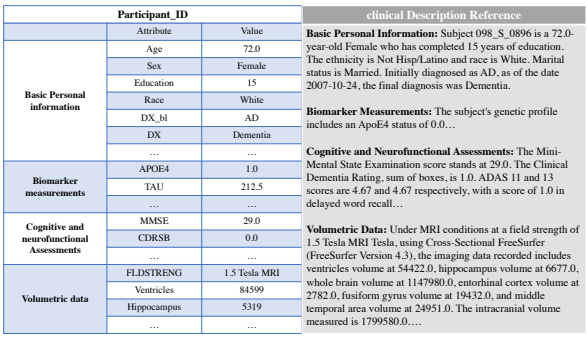
Enhancing Alzheimer's Disease Diagnosis Records with Large Language Models: A Pipeline for Multimodal and Longitudinal EHRs

Anonymous ACL submission

Abstract

Alzheimer's disease (AD) is a neurodegenerative, incurable condition and a leading cause of morbidity among individuals over 65 in the US. The screening and early diagnosis of AD condition is usually based on the patient's electrical health records (HERs), including clinical observations, cognitive tests, patient profiles, and medical-imaging-aided diagnoses. However, above information for researchers is highly fragmented. One of the most critical clinical diagnostic notes is stored in structured tables using specialized terminological formats. This presents significant challenges to the accessibility and readability for non-experts, thereby hindering information processing and the development of general medical AI systems. This work proposes a novel pipeline for processing AD clinical diagnostic information: (1) we collect clinical data from the largest AD dataset of Alzheimer's Disease Neuroimaging Initiative (ADNI)[1], explain abbreviations and terminology, and organize the information in an accessible manner for those without expert knowledge of AD. (2) Leveraging the power of Large Language Models (LLMs), we present a GPT-based method that effectively transforms tabular clinical data into fluent and faithful natural language diagnostic reports, as demonstrated by our experimental results. (3) We further explore the inherently multi-modal nature of medical data, collecting and processing a total of 10387 volumetric T1-weighted MRI scans from ADNI. (4) Finally, we discuss the existing limitations in applying multimodality EHRs for brain disease analysis and propose forward-looking directions to meet the demands of the neuroimaging domain. We expect that this work will provide new insights into the neuroimaging domain and improve AI applications in healthcare. Our code is available at https://anonymous.4open.science/r/ADNI-table-to-Text-2EDB

[1] https://adni.loni.usc.edu/



Participant_ID		
	Attribute	Value
Basic Personal information	Age	72.0
	Sex	Female
	Education	15
	Race	White
	DX_bl	AD
	DX	Dementia
	…	…
Biomarker measurements	APOE4	1.0
	TAU	212.5
	…	…
Cognitive and neurofunctional Assessments	MMSE	29.0
	CDRSB	0.0
	…	…
Volumetric data	FLDSTRENG	1.5 Tesla MRI
	Ventricles	84599
	Hippocampus	5319
	…	…

clinical Description Reference

Basic Personal Information: Subject 098_S_0896 is a 72.0-year-old Female who has completed 15 years of education. The ethnicity is Not Hisp/Latino and race is White. Marital status is Married. Initially diagnosed as AD, as of the date 2007-10-24, the final diagnosis was Dementia.

Biomarker Measurements: The subject's genetic profile includes an ApoE4 status of 0.0…

Cognitive and Neurofunctional Assessments: The Mini-Mental State Examination score stands at 29.0. The Clinical Dementia Rating, sum of boxes, is 1.0. ADAS 11 and 13 scores are 4.67 and 4.67 respectively, with a score of 1.0 in delayed word recall…

Volumetric Data: Under MRI conditions at a field strength of 1.5 Tesla MRI Tesla, using Cross-Sectional FreeSurfer (FreeSurfer Version 4.3), the imaging data recorded includes ventricles volume at 54422.0, hippocampus volume at 6677.0, whole brain volume at 1147980.0, entorhinal cortex volume at 2782.0, fusiform gyrus volume at 19432.0, and middle temporal area volume at 24951.0. The intracranial volume measured is 1799580.0….

