention of ARRIVE and PREPARE guidelines (Percie du Sert et al., 2020; Smith et al., 2018). Furthermore, explicit reporting of sample size calculations is sparse (**Table 1**).

At the **token level**, therapy-related annotations are the most prevalent, as these terms were annotated in both the abstract and methods sections (**Table 2**). Disease and strain entities exhibit high lexical variability, with high ratio of unique textual representations across the corpus. Additionally, many abbreviations are annotated, such as *AD* for Alzheimer's disease (**Supplementary Figure 5**).

3.3.2 Analysis of Annotation Disagreements

We observed several patterns of discrepancies in the multiply-annotated documents selected for the calculation of IAA:

- **Text level annotations:** "Readout" and "Control" were the most challenging document-level classification tasks (**Table 1**). The language describing the readouts varied greatly across papers and was often not explicit. Some annotators selected "other" in these cases, while others attempted to infer a more specific readout type. Furthermore, some annotators selected "histology" in cases when there was clearly no mention of this readout, suggesting a misunderstanding of the concept. Similarly, the presence of control intervention was rarely specified, even though the text sometimes contained a comparison verb (e.g., "improved"). In such instances, where the presence of control is implicit, some annotators marked the presence of control, while others did not, leading to a lower agreement score. The variability in pair-wise IAA is evident from **Supplementary Figure 4**.

- **Sentence level annotations:** At the sentence level, annotators often agreed on the study's overall conclusion but struggled to identify the exact concluding sentence, sometimes confusing it with a summary of findings. Variation in punctuation usage—especially around colons and semicolons—also caused inconsistencies in the selection of annotation spans, resulting in partial agreement. Annotation of "randomization" was meant to refer only to the allocation of animals into experimental groups, but one annotator highlighted other contexts as well.

- **Token level annotations:** At the token level all entities except "therapy-other" achieved a satisfactory level of agreement (**Table 2**). We identified three main discrepancies. First, some annotators occasionally missed entities, either due to human error in reading longer texts or a misunderstanding of the guidelines. For instance gene mentions were sometimes annotated as therapy (Nurr1, Fox2) when, according to the guidelines, they should not be annotated as such. Second, label disagreements arose when different annotators assigned different labels to the same entity. For instance, one annotator consistently labeled antibody therapies as "therapy-other", while it should have been "therapy-drug". Finally, span disagreements occurred when some annotators included a preceding modifier as per the guidelines instructions (e.g., "morphine-induced"), while another did not. Such discrepancies introduce noise in the dataset and reduce the agreement score. The variability in pair-wise IAA is shown in **Supplementary Figure 5**.

4 Discussion and Conclusion

We introduced PreClinIE, an openly available corpus for extracting study rigor indicators and experimental details from published articles describing animal research.

Our annotation process uncovered key challenges. Particularly there was a low agreement

in control/comparator annotations, aligning with findings from the related PICO study (Wang et al., 2022a). At the same time, this study had stronger performance for readout extraction, suggesting that token-level annotation may be more suitable for this task. Additionally, the high disagreement in identifying the exact conclusion sentence suggests that study conclusions might be best evaluated using the full document rather than isolated sentences. Furthermore, we observed that crucial study details often appear exclusively in the methods section, emphasizing the importance of section-aware extraction.

In designing the annotation scheme, we made several pragmatic choices to balance granularity, feasibility, and consistency. For example, while our approach captures individual parameters such as sex, strain, and treatment, it does not explicitly encode relationships between these entities. As a result, reconstructing complex experimental groupings may be challenging in studies involving multiple animal subgroups. Nevertheless, this design simplifies annotation and aligns with our primary goal of extracting key methodological features at scale. Future work could explore incorporating relational annotations to capture richer experimental structures. Additionally, it may be possible to use simpler heuristic rules, for instance, pairing species and model terms that occur within a pre-specified window in the text, to make those links more explicit (Zeiss et al., 2019).

Similarly, we chose to restrict some annotations to single sentences. This constraint reduces cognitive load for annotators. Although it may result in missed information that spans multiple sentences, such as animal welfare statements or methodological clarifications, we find that capturing key information once in the text is sufficient for many downstream applications. Future extensions of the annotation scheme could explore cross-sentence linking or section-level annotation to support more nuanced analysis.

Beyond annotation challenges, our findings highlight a male bias in animal use, a majority of positive conclusion statements indicative of reporting bias, and insufficient reporting of sample size calculations.

These patterns warrant further evaluation, as they suggest systemic issues in study design and reporting that could impact the reliability and reproducibility of preclinical research findings (Beery and Zucker, 2011; Button et al., 2013). As future work, we plan to provide a baseline experiment to illustrate how the dataset can support computational information extraction from preclinical literature. The corpus enables a range of NLP tasks, such as named entity recognition and sentence classification, and can serve as a benchmark for model development in this domain. We hope these efforts will inspire further research in NLP models development and evaluation, ultimately contributing to more transparent and reliable scientific practices.

Limitations

Data Scope. Our developed dataset includes publications focusing mainly on research in neuroscience. This may influence the generalizability of our findings to other areas of research.

