Unpacking Unstructured Data: A Pilot Study on Extracting Insights from Neuropathological Reports of Parkinson's Disease Patients using Large Language Models

Authors:

Oleg Stroganov[1#], Amber Schedlbauer[1], Emily Lorenzen[1#], Alex Kadhim[1], Anna Lobanova[1], David A. Lewis[2], Jill R. Glausier[2].

[1]Rancho BioSciences LLC, 16955 Via Del Campo #220, San Diego, CA 92127

[2] University of Pittsburgh, Department of Psychiatry

[#]email: oleg.stroganov @ranchobiosciences.com

ABSTRACT

Objective

The aim of this study was to make unstructured neuropathological data, located in the NeuroBioBank (NBB), follow FAIR principles, and investigate the potential of Large Language Models (LLMs) in wrangling unstructured neuropathological reports. By making the currently inconsistent and disparate data findable, our overarching goal was to enhance research output and speed.

Materials and Methods

The NBB catalog currently includes information from medical records, interview results, and neuropathological reports. These reports contain crucial information necessary for conducting in-depth analysis of NBB data but have multiple formats that vary across sites and change over time. In this study we focused on a subset of donors with Parkinson's Disease (PD). We developed a data model with combined Brain Region and Pathological Findings data at its core. This approach made it easier to build an extraction pipeline and was flexible enough to convert resulting data to Common Data Elements (CDEs) used by the community.

Results

This pilot study demonstrated the potential of LLMs in structuring unstructured neuropathological reports of PD patients available in the NBB. The pipeline enabled successful extraction of microscopic and macroscopic findings and staging information from pathology reports, with extraction quality comparable to results of manual curation. To our knowledge, this is the first attempt to automatically standardize neuropathological information at this scale. The collected data has the potential to serve as a valuable resource for PD researchers, bridging the gap between clinical information and genetic data, thereby facilitating a more comprehensive understanding of the disease.

INTRODUCTION

Effective data modeling of biological experiment data can have a major impact on downstream data usage, accessibility, and significantly improve research output. Having a robust and sound data model allows FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles to be employed and provide major benefits to research progress [1]. A recent cost-benefit analysis by the European commission on FAIR data suggests that not using FAIR data principles costs the European economy approximately €10.2 billion per year [2]. Thus, improving the quality of data models by applying these principles serves to save a considerable amount of time and resources, further advancing research efforts.

With the advent of high-performance computing and artificial intelligence (AI), technologies such as natural language processing (NLP) and large language models (LLMs) can be used to facilitate data FAIRification of unstructured data [3,4]. LLMs such as OpenAI's Generative Pre-trained Transformers (GPT) [5] can learn from the statistical associations between words in large online text databases to produce human-like text outputs [4]. In the context of biomedical research, LLMs are currently being explored as a means to extract data, identify patterns, and uncover insights that may have been previously hidden [6,7]. While the current gold-standard of data curation is to perform manual curation, this process is time intensive and can introduce errors. Therefore, LLMs may hold value in accelerating data curation and allowing FAIR data principles to be applied, ultimately improving research efficiency.

The NIH-funded NeuroBioBank (NBB, https://neurobiobank.nih.gov/) was established in September 2013 as a national resource for investigators utilizing post-mortem human brain tissue and related biospecimens for their research to understand conditions of the nervous system such as Alzheimer's Disease (AD), Parkinson's Disease (PD), frontotemporal dementia (FTD), and many others. The overall goals of the NBB are to 1) increase the availability of brain tissue from individuals affected and unaffected by brain disorders, 2) facilitate brain tissue distribution and 3) provide a central resource of best practices and protocols to the research community. Comprised of medical records, interview results, and neuropathological reports, the NBB catalog is an invaluable source of data for researchers. The catalog has information on clinical diagnosis, medical history, as well as results of whole genome sequencing. However, key pieces of data – specifically, the results of gross and microscopic examination of brain samples exist primarily as unstructured notes, often in the form of PDF pathological reports. The lack of standardization and inconsistent formats used across sites presents a significant data accessibility challenge, which hinders effective data usability and ultimately, research output. Converting these reports to a standardized format in accordance with FAIR principles would avoid duplicating efforts by different groups to extract data, accelerating research progress. Standard representation of pathology data would provide researchers with powerful tools to understand the mechanisms underlying the development of various pathologies, ultimately leading to improvements in the diagnostics and treatment of these debilitating conditions.

In this study, we investigated the potential of LLMs in unpacking unstructured neuropathological reports, with a focus on a subset of patients with PD. The goal of this work was to provide a framework to improve neuropathological findings, diagnosis, and staging. Our pilot consisted of 822 PD reports which spanned seven different sites and utilized 15 different formats. Reports were first preprocessed and converted to HTML file format, as they were provided in various formats. Information was then extracted from parsed reports using a questionnaire-based method that employed few-shot learning using the *gpt-3.5-turbo* model [8]. The extracted data was then combined, harmonized, and manually reviewed for accuracy.

MATERIALS AND METHODS

Source data

We used 822 PD reports generated from seven NBB sites: University of Maryland, University of Pittsburgh , National Institute of Mental Health (NIMH), University of Miami, Sepulveda, Harvard Brain Tissue Resource Center, and Mount Sinai/Bronx VA Medical Center. This represents approximately 5% of the total number of reports collected by the NBB. The sites provided reports in various file formats such as pdf, docx, or xlsx. Most of the reports contained sections outlining the specimen received, neuropathological diagnosis, macroscopic and microscopic pathological findings, and pathologist comments. Details and formats of the reports differed among the sites: for some sites, such as Maryland and Sepulveda, pathology descriptions were provided as narratives with sequential descriptions of findings. Other sites such as Harvard and Miami, grouped findings by brain regions. Notably, the Mount Sinai site utilized an electronic system to capture information, significantly streamlining the data collection process. In total, we compiled data from 822 neuropathological reports, spanning a 32-year period from 1990 to 2022. Report selection was based on the presence of a PD clinical diagnosis. No additional stratification based on age, gender, or disease stage was made, and all personally identifiable information was redacted by NBB staff.

To facilitate the development of an automatic extraction pipeline, we classified reports based on their formats and created a training set containing 65 reports, which included the most representative reports for each site and format. These 65 reports were manually curated in consultation with the NBB working group and were used as a "gold standard" to develop and improve the NLP pipeline.