Figure 1: An example of a clinical notes

1 Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by a continuum that ranges from undetectable brain changes to significant alterations that impair memory, culminating in severe physical disabilities(Aramadaka et al., 2023). Current research suggests that the clinical stages of AD can be divided into three stages(Frisoni et al., 2010; Jack et al., 2010). First is the pre-symptomatic phase, where individuals are cognitively normal despite having AD pathological changes. The second phase is the prodromal phase, often referred to as mild cognitive impairment (MCI), characterized by the onset of early cognitive symptoms. The third phase is dementia, marked by severe impairments across multiple domains, leading to loss of function(Yu et al., 2023). The pathological changes in the pre-symptomatic phase may begin to develop decades before the earliest clinical symptoms appear in the prodromal phase. While there is currently no cure for AD once established, in the realm of machine learning and deep learning, differentiating AD from the prodromal stage is a significant issue that interests a large amount of researchers making efforts on(Miller et al., 2022; Zhou et al., 2023b; Cai et al., 2023; Feng et al., 2023; Leng et al., 2023; Liang et al., 2023; Zhou et al., 2023a; Zhang et al., 2023).

Previous studies have predominantly focused on analyzing medical brain images like Magnetic Resonance Imaging (MRI) scans and Position Emission Tomography (PET) solely, extracting various features and designing algorithms to diagnose AD and MCI. The recently revised clinical criteria for detecting AD emphasize the importance of using multiple modalities such as core biomarkers and clinical tests for diagnosis(Carrillo and Masliah, 2023). However, much of this information has not yet been incorporated into brain disease studies. Thanks to the widespread implementation of EHRs, a rich array of routine clinical reports about participants are digitally available in a tabular form and offered by the Alzheimer's Disease Neuroimaging Initiative (ADNI)(Jack Jr et al., 2008). These tabular datasets offer rich details including demographic attributes, biomarker measurements, cognitive and neurofunctional assessments, and image information, making it an invaluable resource.

Despite their potential, the complexity of medical terminology often makes EHRs less accessible to non-professionals, limiting their utility. While ADNI ERHs provide complementary information except for medical scans for understanding disease patterns, for example, early researches report that age is the single most significant risk factor for AD and a higher incidence of MCI occurs in females than in males(Aramadaka et al., 2023; Katz et al., 2012), there is an urgent need to develop methodologies that can simplify and harness this wealth of information.

As an essential and widely used data format, tabular data plays a crucial role in structuring information, significantly advancing the fields of information retrieval and knowledge extraction for data mining(Engelmann et al., 2023; 10., 2022; Sui et al., 2023). Recently, LLM based on the Transformer architecture(Vaswani et al., 2017) have become increasingly prominent in addressing a wide range of natural language-related tasks. Given their state-of-the-art performance and board application potential, researchers are also developing LLMs for tabular data processing. Notable examples include TableGPT(Gong et al., 2020) and TABT5(Andrejczuk et al., 2022), which are trained for complex table-to-text generation tasks. Additionally, MediTabWang(Wang et al., 2023) has been specially developed for medical table prediction, showcasing the versatility of LLMs in handling structured data formats. Above successful large language models raise our beliefs that the application value of tabular structured EHRs can be extensively mined and analyzed, offering deep clinical insights into patient care.

Furthermore, medical data is inherently multimodal, making vision-and-language pre-training (VLP) a crucial technique for jointly understanding medical images and texts. Techniques like the Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) treat the text caption as a linguistic representation of the image, pulling the image-text pair closer together in a latent space, have proven effective in aligning these pairs and showing significant benefits in various downstream applications in the medical field, including image generation, segmentation, detection, and classification. Although large-scale image-text pair datasets are essential in medical domain research, however, existing large-scale datasets such as ROCO(Petersen et al., 2010), MedICat(Subramanian et al., 2020), FFA-IR(Li et al., 2021), and MIMIC-CXR(Johnson et al., 2019) and existing medical VLP(Tiu et al., 2022; Huang et al., 2021; Seibold et al., 2022; Pellegrini et al., 2023) are predominantly focus on 2D images. While 2D slices can be adequate for learning, 3D brain scans are crucial for brain disease diagnosis. Currently, there is no mature image-text pairing methodology for brain disease analysis. This paper aims to address this gap by introducing a novel pipline and establishing a Multimodality and Longitudinal dataset for neuroimaging studies. Specifically, our contributions are as follows:

- We collect comprehensive patient tabular data from ADNI, which includes demographic information, biomarker measurements, cognitive assessments, and volumetric data. By analyzing this specialized clinical knowledge, we developed an annotation method that transforms tabular data into information-rich and fluent natural language.

- We present an encoder-decoder based model ADTabGPT(Medical tabular-to-text Generative pre-trained transformers) that can be applied to ADNI table-to-text generation. Meanwhile, we collect and process a total of 11107 volumetric T1-weighted MRI scans from ADNI as the image feature.

- To our best knowledge, this is the first work to explore the full utilization of multimodal and longitudinal EHRs for brain disease analysis.