Annotation Setup. Only a small portion of the dataset was multiply-annotated. We conducted two annotation pilots to harmonize understanding among annotators. However, more multiply-annotated documents and additional training sessions would likely have further improved annotation quality.

Possible need for enrichment of the data. Another challenge is the under-representation of certain classes in our dataset. As an example of the imbalance on the sentence-level annotations, the number of positive study conclusions (625) dwarfs the negative (18), neutral (21) and mixed (15) conclusions. Among document-level annotations, "animal species" class shows that the majority of animals used for experiments are mice and rats, with only a handful of other species found in the dataset (see Table **1**). Although likely reflecting the natural distribution of the conclusions among publications, this imbalance may limit the model performance for those categories. Potential remedies include merging our dataset with related ones, applying targeted data collection strategies to expand coverage and improve class balance, or augmenting the dataset with synthetic data.

Acknowledgments

We thank Vera Lara Bernhard for her support in implementing the confidence interval calculations for the Krippendorff's alpha algorithm.

References

Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew SC Rice, Sophia Ananiadou, Jing

Liao, and Malcolm Robert Macleod. 2019. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews*, 8:1–12.

Annaliese K Beery and Irving Zucker. 2011. Sex bias in neuroscience and biomedical research. *Neuroscience & Biobehavioral Reviews*, 35(3):565–572.

Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Simona Emilova Doneva, Sijing Qin, Beate Sick, Tilia Ellendorff, Jean-Philippe Goldman, Gerold Schneider, and Benjamin Victor Ineichen. 2024. Large Language Models to process, analyze, and synthesize biomedical texts-a scoping review. *Discover*, 4:107.

John Hughes. 2021. krippendorffsalpha: An R package for measuring agreement using Krippendorff's alpha coefficient. *arXiv preprint arXiv:2103.12170*.

Benjamin V Ineichen, Eva Furrer, Servan L Grüninger, Wolfgang E Zürrer, and Malcolm R Macleod. 2024. Analysis of animal-to-human translation shows that only 5% of animal-tested therapeutic interventions obtain regulatory approval for human applications. *PLoS biology*, 22(6):e3002667.

Benjamin V Ineichen, Marianna Rosso, and Malcolm R Macleod. 2023. From data deluge to publomics: How AI can transform animal research. *Lab animal*, 52(10):213–214.

Guowei Li, Luciana PF Abbade, Ikunna Nwosu, Yanling Jin, Alvin Leenus, Muhammad Maaz, Mei Wang, Meha Bhatt, Laura Zielinski, Nitika Sanger, et al. 2017. A scoping review of comparisons between abstracts and full reports in primary biomedical research. *BMC medical research methodology*, 17:1–12.

Joe Menke, Martijn Roelandse, Burak Ozyurt, Maryann Martone, and Anita Bandrowski. 2020. The rigor and transparency index quality metric for assessing biological and medical science methods. *Iscience*, 23(11).

Ines Montani and Matthew Honnibal. 2017. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

Mariana Neves, Antonina Klippert, Fanny Knöspel, Juliane Rudeck, Ailine Stolz, Zsofia Ban, Markus Becker, Kai Diederich, Barbara Grune, Pia Kahnau, et al. 2023. Automatic classification of experimental models in biomedical literature to support searching for alternative methods to animal experiments. *Journal of Biomedical Semantics*, 14(1):13.

Nathalie Percie du Sert, Viki Hurst, Amrita Ahluwalia, Sabina Alam, Marc T Avey, Monya Baker, William J Browne, Alejandra Clark, Innes C Cuthill, Ulrich Dirnagl, et al. 2020. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *Journal of Cerebral Blood Flow & Metabolism*, 40(9):1769–1777.

Yiyuan Pu, Kaitlyn Hair, Daniel Beck, Mike Conway, Malcolm Macleod, and Karin Verspoor. 2024. Intervention extraction in preclinical animal studies of Alzheimer's Disease: Enhancing regex performance with language model-based filtering. In *23rd Meeting of the ACL Special Interest Group on Biomedical Natural Language Processing, BioNLP 2024*, pages 486–492. Association for Computational Linguistics (ACL).

Attila A Seyhan. 2019. Lost in translation: the valley of death across preclinical and clinical divide–identification of problems and overcoming obstacles. *Translational Medicine Communications*, 4(1):1–19.

Adrian J Smith, R Eddie Clutton, Elliot Lilley, Kristine E Aa Hansen, and Trond Brattelid. 2018. PREPARE: guidelines for planning animal research and testing. *Laboratory animals*, 52(2):135–141.

Katy Taylor and Laura Rego Alvarez. 2019. An estimate of the number of animals used for scientific purposes worldwide in 2015. *Alternatives to Laboratory Animals*, 47(5-6):196–213.

Qianying Wang, Jing Liao, Mirella Lapata, and Malcolm Macleod. 2022a. PICO entity extraction for preclinical animal literature. *Systematic Reviews*, 11(1):209.

Qianying Wang, Jing Liao, Mirella Lapata, and Malcolm Macleod. 2022b. Risk of bias assessment in preclinical literature using natural language processing. *Research synthesis methods*, 13(3):368–380.