Data preprocessing and parsing

Neuropathological reports in pdf file format were converted to HTML using the ABBYY FineReader Optical Character Recognition (OCR) tool (https://pdf.abbyy.com/). Similarly, docx reports were converted to HTML using the doc2html Python library (https://github.com/chadwickcole/doc2html). HTML was chosen as the target file format because it preserved information about text styles, which was utilized to mark the beginning of report sections. During preprocessing, the reports were split into sections such as gross pathology, microscopic findings, diagnosis, and sample information. For the reports where tissue information was available, an additional sub-rubric "tissue" was added. An output table was created containing report ID, section type, and section content. The data preprocessing and parsing stage enabled the conversion and organization of neuropathological reports into a structured and machine-readable format, laying the foundation for subsequent NLP pipeline development and data extraction.

Data model development

At the heart of the data model development, we focused on establishing a "region-finding-qualifier" triad. This approach contrasts with many Common Data Elements (CDEs), where the combination of region and finding represents an immutable data element. The rationale behind selecting the "region-finding-qualifier" triad approach was twofold. First, it allowed us to perform named entity recognition on the region, finding, and qualifier independently, which greatly facilitated the development of the NLP pipeline. Second, this approach enabled us to deal with a variety of report styles and older reports that might not capture brain features or findings according to modern standards.

To ensure consistency and standardization, we utilized external ontologies and controlled vocabularies for the model attributes. These sources included the Allen Human Brain Atlas [9], Systematized Nomenclature of Medicine (SNOMED) [10], National Cancer Institute Thesaurus (NCIT) [11], Disease Ontology (DOID) [12], Medical Subject Headings (MeSH) [13], Federal Interagency Traumatic Brain Injury Research (FITBIR) CDEs [14], and National Alzheimer's Coordinating Center (NACC) CDEs [15].

NLP pipeline and postprocessing

We utilized OpenAI GPT-3.5 model [8] as the primary LLM engine for our NLP pipeline development. No additional fine-tuning was performed, and the standard settings were employed. All calls for gpt-3.5-turbo were executed through the command line, as per the recommended guidelines using default model parameters (temperature 1, top P 1, frequency penalty 0, presence penalty 0). We used python for data extraction and R for data harmonization and QC.

Two approaches were adopted for data extraction from neuropathological reports. The first approach was questionnaire-based, in which the input text was directly mapped to the data model through a series of questions and coded answers. This method facilitated a structured approach to obtaining relevant information from the reports.

The second approach focused on the direct extraction of tissue-finding-qualifier triads from the text. However, this required a harmonization step, as the extracted tissue, finding, and qualifier terms were not directly mapped to our data model. To address this challenge, we performed manual harmonization with the assistance of an embedding-based classification technique, using the text-embedding-ada-002 model from OpenAI [16]. This approach allowed us to effectively categorize and map the raw extracted terms into a standardized format compatible with our data model.

Evaluation metrics

To evaluate the quality of data extraction, we compared the results obtained from the automatic extraction pipeline with the manually curated data ("gold standard") for the 65 selected reports. For macroscopic and microscopic findings, we assessed the agreement between the lists of brain regions identified manually and those extracted by the automatic pipeline. In cases where findings intersected, we compared the associated qualifiers. For regions with mismatches, we sampled and analyzed regions that were present exclusively in either the manual curation or the pipeline extraction. Upon examining these differences, we found that not all discrepancies were due to actual errors in the extraction process. Some variations could be attributed to ambiguities in the reports or harmonization efforts during the manual curation process. For neuropathological staging information, we compared the manually curated staging data with the staging information extracted by the automatic pipeline.

By conducting this thorough evaluation, we were able to assess the performance of the extraction pipeline and identify areas for improvement.

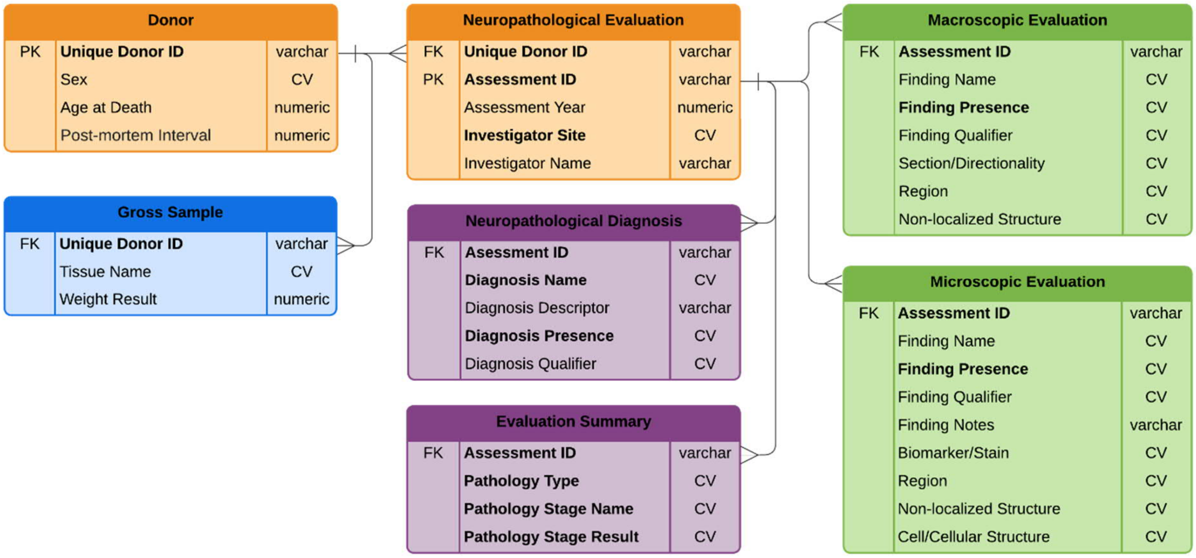
RESULTS

Data model

Overview

In this study, we have developed a comprehensive data model to represent and capture the relevant information from neuropathological reports. Key entities and attributes are shown on the Entity Relationship Diagram (ERD; Figure 1). Full ERD and tabular description of all entities and attributes are provided in Supplementary Information S1 and S2. The primary objective of this data model is to efficiently organize and store the extracted data from the pathology reports, facilitating easy access and analysis for researchers. The data model comprises several key entities, which are interconnected to represent the various aspects of the neuropathological findings. At the core of the data model is the Neuropathological Evaluation entity, which serves as a central hub linking the other entities. Additionally, this entity is connected to the Donor entity, enabling a clear association between the evaluation results and the corresponding donor.