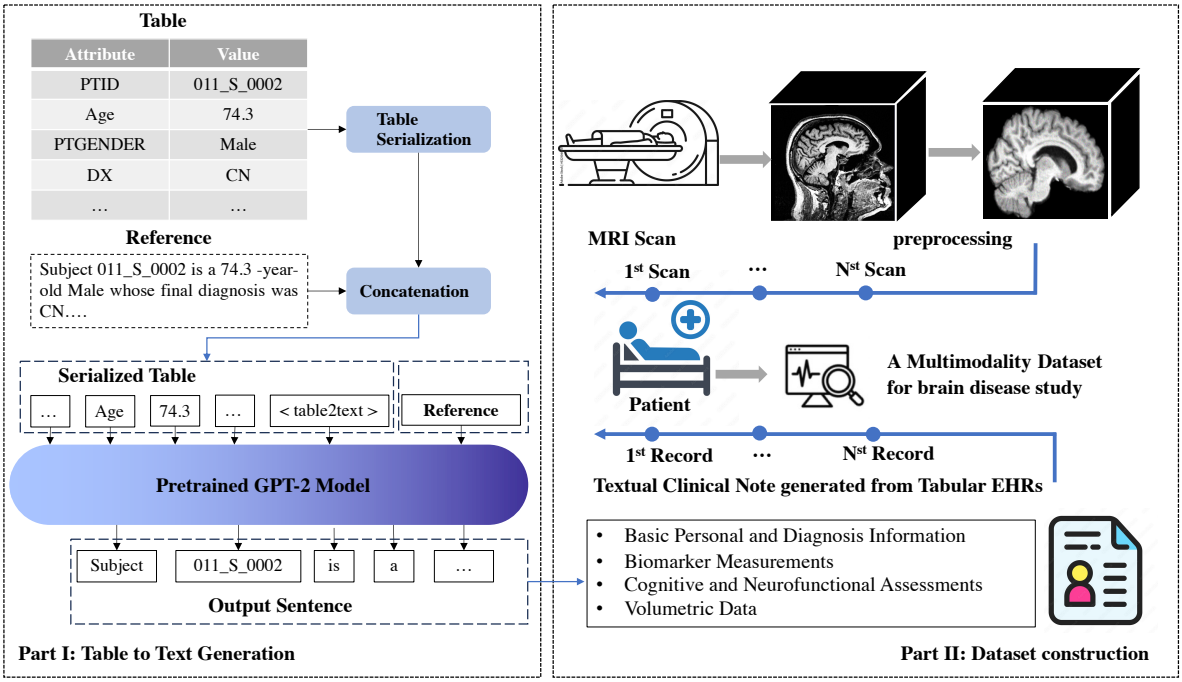


Figure 2: The overall architecture for constructing a multimodal and longitudinal Alzheimer's Disease-related neuroimaging dataset. It encompasses two major parts: **Part I** transforms tabular clinical notes into natural language clinical reports; **Part II** preprocesses the corresponding MRI scans. Ultimately, these two modalities are combined.

2 Data Collection and Analysis

We design a four-step annotation process, as depicted in figure 2. Initially, we collect tabular EHRs and analysis the key variables, and then design ADTabGPT to convert them into natural language clinical reports, as illustrated in part I of Figure 2. Next, as shown in part II of Figure 2, we collect the corresponding MRI scans of patients from ADNI and undertake a series of preprocessing steps. Following above steps, we finally construct a multimodal and longitudinal neuroimaging EHR especially for Alzheimer's disease.

2.1 Data Preparation

The ADNI is a consortium of universities and medical centers in the United States and Canada. It was established to develop standardized imaging techniques and biomarker procedures in Cognitive Normal(CN), MCI, and AD subjects. One of the major goals of ADNI is to create an accessible data repository that acquires clinical, cognitive, and biochemical data(Petersen et al., 2010). In addition to MRI imaging information, the ADNI provides an ocean of tabular data, allowing researchers and clinicians to explore the details such as demographic information(e.g., age, sex, educational level, etc); biomarker measurements (e.g., allele producing the apolipoprotein E [APOE-ε4]); cognitive assessments (e.g., Mini-Mental State Examination, Clinical Dementia Rating Scale, etc); as well as volumetric data (e.g., metrics of the WholeBrain and Hippocampus).