Caroline J Zeiss, Dongwook Shin, Brent Vander Wyk, Amanda P Beck, Natalie Zatz, Charles A Sneiderman, and Halil Kilicoglu. 2019. Menagerie: A text-mining tool to support animal-human translation in neurodegeneration research. *PloS one*, 14(12):e0226176.

Wolfgang Emanuel Zurrer, Amelia Elaine Cannon, Ewoud Ewing, David Brüschweiler, Julia Bugajska, Bernard Friedrich Hild, Marianna Rosso, Daniel Salo Reich, and Benjamin Victor Ineichen. 2024. STEED: A data mining tool for automated extraction of experimental parameters and risk of bias items from in vivo publications. *PloS one*, 19(11):e0311358.

A Corpus Details and Statistics

A.1 Corpus Overview

Figure 1 shows the time range and journals represented in the corpus.

A.2 Annotations Overview

Figure 3 outlines the top 10 annotations across the different document-level categories and their distribution by abstract and methods. Rodent models, particularly rats and mice, dominate the dataset, with other species such as rabbits, guinea pigs, dogs, and monkeys appearing infrequently. Experimental outcomes are diverse, with histology and behavior among the most common readouts, often annotated together, indicating a tendency to explore multiple endpoints, as well as possible annotation challenge. The majority of studies include a control group, though fewer are explicitly mentioned in abstracts. For animal sex, a male bias is evident, as well as lack of reporting of animal sex in the abstract.

Figure 5 focuses on entity-level annotations in the corpus. Therapy-related drug entities are the most frequently annotated, with levodopa (109 instances) and L-DOPA (86 instances) leading the list, followed by commonly studied compounds such as morphine, MK-801, and cannabidiol (CBD). Beyond pharmacological interventions, other therapy entities include treatments like exercise, acupuncture, and curcumin. Among disease entities, Alzheimer's disease (96 instances) and stroke (103 instances) are well-represented, while neurodegenerative and neurological conditions such as Parkinson's disease (88 instances), epilepsy (48 instances), and spinal cord injury (81 instances) also feature prominently. Regarding animal strains, Sprague-Dawley (239 instances) and Wistar (207 instances) are the most frequently reported. However, annotations for animal numbers show substantial variability. The frequent presence of abbreviations (e.g., AD for Alzheimer's disease and SD for Sprague-Dawley) suggests that entity disambiguation is critical for accurate text interpretation.

A.3 Inter-Annotator Agreement Scores

We report IAA for document-level (**Figure 6**, left), as well as sentence-level and entity-type annotations (**Figure 6**, right). For the latter, we compute and report Krippendorff's Alpha on the level of tokens (words). This allows to capture partial agreement, when annotators agree on the label but disagree on its span, i.e. where exactly it starts and ends in the text.

B Annotation Guidelines

See full document here Annotation Guidelines (v5).

B.1 Inclusion Criteria for Papers

Before starting the annotation, ensure the paper meets the following eligibility criteria:

1. **Experimental study in animals** (excluding humans).

2. The study tests an **intervention** with the goal of improving animal health. The intervention should be externally applied (e.g., gene knockout does not qualify).

    - Apply criterium generously; include studies where the exact purpose of a drug treatment is not explicitly stated (e.g., testing different substances in animals without claiming a therapeutic benefit).
    - Exclude studies assessing the effect of endogenous substances (e.g., endogenously excreted miRNA-107).

3. The study models a **neurological or psychiatric disease**.

    - Apply criterium generously, including studies assessing pain in osteoporosis or mentioning neurological complications in systemic diseases such as cryptococcosis.

If any of these criteria are not met, exclude the study (no annotation required). If pertinent, exclude at the **abstract level** to ensure all related text sections (abstract, methods, and results) are omitted.

B.2 General Rules on Annotation

B.2.1 Token Annotation

1. **Consider Context:** Identify Population (P), Intervention (I), Control (C), and Outcomes (O).

2. **Annotate Only Relevant Information:**

    - Example: If a study uses male mice but suggests repeating experiments in rats, only mice should be annotated.

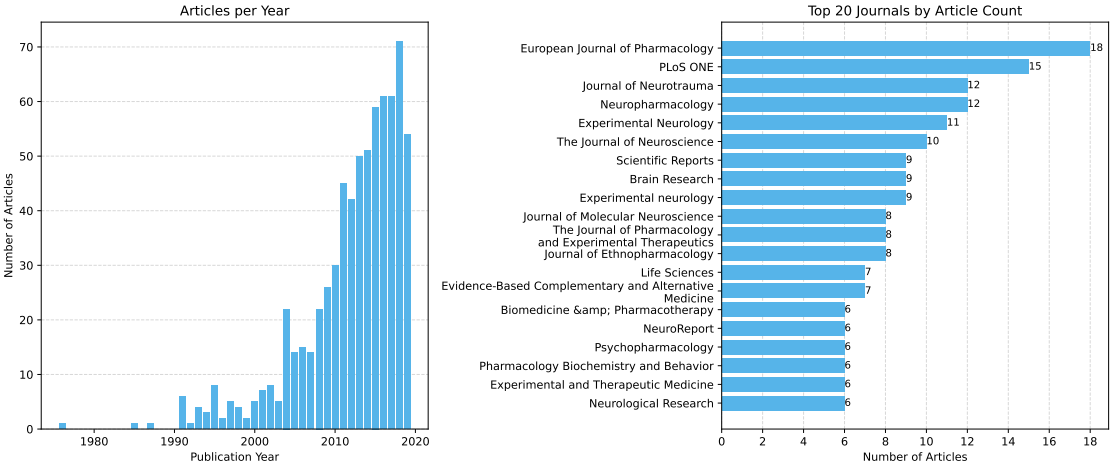

Figure 1: Distribution of articles in the corpus by (left) publication year and (right) journal.