The Neuropathological Evaluation entity is further linked to four main sub-entities: Evaluation Summary, Neuropathological Diagnosis, and Macroscopic and Microscopic Evaluations. The Evaluation Summary entity encompasses the staging information (e.g., Braak and Del Tredici stage for PD [17], ABC score according to NIA-AA 2012 consensus guidelines [18], and others [19–22]). The Neuropathological Diagnosis entity lists all neuropathological diagnoses identified in the report, serving as a comprehensive catalog of the patient's conditions.



Donor

PK	**Unique Donor ID**	varchar
	Sex	CV
	Age at Death	numeric
	Post-mortem Interval	numeric

Gross Sample

FK	**Unique Donor ID**	varchar
	Tissue Name	CV
	Weight Result	numeric

Neuropathological Evaluation

FK	**Unique Donor ID**	varchar
PK	**Assessment ID**	varchar
	Assessment Year	numeric
	Investigator Site	CV
	Investigator Name	varchar

Neuropathological Diagnosis

FK	**Assessment ID**	varchar
	Diagnosis Name	CV
	Diagnosis Descriptor	varchar
	Diagnosis Presence	CV
	Diagnosis Qualifier	CV

Evaluation Summary

FK	**Assessment ID**	varchar
	Pathology Type	CV
	Pathology Stage Name	CV
	Pathology Stage Result	CV

Macroscopic Evaluation

FK	**Assessment ID**	varchar
	Finding Name	CV
	Finding Presence	CV
	Finding Qualifier	CV
	Section/Directionality	CV
	Region	CV
	Non-localized Structure	CV

Microscopic Evaluation

FK	**Assessment ID**	varchar
	Finding Name	CV
	Finding Presence	CV
	Finding Qualifier	CV
	Finding Notes	varchar
	Biomarker/Stain	CV
	Region	CV
	Non-localized Structure	CV
	Cell/Cellular Structure	CV

Figure 1 – Entity Relationship Diagram for Neuropathological data model. Key entities and attributes are shown. Relationships between entities follow standard crow foot notation. Color scheme corresponds to conceptual subschemas: orange – donor conceptual subschema, blue – biological specimen conceptual subschema, purple – case diagnosis conceptual subschema, green – pathology case conceptual subschema. Attributes in bold are mandatory. PK – primary key, FK – foreign key, CV – attribute values taken from controlled vocabulary.

The Macroscopic and Microscopic Evaluations entities capture the detailed findings from the gross and microscopic examinations of the brain samples, respectively. These entities store both positive and negative findings, ensuring a complete representation of the pathological landscape. They are essential for understanding the specific pathological abnormalities present in the sample and their implications on the patient's condition.

Data dictionary

To accurately capture the anatomic location of pathology findings, we utilized the Allen Human Brain Atlas (AHBA) [9] as the foundation for our data model's representation of brain regions. The AHBA offers comprehensive coverage of brain structures; however, certain adjustments and extensions were necessary to address the specific needs of our study. Firstly, the AHBA does not encompass the vascular system. To address this, we added major arteries to the data dictionary and provided corresponding links to external ontologies such as MeSH [13] or UBERON [23]. Similarly, we included adjacent structures that are not part of the brain, such as the skull and scalp, with appropriate links to external ontologies.

Another challenge we encountered was the presence of hyperspecific and hypospecific regions in the neuropathological reports. Hyperspecific regions, such as the *CA1/CA2 junction* or *calvarial dura*, contain a level of detail absent in the AHBA. Conversely, hypospecific regions, such as the *visual cortex* or *olivary nucleus*, represent groups of brain regions that do not have a corresponding entity in the AHBA. In some cases, the reports used terminology for brain regions that only exist for non-human species, such as the *caudal medullary velum* (rat) or *occipital gyrus* (macaque). To address these issues, we added these regions to our data dictionary and, where possible, provided links to external ontologies and parent regions.

In instances where the report specified a particular part of a brain region, we captured this information using a combination of “Region” and “Section/Directionality”. For example, *posterior occipital cortex* was mapped to a combination of the *occipital cortex* region (AHBA id:3614) and *posterior* directionality. Lastly, we acknowledged that many reports did not associate specific regions with certain findings. For example, reports may mention finding in blood vessels, gray matter, or lesions, without indication of where exactly the finding is located. To accommodate these cases, we included a "Non-localized Structure" attribute in our data model.

In the development of our data model, we aimed to effectively capture and represent macroscopic and microscopic finding names and their associated qualifiers. Generally, finding names encompass descriptions of the observation, such as calcification, atrophy, necrosis, or abnormal coloration, while qualifiers provide optional details regarding severity, quantity, color, shape, and other properties of the findings. We used findings and qualifiers from existing neuropathology CDEs such as those supplied by FITBIR and NACC and extended the dictionary with information from reports. Our initial approach involved separating finding names and qualifiers into basic repeating elements to reduce the number of distinct values in the dictionary and streamline data extraction and harmonization. However, after consultations with the NBB working group, we made certain exceptions. For example, instead of separating *diffuse plaques* into the finding *plaque* and the qualifier *diffuse* we maintained it as a single finding. This decision was made to preserve the specificity and clarity of certain findings.

The final data model comprised 183 macroscopic findings, 416 microscopic findings, and 333 qualifiers. To ensure consistency and interoperability, we mapped the findings to established external ontologies, such as SNOMED and NCIT whenever possible.

Data Extraction Pipeline

Overall description

The overall data extraction pipeline (Figure 2) involves six crucial steps. In the first step, the neuropathological reports in pdf and docx file formats are converted into HTML file format, which efficiently preserves the styling information. This conversion allows for easier parsing and extraction of relevant data in subsequent steps. In the second step, the HTML documents are split into distinct sections, such as gross pathology, microscopic findings, diagnosis, and sample information. For reports containing available tissue information, an additional subcategory titled “tissue” is incorporated. An output table is generated, encompassing report ID, section type, and section content, which serves as a structured representation of the data extracted from the reports.

The third step involves feeding the machine-readable data from the output table into the data extraction process. Some information, such as brain weight, donor age, and sex, is already structured and can be effortlessly extracted using pattern matching. To extract most of the other information, we employed two approaches: a questionnaire-based method and a few-shot learning direct approach, both utilizing GPT-based models from OpenAI. In the fourth step, all extracted information is combined and reshaped to create a unified dataset. This dataset then undergoes a harmonization process in step five, where the data is mapped and aligned with the developed data model specifically tailored for neuropathological conditions.

The final step consists of assessing the quality of the extracted data. Values are compared to both the data model and manually curated data to ensure accuracy and consistency across the dataset. Any discrepancies or issues identified during the quality assessment are addressed to refine the data extraction pipeline further.