This paper primarily focuses on the *ADNIMERGE Table*, which merges several key variables from various case report forms and biomarker laboratory summaries across the ADNI protocols. The table encompasses 16,421 diagnostic records from 2,430 subjects, of which 2,409 have a diagnosis. the average age is 73.2 ± 7.0 years, and the average educational attainment is 16.1 ± 2.8 years. The data includes 4,020 diagnoses of CN individuals, 4,989 diagnoses of MCI, and 4,989 diagnoses of AD. The data comprises samples from three major racial groups: 2,143 White, 179 Black, and 58 Asian patients. Additionally, 50 samples include multiple racial backgrounds or are categorized as Unknown; American Indian/Alaskan Native; or Hawaiian/Other Pacific Islander. Gender-wise, females represent 52.4% of the subjects, totaling 1273, with the remainder being males. The ethnic distribution includes 2,302 non-Hispanic/Latino, 116 Hispanic/Latino, and 12 Unknown. From a

	CN group	MCI group	AD group
Gender (F/M)	475/354	324/448	345/463
Age(year)	71.4±6.8	73.0±7.5	74.4±7.5
Education(year)	16.6±2.5	15.9±2.8	15.6±2.8
APOE-$\varepsilon 4_{0/1/2}$	511/202/20	411/226/56	374/274/134

Table 1: Basic demographic and biomarker statistics.

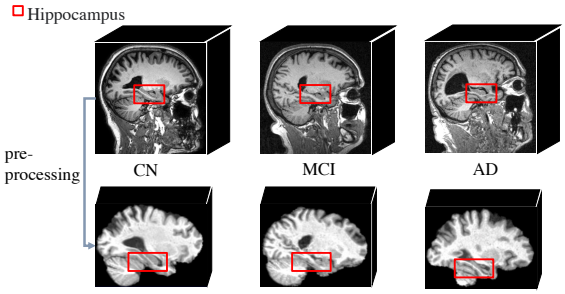


Figure 3: Brain morphology (hippocampus volume) contrast vary as a function of Alzheimer's disease. The upper row are original MRI images and the lower are pre-processed(brain extraction and registration) images.

marital status perspective, there are 1,824 married, 269 widowed, 228 divorced, 99 never married, and 9 unspecified. Regarding the APOE-$\varepsilon 4$ risk factor, 1,199 individuals do not carry the allele, 803 carry one copy, and 211 carry two copies. table 1 presents a basic demographic and biomarker statistics based on diagnostic. We also supplement the details of the distribution in the supplementary material. Additionally, ADNI provides a dictionary that elucidates the meaning of each variable. The details can also be found in the supplementary material table.

We collected 10387 T1-weighted MRI scans from ADNI, which includes 1510 AD scans (791 males, 719 females; 76.27 ± 7.86 years), 3632 CN scans (1843 males, 1791 females; 77.75 ± 5.73 years), and 5965 MCI scans (3758 males, 2198 females; 75.47 ± 7.53 years). Following established works before(Zhang et al., 2021; Lyu et al., 2022), we performed a series of pre-processing steps as including bias field correction and skull stripping. All T1-weighted images are in NIFTI format and have been registered to the standard Montreal Neurological Institute (MNI) template using FSL-flirt linear registration tools and have a uniform view of 197*233*189.

These T1-weighted MRI scans provide critical imaging perspectives for the analysis of brain diseases. For example, Post-mortem investigations have shown that structures in the medial temporal lobe, particularly the entorhinal cortex and hippocampus, are the first to change in AD and can be observed through MRI imaging techniques(Aramadaka et al., 2023). Fig 3 shows how brain morphology evolves with disease progression, from which we can observe a significant decrease in hippocampus volume.

3 ADTabGPT

3.1 Problem Formulation

In this paper, we formulate each training instance (T, R) represented as a pair of linearized table T and their associated reference text R. The table T consists of a set of tabular data elements related to patients' clinical notes, denoted as $T = t_1, \ldots t_n$, where each element $t_i = a_i, v_i$ corresponds to an attribute(e.g., PTID)-value(e.g., 011_S_0002) pair and can be viewed as a sequence of words. Correspondingly, $R = r_1, \ldots r_m$ represents the set of reference texts, with each r_i providing a narrative description that correlates with table T. The table-to-text model is expected to generate both fluent and faithful natural language descriptions of patients' clinical information based on the given tables.