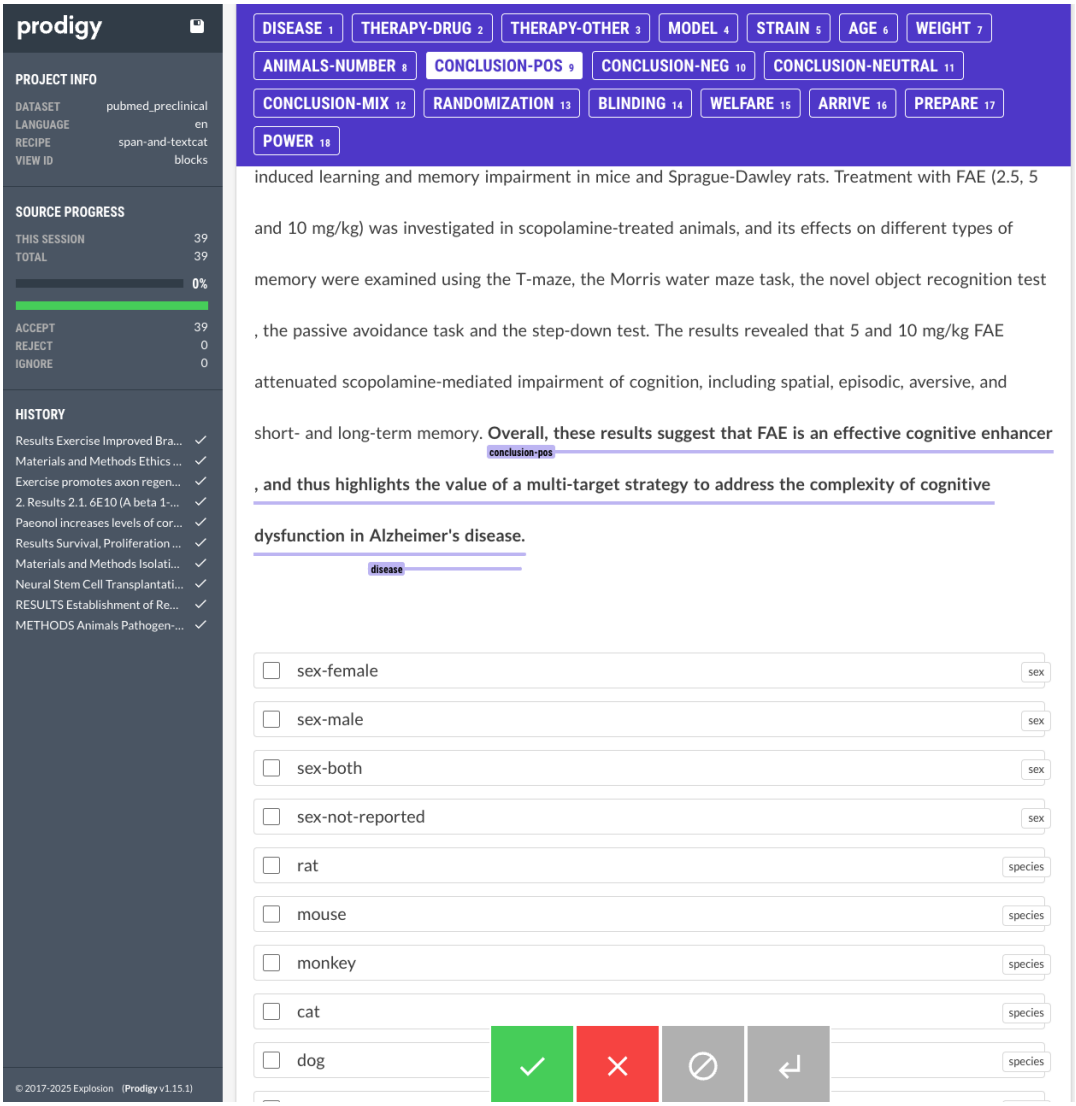

Figure 2: Annotation example shown in the annotation tool Prodigy.