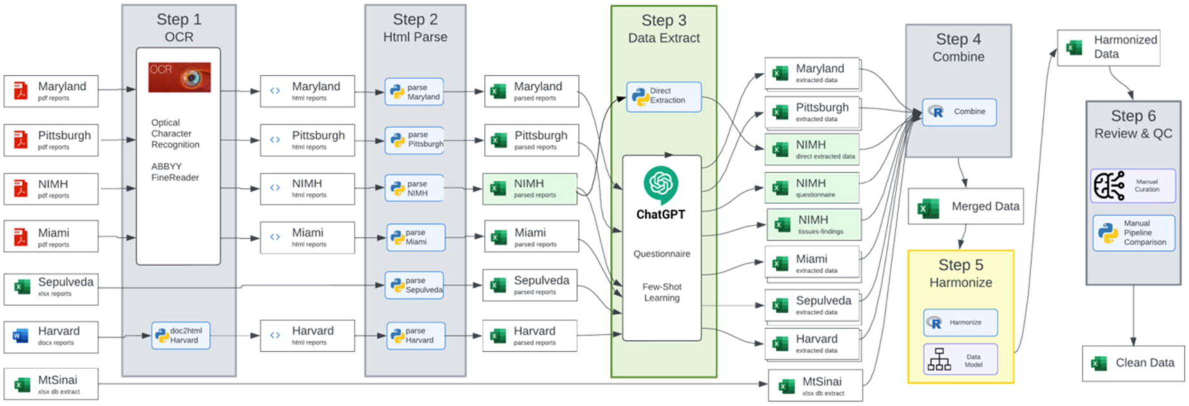
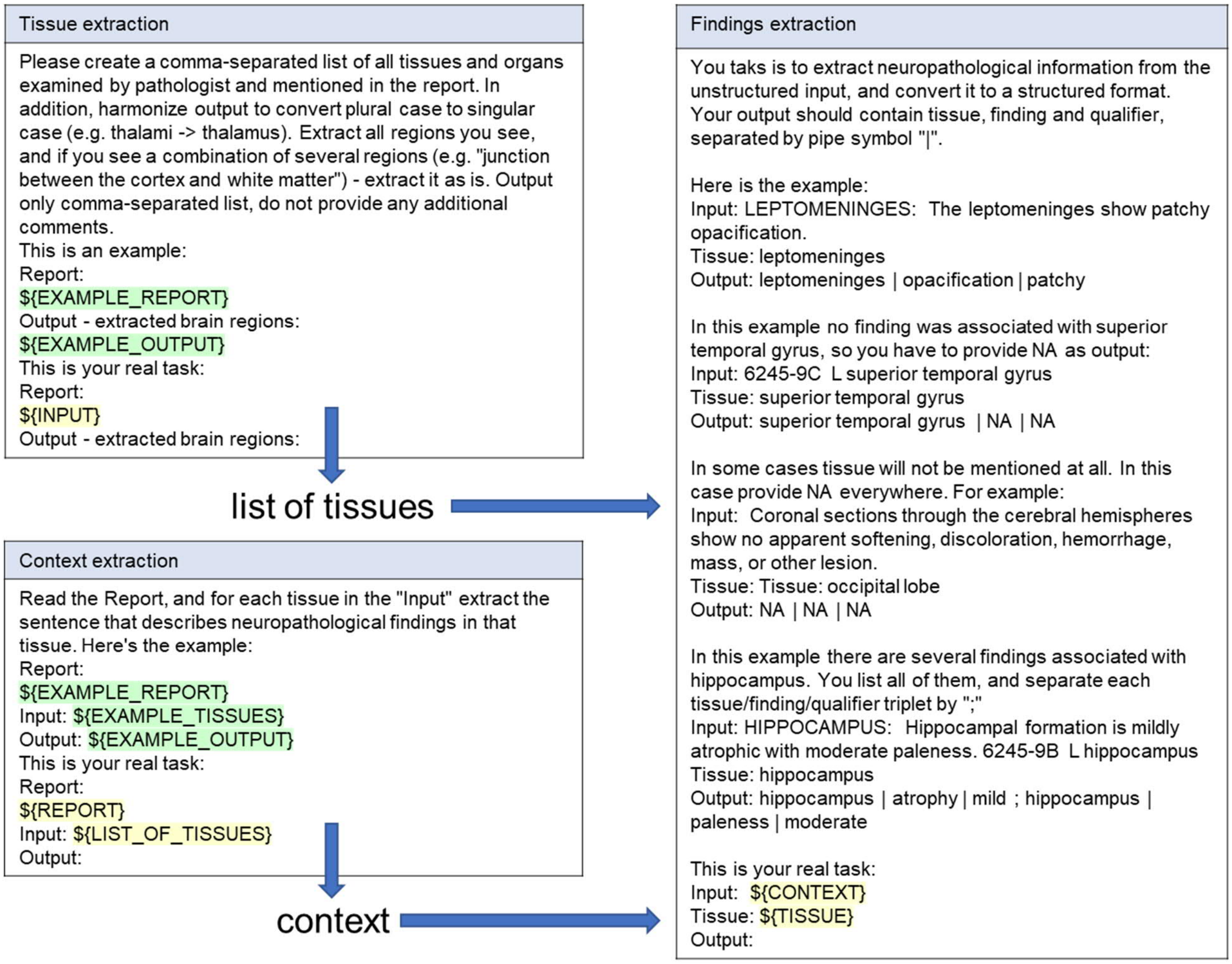


Figure 2 – Schematic representation of data extraction pipeline.

Direct approach for data extraction

The data extraction process adapted a direct approach of traditional Named Entity Recognition (NER) techniques. Classical NER methods, such as N-gram phonetic search [24], perform optimally when dictionaries are well-defined and comprehensive. However, in our case, we could not rely on the AHBA for region extraction, as it did not encompass all the regions we intended to extract. Furthermore, the dictionary for findings was non-existent, rendering NER tools unsuitable for extracting findings and qualifiers. Consequently, we employed LLMs in the initial step of data extraction to identify all mentioned brain regions. To improve recall, region extraction was executed twice, and tissue lists from both runs were consolidated. Subsequently, for each mentioned tissue, LLMs extracted associated findings and qualifiers. We guided the model using a few-shot learning approach, providing request and response examples to assist in handling complex or ambiguous cases.

Initially, the davinci-03 model was employed, which was not specifically fine-tuned for user requests. In the final version, we used the gpt-3.5-turbo model. Since gpt-3.5-turbo was trained to respond to direct user requests, we supplemented the prompts with explicit instructions regarding the desired output format. Examples and the overall protocol can be found in Figure 3.