3.2 Table Transformation

In contrast to the rich corpus data employed in Pre-trained Language Models (PLMs), tabular data present complex topology structures with sparse narrative descriptions, making them quite distinct from natural language representations. Fig.1 illustrates an example of tabular clinical notes and the human language set. To align with the sequential nature of language model, this paper employs a key-value pair template "PTID: 011_S_0002" as "PTID 011_S_0002" to structure the table into a sequence S. Subsequently, all of the key-value pairs are concatenated into a single sentence, that is, "PTID 011_S_0002 Age 74.3 PTGENDER Male ...", while the corresponding textual description T is generated by another template: "Subject 011_S_0002 is a 74.3-year-old Male". It should be noted that the original table may exhibit missing patient information, such as absent scan results or cognitive assessments, which are designated as "unknown". Following the approach used in TableGPT(Gong et al., 2020), the structured table

S and the natural language description T are linked with a special token <*table2text*> to encode the overall table information and signal the model to commence text generation. Another special token <*endoftext*> marks the end of the entire sequence.

This methodology enables our model to encode the structured patients' clinical notes and to learn to generate their natural language descriptions. The training objective of the language model is to maximize the likelihood of generating accurate textual clinical notes, that is, $\max P(T \mid S)$.

3.3 Annotation Challenges & Quality Evaluation

We find that one of the major challenges to generating patients' clinical diagnostic reports from the given table is the presence of numerous missing values, primarily in biomarker measurements (e.g., ABETA, TAU, PTAU). To ensure the completeness and fluency of the generated reports, we set these missing value as "unknown". For instance:

"The subject's genetic profile includes an ApoE4 status of unknown. Neuroimaging with FDG-PET shows average uptake in the angular, temporal, and posterior cingulate regions being unknown. Cerebrospinal fluid analysis reveals AAβ 42 levels at unknown, total tau protein levels at unknown, and phosphorylated tau levels at unknown."

Furthermore, to evaluate whether the clinical reports indeed reflect the patient's potential for dementia, we split the annotated reports corpus (excluding the diagnosis results) into a training set (80%) for training the classifier and a test set (20%) for evaluation. We simply add a linear layer plus a binary softmax layer on the BERT model as the AD classifier. The accuracy of the classifier tested on the binary classification (CN, AD) reached 99%, and for the ternary classification (CN, MCI, AD) achieved 90%, shown as table 2. These results not only confirm the contribution of EHRs to AD diagnosis but also highlight the significant potential of ADNI tabular data for brain disease analysis.

Model	precision	recall	f1-score
$\text{BERT}_{CN/AD}$	0.99	0.99	0.99
$\text{BERT}_{CN/MCI/AD}$	0.90	0.89	0.90

Table 2: Performance of the BERT-based classifier.

4 Experiment

4.1 Baselines

We conduct experiments by fine-tuning following state-of-art text generation methods:

Pointer Generator(See et al., 2017) A LSTM-based seq2seq model incorporates a copy mechanism and was originally designed for text summarization. It has also been adapted for data-to-text applications(Gehrmann et al., 2018).

BERT-to-BERT(Rothe et al., 2020) A transformer-based encoder-decoder model initialized with BERT(Devlin et al., 2018).

BART(Lewis et al., 2019) A pre-trained denoising autoencoder with standard Transformer-based architecure model.

T5(Raffel et al., 2020) A transformer-based model is pre-trained to convert all textual language problems into text-to-text format and proves its effectiveness.

4.2 Evaluation Metrics

Automatic Metrics We employ two automatic metrics, BLEU(Papineni et al., 2002) and PARENT(Dhingra et al., 2019). BLEU is widely used to evaluate machine-generated text quality and PARENT is specifically designed for data-to-text tasks, which evaluates alignment by comparing n-grams from both the reference and generated texts to the source table.

Human Evaluation Since clinical reports require good readability and need to accurately reflect the patient's various indicators, automatic metrics are not adequate to assess the quality of generated reports. We also perform human evaluation based on the following three criteria: 1) **fluency** indicates how well the report is structured; it evaluates whether the expressions are grammatical and fluent. 2) **Consistency** indicates whether the report is consistent with the data provided; it checks for discrepancies between the reported data and the actual measurements. 3) **Relevance** evaluates whether the report includes all necessary information pertinent to a clinical diagnosis, focusing on the inclusion of critical data and avoiding extraneous details. Above criteria ensure that each report is not only accurate but also practical for medical professionals diagnosing and treating patients in real-world clinical settings. Three experts were instructed to rate the three criteria on a scale from 1 to 5, where 1 denotes the worst and 5 represents the best possible score.