- Example: If isoflurane is used for anaesthesia but not as a treatment, do not an-

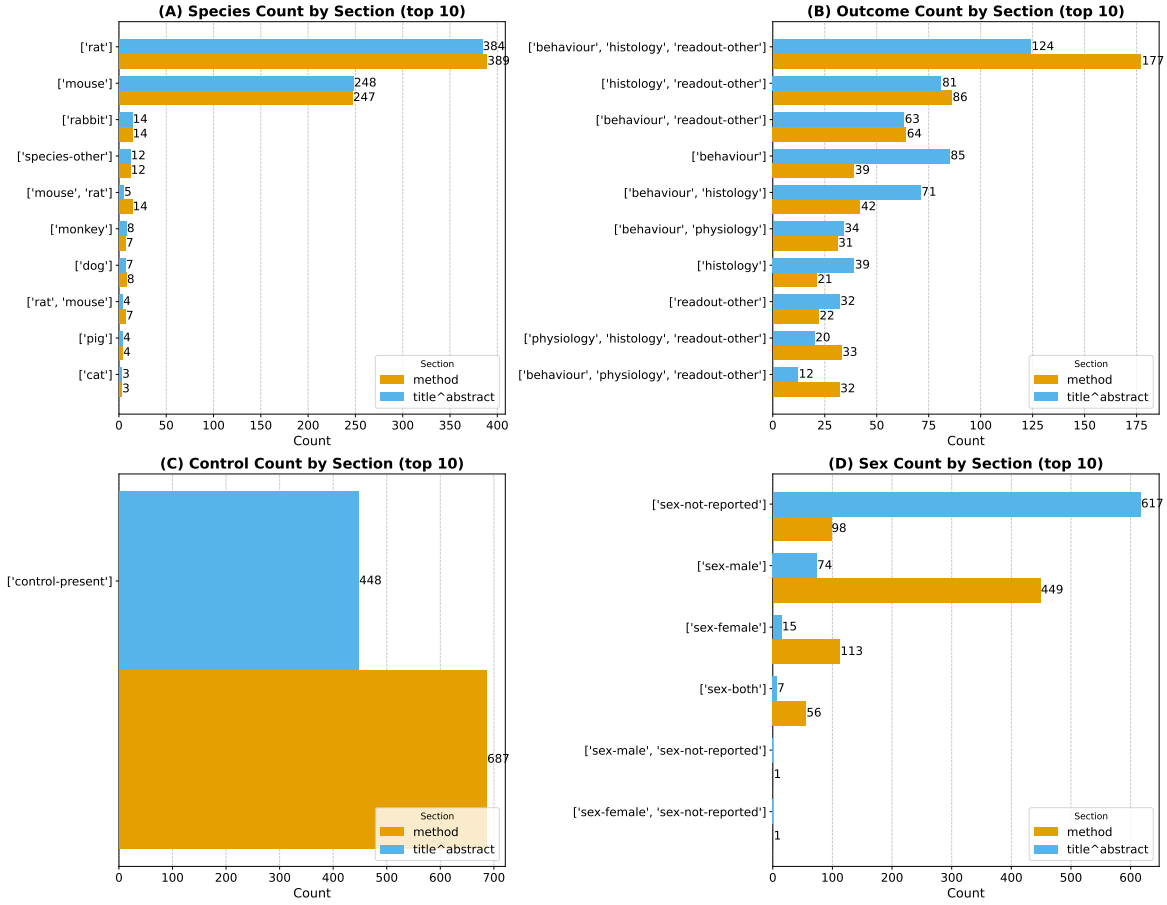

Figure 3: Top 10 most frequent document-level annotations for (A) Species, (B) Outcomes, (C) Control, and (D) Sex categories.

notate it.

B.2.2 Annotation Tasks

- **Text Annotation:** Entire text is classified based on predefined labels.
- **Sentence Annotation:** Entire sentence (including punctuation like colons and question marks) is annotated.
- **Token Annotation:** Specific words or phrases are annotated.

B.2.3 Additional Annotation Rules

1. Sentence annotations should **exclude references at the end of sentences**.
2. Include incorrect spelling/grammar if relevant.
3. Avoid mixing terms with and without brackets in annotations (e.g., annotate *oral appliance* and *OA* separately).
4. Do not annotate tokens where annotation would require inclusion of punctuation due to interface limitations.
5. Overlapping annotation between different tags is allowed.
6. Annotate each parameter once per section (i.e., once in abstract and once in methods); conclusions should only be annotated in the abstract.
7. Be careful in selecting the correct label, as incorrect annotation affects inter-rater agreement.
8. Do not annotate punctuation at the end of a sentence.
9. Ignore manuscript parts mistakenly included in the annotation interface (e.g., misplaced discussion sections).