Tissue extraction

Please create a comma-separated list of all tissues and organs examined by pathologist and mentioned in the report. In addition, harmonize output to convert plural case to singular case (e.g. thalami -> thalamus). Extract all regions you see, and if you see a combination of several regions (e.g. "junction between the cortex and white matter") - extract it as is. Output only comma-separated list, do not provide any additional comments.
This is an example:
Report:
${EXAMPLE_REPORT}
Output - extracted brain regions:
${EXAMPLE_OUTPUT}
This is your real task:
Report:
${INPUT}
Output - extracted brain regions:

Context extraction

Read the Report, and for each tissue in the "Input" extract the sentence that describes neuropathological findings in that tissue. Here's the example:
Report:
${EXAMPLE_REPORT}
Input: ${EXAMPLE_TISSUES}
Output: ${EXAMPLE_OUTPUT}
This is your real task:
Report:
${REPORT}
Input: ${LIST_OF_TISSUES}
Output:

Findings extraction

You taks is to extract neuropathological information from the unstructured input, and convert it to a structured format. Your output should contain tissue, finding and qualifier, separated by pipe symbol "|".

Here is the example:
Input: LEPTOMENINGES: The leptomeninges show patchy opacification.
Tissue: leptomeninges
Output: leptomeninges | opacification | patchy

In this example no finding was associated with superior temporal gyrus, so you have to provide NA as output:
Input: 6245-9C L superior temporal gyrus
Tissue: superior temporal gyrus
Output: superior temporal gyrus | NA | NA

In some cases tissue will not be mentioned at all. In this case provide NA everywhere. For example:
Input: Coronal sections through the cerebral hemispheres show no apparent softening, discoloration, hemorrhage, mass, or other lesion.
Tissue: Tissue: occipital lobe
Output: NA | NA | NA

In this example there are several findings associated with hippocampus. You list all of them, and separate each tissue/finding/qualifier triplet by ";"
Input: HIPPOCAMPUS: Hippocampal formation is mildly atrophic with moderate paleness. 6245-9B L hippocampus
Tissue: hippocampus
Output: hippocampus | atrophy | mild ; hippocampus | paleness | moderate

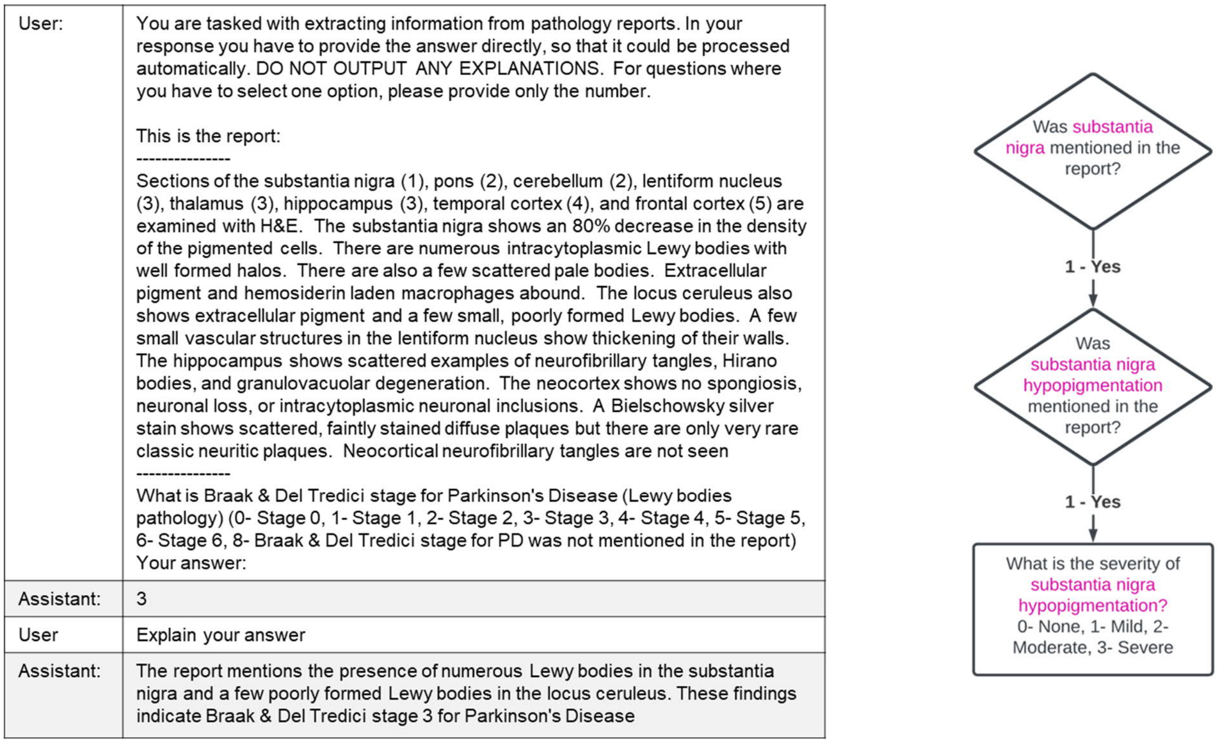
This is your real task:
Input: ${CONTEXT}
Tissue: ${TISSUE}
Output:

Figure 3. Workflow for direct data extraction. The $ designates text that is dynamically substituted by the data from reports, or results obtained on previous steps.

While the direct approach effectively captured information present in the reports, it necessitated subsequent harmonization, as the LLM was not cognizant of specific dictionaries employed in later stages. Moreover, the model did not consistently differentiate between finding names and qualifiers, resulting in discrepancies such as Finding Name *size decreased* versus the combination of Finding Name *size* and Finding Qualifier *decreased*. Therefore, a crucial harmonization step was essential to render the extracted information useful and consistent. Nevertheless, the developed few-shot learning approach successfully navigated ambiguous information and the absence of dictionaries, yielding semi-structured raw output that could be harmonized downstream.

Questionnaire-based approach for data extraction

The questionnaire-based approach emulates data abstraction by filling electronic forms. In this method, a series of questions are posed to the text, with coded answers provided. The goal is not to extract every piece of data but to focus on what is most important. We employed a similar approach where, instead of a human operator, it was the LLM that answered the questions. Questions were derived from the NACC questionnaire, which covers findings relevant to establishing diagnoses of AD and PD. In brief, the LLM was provided with instructions to answer the questions, a context (specific section of the report), and the questions themselves. These instructions or questions could be direct (e.g., "Provide the whole brain weight, in grams.") or dictionary-based (e.g., "What is the severity of cerebral cortex atrophy?" with coded answer options). Answers to dictionary-based questions were mapped directly to the data model, so no additional harmonization was required.