Model	BLUE-1/2/3	PARENT	Fluency	Consistency	Relevance
Pointer Generator	31.5/29.1/15.8	28.0	1.3	1.7	1.7
BERT-to-BERT	35.8/30.1/25.5	35.7	2.3	1.7	1.7
T5-small	56.6/40.6/30.6	60.3	3.3	2.0	2.3
T5-base	65.5/50.4/44.3	72.5	4.0	2.0	2.7
BART	76.0/65.0/55.8	88.1	4.7	3.0	3.7
ADTabGPT	**98.8/98.9/98.5**	**97.8**	**5.0**	**4.7**	**5.0**

Table 3: Results of automatic metrics and human evaluation.

4.3 Experiment Settings

We randomly split 16421 health records into train (70%), dev (15%) and test (15%) with no overlap. We download pre-trained GPT2 models from huggingface to avoid starting training from scratch. For optimizer, we adopt the OpenAI AdamW optimizer with 100 warm steps. The batchsize is set to 4 and we fine-tune the model 10 epochs with learning rate is set to 2e-4. All experiments are conducted on a Nivida RTX TITAN GPU.

4.4 Results and Analysis

4.4.1 Automatic Evaluation & Human Evaluation

As shown in Table 3, **first**, from an overall perspective, methods based on Large Language Models (i.e. T5 and BART) outperform the other approaches, with our ADTabGPT achieving the best results. Both metrics for methods based on LLM are scored highly. **Second**, we observed that our ADTabGPT-generated clinical notes almost identical to the references, achieving state-of-the-art results in both automatic and human evaluations. **Third**, we selected 200 generated clinical reports for human evaluation, and the results are presented in Table 3. Our method surpasses five baseline models across all three human-centric metrics. The most significant improvements are observed in consistency and relevance, demonstrating that our method can generate not only fluent diagnostic reports but also accurately reflect various patient indicators.

These results indicate that LLM are adept at handling table-to-text tasks, even in the highly specialized domain of neuroimaging. This reinforces our belief in the potential of using such techniques to mine valuable variables from ADNI tabular EHRs for Alzheimer's disease analysis. Further, although the automatic evaluation shows high scores for the generated clinical reports, there are still existing issues, which will be analyzed in detail in the next case study section.

4.4.2 Case Study

We show in table 4 some clinical note examples generated by T5,BART, and our ADTabGPT. As we can see, both methods perform well in the initial sections of the diagnostic reports, Basic Personal Information. However, with longer documents, T5 and BART tend to produce various errors. The portions marked in orange indicate hallucinations, meaning the outputs cannot be verified against the source table data. The parts marked in red signify more severe, multiple errors. Beyond hallucinations, these include value errors, such as altering the patient's measurements and test results, grammatical error, and missing critical details. This is particularly severe in the Cognitive and Neurofunctional Test Results section of T5's outputs, where a significant amount of information is missing, affecting the understanding of patients' diagnostic results. The sections marked in green highlight anomalous symbols in the diagnostic results, although the values are correct. This issue only appears in the BART-generated outputs. The occurrence of these problems may be due to the limited capacity of large models to handle long documents, resulting in errors when the patient's diagnostic report is lengthy. Our ADTabGPT demonstrates perfect ability to generate fluent and faithful clinical reports in this example.

4.4.3 EHR case

We provide an example of a patient's diagnosis to detail our Multimodal and Longitudinal Neuroimaging EHR. As illustrated in Figure 4, we can observe the progression of this subject's condition from both diagnostic reports and MRI imaging perspectives. In this case, the patient's records span from 2006-01-19 to 2023-04-04, with the condition progressing from MCI to a confirmed diagnosis of AD.

Attribute	Value
PTID	002_S_4237
Age	80.9
Sex	Female
...	...

ADTabGPT:(consistent with the reference)

Basic Personal Information: Subject 002_S_4237 is a 80.9-year-old female who has completed 13 years of education. The ethnicity is Not Hisp/Latino and race is white. Marital status is never married. Initially diagnosed is early mild cognitive impairment, and on the date of 2007-10-14, the final diagnosis is mild cognitive impairment.

Biomarker Measurements: The subject's genetic profile includes an ApoE4 status of 0.0. Neuroimaging with FDG-PET shows average uptake in the angular, temporal, and posterior cingulate regions being 1.17185. Cerebrospinal fluid analysis reveals Aβ 42 levels at more than 1700, total tau protein levels at 302.1, and phosphorylated tau levels at 25.66.