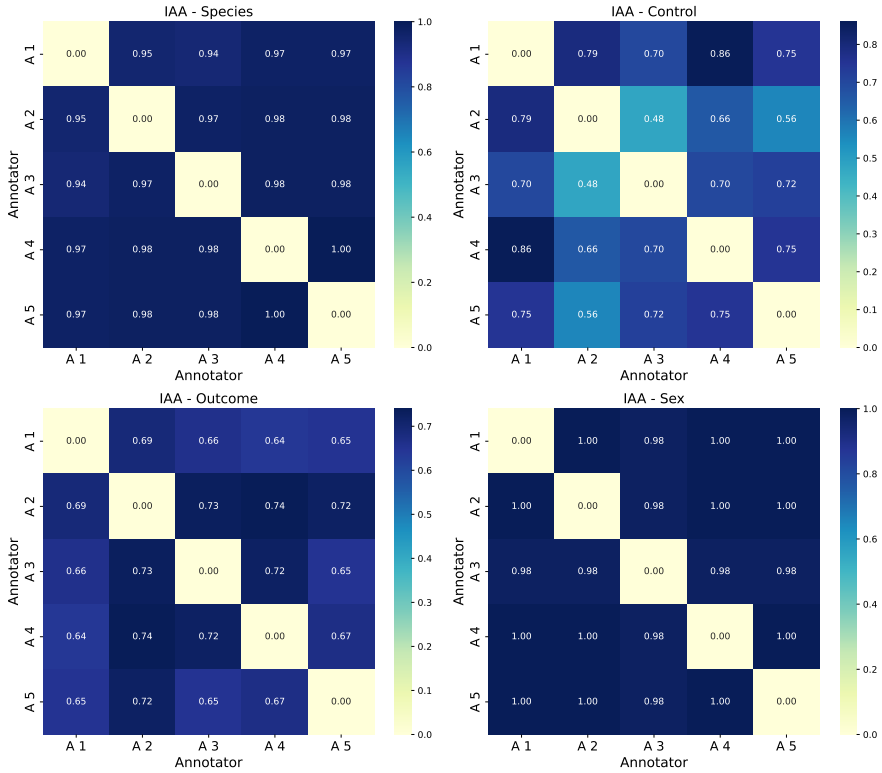

Figure 4: Cohen's Kappa scores for inter-annotator pairs for the categorical annotations in the overlapping articles.

B.3 Experimental Parameters

B.3.1 Animal species

Type: Population
Task: Text
Definition: The animal species used to test the intervention of interest.
Examples: Rats, mice, monkeys, rabbits, etc.
Location in Paper: Abstract, Methods
Comments:

- Most studies use rats or mice, while monkeys, pigs, cats, dogs, and rabbits are rarer.
- A study could use more than one species.

B.3.2 Animal strain

Type: Population
Task: Token
Definition: The animal strain further defining the animal species. A strain is a genetic variant, a subtype, or a culture within a biological species.
Examples: BALB/cJ (mouse), C57BL/6J (mouse), DBA/2J (mouse), Lewis (rat), Sprague-Dawley (rat).
Location in Paper: Abstract, Methods
Comments:

- A study could use more than one strain.
- Only annotate the strain (not the species).
- Be careful to separate strain from transgenic identification.

B.3.3 Animal sex

Type: Population
Task: Text
Definition: The animal sex further defining the animal species.
Examples: Male, female, both sexes.
Location in Paper: Abstract, Methods
Comments:

- A study could use either male, female, or both sexes.
- Some studies do not report the sex used.
- Only label the sex used to test the drug of interest.

B.3.4 Diseases mentioned

Type: Population
Task: Token
Definition: Diseases of interest related to the used animal model(s).
Examples: Multiple sclerosis, stroke, spinal cord