User:	You are tasked with extracting information from pathology reports. In your response you have to provide the answer directly, so that it could be processed automatically. DO NOT OUTPUT ANY EXPLANATIONS. For questions where you have to select one option, please provide only the number. This is the report: --------------- Sections of the substantia nigra (1), pons (2), cerebellum (2), lentiform nucleus (3), thalamus (3), hippocampus (3), temporal cortex (4), and frontal cortex (5) are examined with H&E. The substantia nigra shows an 80% decrease in the density of the pigmented cells. There are numerous intracytoplasmic Lewy bodies with well formed halos. There are also a few scattered pale bodies. Extracellular pigment and hemosiderin laden macrophages abound. The locus ceruleus also shows extracellular pigment and a few small, poorly formed Lewy bodies. A few small vascular structures in the lentiform nucleus show thickening of their walls. The hippocampus shows scattered examples of neurofibrillary tangles, Hirano bodies, and granulovacuolar degeneration. The neocortex shows no spongiosis, neuronal loss, or intracytoplasmic neuronal inclusions. A Bielschowsky silver stain shows scattered, faintly stained diffuse plaques but there are only very rare classic neuritic plaques. Neocortical neurofibrillary tangles are not seen --------------- What is Braak & Del Tredici stage for Parkinson's Disease (Lewy bodies pathology) (0- Stage 0, 1- Stage 1, 2- Stage 2, 3- Stage 3, 4- Stage 4, 5- Stage 5, 6- Stage 6, 8- Braak & Del Tredici stage for PD was not mentioned in the report) Your answer:
Assistant:	3
User	Explain your answer
Assistant:	The report mentions the presence of numerous Lewy bodies in the substantia nigra and a few poorly formed Lewy bodies in the locus ceruleus. These findings indicate Braak & Del Tredici stage 3 for Parkinson's Disease

Was substantia nigra mentioned in the report?
1 - Yes
Was substantia nigra hypopigmentation mentioned in the report?
1 - Yes
What is the severity of substantia nigra hypopigmentation? 0- None, 1- Mild, 2- Moderate, 3- Severe

Figure 4. An example of a typical questionnaire prompt with responses and follow-up. Flowchart of the questions aimed to discriminate between absence of information, negative findings, and positive findings.

Assessing data extraction quality for data extracted by the questionnaire revealed several failure modes. First, the LLM was not able to discriminate between cases where the region was not mentioned and cases where no abnormalities were found. Moreover, the LLM tended to deviate from direct answers and make conclusions that were close but not exactly answering the questions. For example, when asked about the Braak & Del Tredici stage of PD [17], the model sometimes attempted to interpret the extent of findings and give stage assessments, rather than reporting that the stage was not mentioned (Figure 4). To address these issues, we modified the flow of the questionnaire and added intermediate questions such as "Is X mentioned in the report? (0- No, 1- Yes)", "Is there evidence of X? (0- No, 1- Yes)", and only then "What is the severity of X?". This approach enabled us to distinguish more precisely between present and absent findings and reduce hallucinations.

Quality Assessment

Overall quality assessment results

To assess the quality of data extraction, we compared the data extracted from 65 reports by the pipeline to the data extracted through manual curation. For macroscopic and microscopic findings, we evaluated the agreement between the lists of brain regions identified manually and those extracted by the automatic pipeline. In cases where findings intersected, we compared the associated qualifiers. For regions with mismatches, we sampled and analyzed regions that were exclusively present in either the manual curation or pipeline extraction. For neuropathological staging information, we compared the manually curated staging data with the staging data extracted by the automatic pipeline. Overall metrics are reported in Table 1.

Table 1. Overall QC metrics for Macroscopic and Microscopic evaluation (comparison of sets of regions) and Evaluation summary (comparison of staging information). C and P denote information extracted by Curation and Pipeline respectively. C∩P denotes records intersection of sets of records, C∪P denotes union of sets.

	Jaccard Index (accuracy)	Sensitivity	Precision
	(C ∩ P) / (C ∪ P)	(C ∩ P) / C	(C ∩ P) / P
Macroscopic evaluation	64.6%	83.6%	73.9%
Microscopic evaluation	55.9%	76.5%	67.5%
Evaluation summary	64.6%	71.9%	86.4%

To further characterize problems with data extraction, we sampled up to 10 mismatches of both types (data present in manual curation only, data present in pipeline extraction only) for every site for regions in macroscopic evaluations. Mismatches were categorized based on the type of problem: issues with pipeline data extraction, issues with manual curation, or issues with harmonization. Then, types of issues were identified (Table 2).

Table 2 – Breakdown of comparison between manually extracted data and data extracted by pipeline by category for macroscopic evaluation. Error percentage is estimated from sampling of up to 20 mismatching records per site (10 records with curation only, 10 records with pipeline only region).

Category	Description	Estimated %
Match	Results of curation matches results of the pipeline	64.6%
Pipeline issues	Any issues due to incorrect or imprecise data extraction by pipeline. This includes the following sub-categories:	9.9%
Pipeline errors	Information was extracted by pipeline incorrectly	3.7%
Pipeline recovery	Brain regions are present in report but were not extracted by the pipeline	2.5%
Pipeline precision	Regions were extracted by pipeline with insufficient precision	1.9%
Pipeline missing/normal	Findings were extracted by pipeline as normal, but were missing in report or vice versa	1.9%
Curation issues	Any issues due to incorrect or imprecise data extraction by curators. This includes the following sub-categories:	9.6%
Curation recovery	Brain regions are present in report but were not extracted by curators	5.2%
Curation precision	Regions were extracted by curators with insufficient precision	4.4%

Harmonization issues	Mismatch due to information being extracted in different ways, or mapped to different terms	15.8%

Issues with the pipeline data extraction comprised 9.9% of the total amount of problems. In 9.6% of cases, information extracted by the pipeline was more accurate than information produced by curators. Additionally, 15.8% of mismatches were due to issues with harmonization, where both curation and pipeline approaches produced results that could be considered valid – the same data was interpreted in different ways. This analysis suggests that the total accuracy of data extraction for macroscopic findings surpasses 74.2% (the number of matching records plus the number of issues due to curation problems).