Cognitive and Neurofunctional Test Results: The Mini-Mental State Examination score stands at 29.0. The Clinical Dementia Rating, sum of boxes, is 5.0. ADAS 11 and 13 scores are 12.0 and 12.0 respectively, with a score of 5.0 in delayed word recall. The Rey Auditory Verbal Learning Test results are as follows: immediate recall at 25.0, learning score at 9.0, forgetting score at 9.0, and percent forgetting at 9.0. Logical Memory test for delayed recall scored 6.0. The Digit Symbol Substitution test resulted in a score of Unknown. Trails B testing showed a score of 253.0. The modified Preclinical Alzheimer's Cognitive Composite scores using Digit Symbol Substitution and Trails B are -4.64725 and -6.38287 respectively. The Functional Activities Questionnaire rated the subject as 0.0.

Imaging Data Points: Under MRI conditions at a field strength of 3 Tesla, using Cross-Sectional FreeSurfer (5.1), the imaging data recorded includes ventricles volume at 92447.0, hippocampus volume at 6102.0, whole brain volume at 976304.0, entorhinal cortex volume at 3915.0, fusiform gyrus volume at 16367.0, and middle temporal area volume at 18784.0. The intracranial volume measured is 1516250.0.

T5-base:

Basic Personal Information: Subject 002_S_4237 is a 80.9-year-old Male who has completed 13 years of education. The ethnicity is Not Hisp/Latino and race is white. Marital status is never married. Initially diagnosed is early mild cognitive impairment, and on the date of 2007-10-14, the final diagnosis is mild cognitive impariment.

Biomarker Measurement: The subject's genetic profile icludes an ApoE4 status of 0.0. Neuroimaging with FDG-PET shows average uptake in the angular, temporal, and posterior cingulate regions being dang gyrus. Cerebrospinal fluid analysis reveals Aβ 42 levels at spital to help wiyh the process of choosing he prefers the most.

Cognitive and Neurofunctional Test Results: The The Mini-Mental State Examination score stands at 4 and 13 in delay word recall. The Clinical Dementia Rating, sum of box, is 4. ADAS 11 and 13 scores are 9.0 and 9.0-PATH at 3.0.

Using Cross-Sectional FreeSurfer (5.1), **the Imaging Data Point:** Under MRI conditions are –4.64725 and 2-6.38287 respectively, with hippocampus volume at 6102.0. Logical Memory test at 6.0. The Digit Symbol Substitution test result in uncontrolled recall. Trails B testing showed obstructive Learning Test results are as follows: immediate recall at 225.0, learning score at 59.0 and percent forgetting at 7.0 in entorhinal cortex volume starting at 976304.0, fusiform gyrus density at 16367.0; the intracranial volume measured is 1516250.0

BART:

Basic Personal Information: Subject 002_S_4237 is a 80.9-year-old female who has completed 13 years of education. The ethnicity is Not Hisp/Latino and race is white. Marital status is never married. Initially diagnosed is early mild cognitive impairment, and on the date of 2007-10-14, the final diagnosis is mild cognitive impairment.

BART:

Biomarker Measurements: The subject's genetic profile includes an ApoE4 status of 0.0. Neuroimaging with FDG-PET shows average uptake in the angular, temporal, and posterior cingulate regions being 1.17185. Cerebrospinal fluid analysis reveals Aβ 42 levels at more than 1700, total tau protein levels at 302.1, and phosphorylated tau levels at 25.66.

Cognitive and Neurofunctional Test Results: The Mini-Mental State Examination score stands at 29.0, with a score of 5.0 in elay word recall. The Rey Auditory Verbal Learning Test results are as follows: immediate recall at 25.0, learning score at 9.0%, forgetting score at 9.0, and percent forgetting at 9.0%. Logical Memory test for delayed recall scored 6.0%. The Digit Symbol Substitution test resulted in a score of Unknown. Trails B testing showed a score of 253.0 and a score for modified Preclinical Alzheimer's Cognitive Composite scores of 253. The Functional Activities Questionnaire rated the subject as -6.38287.

Imaging Data Points: Under MRI conditions at a field strength of 3 Tesla, using Cross-Sectional FreeSurfer (5.1), the imaging data recorded includes ventricles volume at 92447.1, hippocampus volume at 6102.0, whole brain volume at 976304.0), entorhinal cortex volume at 3915.0,\" fusiform gyrus volume at 16367.0\", and middle temporal area volume at 18784.0\", The intracranial volume measured is 1516250.0\ ".

Table 4: A sample tabular clinical diagnostic notes and the textual generated clinical diagnostic reports by our ADTabGPT and baslines. Text highlighted in orange indicates hallucinations, green text signifies errors involving anomalous symbols, and red text denotes multiple errors, including value errors,grammatical mistakes, and missing critical details.