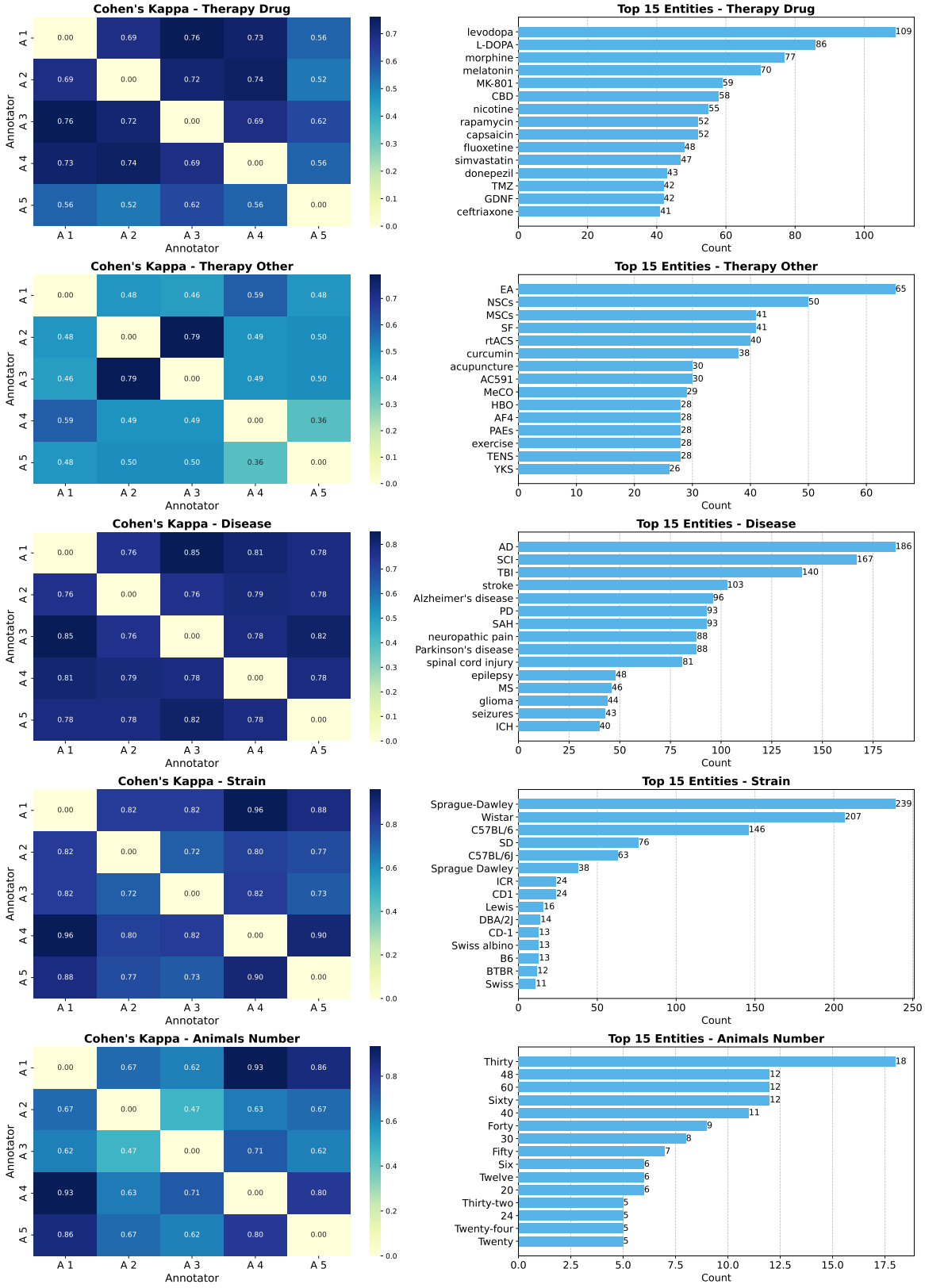

Figure 5: Cohen's Kappa scores for inter-annotator pairs for the NER annotations in the overlapping articles (left column). Top 15 most frequent NER entity text spans in the full dataset (right column).

injury, etc.

Location in Paper: Title, Abstract

Comments:

- Annotate only diseases relevant to the study.

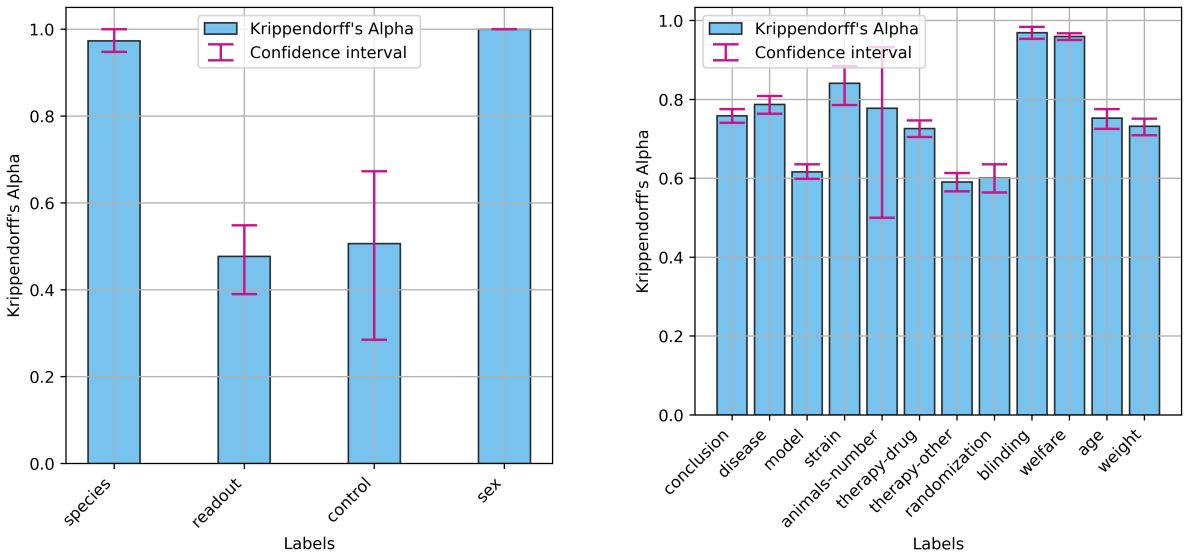

Figure 6: Krippendorff's alpha scores for different annotation levels. (Left) Krippendorff's alpha for document-level annotations. (Right) Krippendorff's alpha for token- and sentence-level annotations. Alpha score is computed on a per-token basis.

- Do not annotate disease models like MCAO or EAE.
- Include abbreviations (e.g., MS).
- Annotate more and less specific mentions (e.g., Alzheimer's disease and dementia).

B.3.5 Number of animals used in total

Type: Population
Task: Token
Definition: The total number of animals used in the study.
Examples: "A total of 968 animals (618 mice and 350 rats) were used."
Location in Paper: Abstract, Methods
Comments:

- Only annotate the exact total number.
- The number might be reported more than once.

B.3.6 Therapy

Type: Intervention
Task: Token
Definition: The therapeutic intervention tested.
Examples: Electroacupuncture, melatonin therapy.
Location in Paper: Title, Abstract, Methods
Comments:

- Two labels:
  - Drug (e.g., a small molecule, siRNA).
  - Non-drug (e.g., exercise, herbal extracts).
- Control treatments should be annotated.
- Do not annotate dosing or application information.