Pipeline data extraction drawback

We identified four types of pipeline errors. The most frequent cases involved incorrect data extraction (3.7%). In many instances, these errors resulted from the incorrect extraction of context. For example, the LLM accurately identifies "cerebral hemisphere" as one of the brain regions mentioned in the following extract:

"Coronal sections of the left hemisphere at the anterior frontal, striatal, and lentiform-thalamic-substantia nigra levels, and the midpons-cerebellum are examined. There is no softening, discoloration, hemorrhage, mass, or other lesion. Moderate cortical atrophy is seen in the frontal and temporal lobes."

The model then proceeds to extract the following context: *Coronal sections of the left hemisphere at the anterior frontal, striatal, and lentiform-thalamic-substantia nigra levels, and the midpons-cerebellum are examined. Moderate cortical atrophy is seen in the frontal and temporal lobes.*", which is then converted into a region, finding, and qualifier combination of "*cerebral hemisphere, atrophy, moderate*". This is less precise than the "*frontal cortex, atrophy, moderate*" extracted by curators.

In 2.5% of cases, the pipeline did not extract region information. Many of these errors occurred when the tissue was correctly identified, but the pipeline had trouble producing results in the desired format. For example, "*The lateral cerebral ventricle is normal in size and shape.*" Was incorrectly converted to "lateral cerebral ventricle | normal size; normal shape | NA". In the correct format, region/finding/qualifier triplets should be separated by semicolons, and the correct output should look like "lateral cerebral ventricle | normal size | NA; lateral cerebral ventricle | normal shape | NA".

Imprecise extraction accounted for 1.9% of the pipeline extraction problems. Typical examples included extraction of "cerebral cortex" without specifying the pathology location in more detail. In many cases, the pipeline extracted both general regions (*cerebral cortex*) and specific regions (e.g., *frontal cortex*). Additionally, some questions from the questionnaire asked about the presence of pathology in general regions and were therefore mapped to these general areas. Moreover, the coded list of answers forced the LLM to perform qualifier mapping. For instance, "*minimal*" was typically mapped to "*mild*", whereas "*mild to moderate*" could be mapped to either "*mild*" or "*moderate*," without a systematic approach to the mapping process. Similarly, pathological evaluations often contained ranges of stages (e.g., Braak II-III). The LLM randomly collapsed the range to one of the stages.

Lastly, 1.9% of pipeline extraction problems stemmed from confusion between findings that were not reported and cases where no abnormalities were found in specific regions. The LLM would easily become confused when the report contained a full list of sections and tissues that were analyzed, assuming that if a tissue was reported but no abnormality was explicitly mentioned, the tissue was normal. Initially, the percentage of these errors was much higher, so we had to remove the list of regions at the preprocessing stage whenever it was possible to identify such a list using style markers or headers.

Harmonization issues and problems with manual curation

Harmonization issues accounted for the largest portion (15.8%) of mismatches between manually curated data and data extracted by the pipeline. Common examples of these issues include discrepancies in naming general regions. For instance, "*cerebrum*" is often interchanged with "*cerebral hemispheres*" by both the pipeline and manual curation.

Another example illustrating the differing approaches between curators and the pipeline can be seen in cases where a list of tissues is reported. For example, a report might contain the following information: "*Neostriatum*: (*caudate nucleus*, *putamen*, and *nucleus accumbens*): Unremarkable." Curators interpreted this as "*neostriatum*, *unremarkable*," whereas the pipeline extracted all four mentioned tissues separately (*neostriatum*, *caudate nucleus*, *putamen*, and *nucleus accumbens*) and reported all of them as unremarkable.

In some cases, harmonization problems arose from instances where a finding or region could be represented in different ways, such as "*junction between cortex and white matter*" versus a combination of "*cortex*" and "*white matter*", or "*tonsillar herniation*" versus a combination of "*cerebellar tonsil*" and "*herniation*".

Lastly, a significant portion of mismatches (9.6%) could be attributed to either imprecise extraction by curators (5.2%) or information not being extracted at all (4.4%). These issues are more prevalent for sites that provide information-rich reports (e.g., Harvard and Miami), as the sheer amount of information in these reports makes manual curation more prone to errors.

DISCUSSION

In this study, we explored the capabilities of LLMs, such as GPT-based models, for data extraction from neuropathological reports. The first step for successful data extraction involves building a flexible data model that can accommodate a variety of data while facilitating extraction and subsequent harmonization. We achieved this by centering our model on regions, findings, and qualifiers, each of which can be varied independently. This approach deviates from the approach used to build CDEs, in which the central conceptual entity consists of predefined sets of regions and findings. While the CDE approach is more suitable for standardizing crucial information in electronic form, it is less flexible and less suitable for automated data extraction. Moreover, different electronic systems may contain sets of CDEs which are not compatible. By deconstructing CDEs into triads of regions, findings, and qualifiers, data becomes more flexible and can be easily converted, as demonstrated in our approach where we merged data from electronic system used to capture information in Mount Sinai with data extracted from pathology reports from other sites.

This study has demonstrated the significant potential of LLMs in structuring unstructured neuropathological reports. LLMs can handle ambiguous information and the absence of predefined dictionaries, which is essential when dealing with diverse and complex data. The reasoning capabilities of LLMs make it possible to extract complicated relationships and infer information that would be unreachable for standard methods. In certain cases where the amount of information is overwhelming, LLMs outperform manual data extraction and retrieve information more reliably.

Despite the benefits of LLMs, there are limitations to their application in data extraction from neuropathological reports. For instance, LLMs may struggle to differentiate between cases where a region is not mentioned and cases where no abnormalities were observed. Additionally, LLMs may inaccurately extract context or become confused when dealing with complex report structures. These limitations can lead to discrepancies between manually curated data and data extracted by the pipeline. Furthermore, LLMs may occasionally deviate from direct answers or make conclusions that are close but not exactly answering the questions.

The present study serves as a pilot effort, focusing on donors with PD, which represents less than 5% of the total number of pathology reports in the NBB. Although a significant portion of donors with PD have other comorbidities (e.g., AD and Cerebrovascular Disease), extending the pipeline to work with other patients would necessitate modifications to both the data model and the extraction workflow. For instance, entities describing brain tumors or specific findings related to traumatic brain injury could be incorporated into the model. The data dictionary could be expanded to encompass findings more characteristic of other pathologies, and corresponding evaluation staging information should be included. Moreover, certain sites (Pittsburgh and NIMH) were sparsely represented in the PD datasets we examined, and integrating additional reports from these sites might require modifications to the preprocessing algorithm. Additionally, the NACC questionnaire used in the data extraction pipeline should be extended with questions relevant to all types of pathologies present in the broader population of reports.

While the current data extraction approach achieved 74.2% accuracy and was comparable with manual curation results, we suggest several avenues for further improvement and scaling up. The LLM used in this study, GPT-3.5, has been superseded by a more powerful model, GPT-4 [25]. Another approach may involve fine-tuning the LLM to better understand the specific domain of neuropathology and further adapting it to handle complex and diverse data formats. Additionally, incorporating domain knowledge and expert feedback into the

LLM training process could enhance the model's ability to accurately interpret and extract information from pathology reports.

A significant portion of time was devoted to manual harmonization of data and mapping from LLM output to data dictionaries. For some attributes such as brain regions and qualifiers, we expect the harmonization efforts to scale sub-linearly, as the current set of reports already covered a significant variety of values. Other attributes, such as finding names, might still require substantial efforts to harmonize, as the data is dependent on the type of pathology. LLMs and related NLP capabilities, such as using embeddings for mapping between information provided by LLMs and predefined ontologies, might be employed to expedite the harmonization process.

CONCLUSIONS

In this study, we have developed a data model and data extraction pipeline that leverages LLMs to structure unstructured neuropathological reports from the NBB, specifically focusing on PD donors. To our knowledge, this is the first attempt to automatically standardize neuropathological information at this scale. The pipeline and data model can be repurposed and extended to accommodate other pathological conditions, making it a versatile tool for researchers. Furthermore, the collected data has the potential to serve as a valuable resource for PD researchers, bridging the gap between clinical information and genetic data, and thereby facilitating a more comprehensive understanding of the disease.

REFERENCES

1. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018. doi:10.1038/sdata.2016.18

2. Cost of not having FAIR research data. Cost-Benefit analysis for FAIR research data. European Commission, Directorate-General for Research and Innovation; 2018. Available: http://publications.europa.eu/resource/cellar/d375368c-1a0a-11e9-8d04-01aa75ed71a1.0001.01/DOC_1

3. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. Nat Rev Phys. 2023;5: 277–280. doi:10.1038/s42254-023-00581-4

4. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5: 194. doi:10.1038/s41746-022-00742-2

5. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv; 2020. Available: http://arxiv.org/abs/2005.14165

6. ChatGPT: what does it mean for scientific research and publishing? BJU Int. 2023;131: 381–382. doi:10.1111/bju.15995

7. Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-tuning large neural language models for biomedical natural language processing. Patterns N Y N. 2023;4: 100729. doi:10.1016/j.patter.2023.100729

8. Introducing ChatGPT and Whisper APIs. [cited 14 Jul 2023]. Available: https://openai.com/blog/introducing-chatgpt-and-whisper-apis

9. Allen Human Brain Atlas. [cited 14 Jul 2023]. Available: http://braininfo.org/

10. Systematized Nomenclature of Medicine. 14 Jul 2023. Available: https://www.snomed.org/

11. National Cancer Institute Thesaurus. [cited 14 Jul 2023]. Available: https://ncithesaurus.nci.nih.gov/ncitbrowser/

12. Disease Ontology. [cited 14 Jul 2023]. Available: https://disease-ontology.org/

13. Medical Subject Headings. [cited 14 Jul 2023]. Available: https://meshb.nlm.nih.gov/

14. Chronic TBI-related Neurodegenerated Codes. [cited 14 Jul 2023]. Available: https://fitbir.nih.gov/chronic-tbi-related-neurodegeneration-cdes

15. National Alzheimer's Coordinating Center. [cited 14 Jul 2023]. Available: https://naccdata.org/

16. New and improved embedding model. 15 Dec 2022 [cited 14 Jul 2023]. Available: https://openai.com/blog/new-and-improved-embedding-model

17. Braak H, Del Tredici K, Rüb U, de Vos RAI, Jansen Steur ENH, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging. 2003;24: 197–211. doi:10.1016/s0197-4580(02)00065-9

18. Hyman BT, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Carrillo MC, et al. National Institute on Aging-Alzheimer’s Association guidelines for the neuropathologic assessment of Alzheimer’s disease. Alzheimers Dement J Alzheimers Assoc. 2012;8: 1–13. doi:10.1016/j.jalz.2011.10.007

19. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol (Berl). 1991;82: 239–259. doi:10.1007/BF00308809

20. Thal DR, Rüb U, Orantes M, Braak H. Phases of A beta-deposition in the human brain and its relevance for the development of AD. Neurology. 2002;58: 1791–1800. doi:10.1212/wnl.58.12.1791

21. Vonsattel JP, Myers RH, Stevens TJ, Ferrante RJ, Bird ED, Richardson EP. Neuropathological classification of Huntington’s disease. J Neuropathol Exp Neurol. 1985;44: 559–577. doi:10.1097/00005072-198511000-00003

22. McKeith IG, Boeve BF, Dickson DW, Halliday G, Taylor J-P, Weintraub D, et al. Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. Neurology. 2017;89: 88–100. doi:10.1212/WNL.0000000000004058

23. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012;13: R5. doi:10.1186/gb-2012-13-1-r5

24. Jelinek F, Merialdo B, Roukos S, Strauss M. A dynamic language model for speech recognition. Proceedings of the workshop on Speech and Natural Language - HLT ’91. Pacific Grove, California: Association for Computational Linguistics; 1991. pp. 293–295. doi:10.3115/112405.112464

25. OpenAI. GPT-4 Technical Report. arXiv; 2023. Available: http://arxiv.org/abs/2303.08774

ACKNOWLEDGEMENTS

Data dictionary developed in this study was built using the controlled access datasets distributed from the DOD- and NIH-supported Federal Interagency Traumatic Brain Injury Research (FITBIR) Informatics Systems. FITBIR is a collaborative biomedical informatics system created by the Department of Defense and the National Institutes of Health to provide a national resource to support and accelerate research in TBI.

Brain pathology reports were provided by NIH NeuroBioBank, including Mount Sinai Neurobiobank, University of Miami Brain Endowment Bank, University of Maryland Brain and Tissue Bank, Harvard Brain Tissue Resource Center, The Human Brain and Spinal Fluid Resource Center, Brain Tissue Donation Program at the University of Pittsburgh, National Institute of Mental Health's Human Brain Collection Core.

Authors would like to thank Dr. Harry Haroutunian, Dr. William Scott, Dr. Derek Oakley, Dr. Pavan Auluck, Dr. Stefano Marenco, and NIH BioBank directors for their advice in development of the data model.

FUNDING

The work described herein was created with funding by NINDS (Contract: HHSN316201200128W-75N95022F00001)

SUPPORTING INFORMATION

S1 – Neurological Data Model – entities and attributes specification (in Excel file format)

S2 – Entity Relationship Diagram for Neurological Data Model