B.3.7 Control mentioned

Type: Control
Task: Text
Examples: "Group 1 (control) received saline."
Location in Paper: Abstract, Methods
Definition: Whether the control group/treatment was mentioned.
Levels:

- Control yes
- Control not reported

B.3.8 Readouts

Type: Outcome
Task: Text
Definition: The readouts used to assess intervention efficacy.
Examples: "We used Nissl staining and MRI to assess stroke volume."
Location in Paper: Abstract, Methods
Levels:

- Behavior (e.g., rotarod, seizure).
- Imaging (e.g., MRI, PET).

- Histology (e.g., Nissl staining, H&E).
- Physiology (e.g., blood pressure, EEG).
- Other (e.g., PCR, Western blot).

B.3.9 Study conclusion

Type: Outcome
Task: Sentence
Definition: The main finding of the study, i.e., the overall effect of the intervention.
Examples: "Our findings suggest a potential therapeutic role for Galantamine in attenuating hyperoxia-induced brain injury."
Location in Paper: Abstract
Levels:

- Positive
- Negative
- Neutral
- Mixed

B.3.10 Animal disease model

Type: Population
Task: Sentence
Definition: The animal model mimicking a neurological or psychiatric condition.
Examples: "EAE was induced by immunizing female Lewis rats with MOG55-66."
Location in Paper: Abstract

B.3.11 Animal age

Type: Population
Task: Sentence
Definition: The age of animals used.
Examples: "12-week-old female C57BL/6 mice were used."
Location in Paper: Methods

B.3.12 Animal weight

Type: Population
Task: Sentence
Definition: The weight of animals used.
Examples: "Male Albino Swiss (20–25 g) mice were used."
Location in Paper: Methods

B.4 Parameters Related to Study Quality

B.4.1 Randomization

Type: Study Quality
Task: Sentence
Definition: Whether the experimental setup used randomization of animals.
Examples: "We randomly divided the experimental rats into five groups with six animals per group as follows: . . . "
Location in Paper: Abstract, Methods
Comments:

- Only applies to circumstances describing the randomization of animals into (treatment) groups.
- Does NOT apply to other instances of randomization (e.g., “we analyzed 5 random fields of view”).
- In most cases, only one sentence describes randomization, but more than one could be annotated if different species are described separately.
- Together with blinding, it is one of the most critical study quality items.
- Annotate the entire sentence.
- If unsure, be generous with annotation.

B.4.2 Blinding

Type: Study Quality
Task: Sentence
Definition: Whether the experimental setup used blinding of experimenters.
Examples:

- "Experimenters were blinded to the treatment group."
- "Researchers were unaware of the treatment of the animals."
- "All behavioral measurements were made by an observer unaware of the treatment."

Location in Paper: Abstract, Methods
Comments:

- Blinding can occur at any step: during treatment, analysis, or both.
- Typically, only one sentence describes blinding, but multiple sentences may exist for different species/experiments.

- Together with randomization, it is a crucial study quality item.
- Annotate the entire sentence.
- If unsure, be generous with annotation.

B.4.3 Animal Welfare Statement

Type: Study Quality
Task: Sentence
Definition: Whether the animal study complies with local, regional, national, or international animal welfare guidelines.
Examples:

- "On October 29, 2019, the institutional Ethics Committee at NODCAR and Faculty of Pharmacy, Cairo University, approved all animal procedures."
- "All mice were maintained under specific pathogen-free conditions and used for experimentation according to protocols approved by the Swiss Federal Veterinary Office."

Location in Paper: Abstract, Methods
Comments:

- Usually, only one sentence describes animal welfare, but multiple sentences may exist for different species/experiments.
- Statements should mention compliance with guidelines/regulations or approval by an ethics committee.
- Commonly reported.

B.4.4 ARRIVE Guidelines

Type: Study Quality
Task: Sentence
Definition: Whether the study follows the ARRIVE guidelines, which provide standards for reporting methodological details in animal experiments.
Examples:

- "In the current study, we handled the animals consistently in accordance with the ARRIVE guidelines."
- "All studies involving animals are reported in accordance with the ARRIVE guidelines for reporting experiments involving animals."

Location in Paper: Abstract, Methods
Comments:

- Can be identified by searching for "ARRIVE" (always in uppercase).
- Rarely reported.

B.4.5 PREPARE Guidelines

Type: Study Quality
Task: Sentence
Definition: Whether the study follows the PREPARE guidelines.
Examples: Search for "PREPARE" in the document.
Location in Paper: Abstract, Methods
Comments:

- Can be identified by searching for "PREPARE" (always in uppercase).
- Very rarely reported.

B.4.6 Sample Size Calculation

Type: Study Quality
Task: Sentence
Definition: Whether the study conducted a prior sample size calculation to determine how many animals were required for the experiments.
Examples:

- "We conducted an a priori sample size calculation."
- "The power was calculated based on prior estimates."

Location in Paper: Abstract, Methods
Comments:

- Very rarely reported.