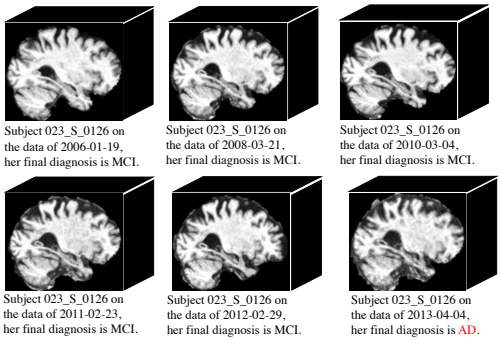


Figure 4: An example of patient's multimodal and longitudinal EHR

5 Conclusion

Given the critical need for multimodal and longitudinal EHRs in brain-focused studies, we introduce the first vision-language brain dataset including 3D brain volume and comprehensive patient clinical notes. By collecting and analyzing ADNI tabular data, we developed a GPT-based method to transform them into information-rich and fluent natural language clinical reports. These reports encompass patients' demographic information, biomarker measurements, cognitive assessments, and volumetric data. Using BERT for disease classification achieved an accuracy of 98.89%, which not only confirms the contribution of tabular EHRs to AD diagnosis but also highlights the significant potential of ADNI tabular data for brain disease analysis. Additionally, we collected 10,387 volumetric T1-weighted MRI scans from ADNI and adopted a series of preprocessing steps, which provide critical imaging perspectives for the analysis of brain diseases. Through the aforementioned annotation and preprocessing methods, we ultimately established a multimodal and longitudinal EHRs dataset that includes 3D brain image volumes and corresponding clinical notes. While ADNI dataset is publicly accessible, it requires approved access. To support future research in this area, we make our processing methods and code publicly available at `https://anonymous.4open.science/r/ADNI-table-to-Text-2EDB`

Limitations

Data scarcity The ADNI dataset is extensive, and we have selected ADNIMERGE for our preliminary work. This table merges several key variables from various case report forms and biomarker laboratory summaries across the ADNI protocols. However, there are additional valuable tables, such as those detailing family history, drug history, and potential causes of AD (e.g., Frontotemporal Dementia, Major Depression, Parkinson's Disease, Huntington's Disease, etc.) Understanding these notes requires high levels of specialized knowledge. In our future work, we aim to integrate more of this

information to form a more comprehensive diagnostic report.

Multi-scale features A key distinction between medical and natural images lies in the significance of multi-scale features. Captions for natural images are typically concise, providing an overall description. In contrast, medical reports consist of multiple aspects and contain significantly more detailed information, as illustrated in 1. These non-image clinical informations are also pay a crucial role in the medical domain, especially for AD diagnosis. For instance, advanced age and specific genes, like APOE4, are both significant risk factors for AD but cannot be observed from images alone. Unlike exsiting radiology reports, the clinical reports in our dataset not only describe the volumetric aspects of T1 scans but also include patients' demographic information, biomarker measurements, and cognitive assessments. How to effectively integrate these clinical details into a unified embedding space and align them effectively is a direction we will explore in our future research.

Inconsistency between pre-training and application A large number of research demonstrates that CLIP's image encoder, initially pre-trained on nature images, also achieves impressive performance in the medical domain. However, most existing medical VLP models are designed to work with 2D images, which may significantly limit their effectiveness in the neuroimaging domain, given the complex and sensitive nature of brain tissue. An important question arises: How can we effectively extend a 2D image encoder to extract features from 3D medical? To address this, we will focus on developing 3D VLP-driven models, specifically for brain analysis.

Medical VLP fairness As previously reported in Section 2, there is significant imbalance in the distribution of fine-grained attributes such as race, ethnicity, and marital status within the data. Although the ADNI dataset is widely used, it was not primarily designed for fairness. Historical instances of bias in various technologies highlight the critical need for fairness in machine learning(Wang et al., 2022a; Dehdashtian et al., 2023). Large Vision-Language Pre-training models, which influence diagnostic and treatment decisions, can perpetuate healthcare disparities and result in adverse outcomes if they exhibit bias. Therefore, improving dataset quality and ensuring the fairness of model algorithms are not just ethical and legal imperatives, but also crucial for achieving healthcare equity(Luo et al., 2024). While VLP models have significantly advanced various medical tasks and propelled the development of AI in healthcare, one of our key goals, as well as a direction for future efforts is to ensure these models provide fair and unbiased diagnostic results across different races, genders, and socioeconomic statuses.

Beyond image-text alignment The philosophy of CLIP is centered on aligning different modalities, specifically images and text, by enabling the model to understand and establish meaningful connections between visual and textual content. In the neuroimaging domain, images often involve various modalities such as MRI and PET scans. Specifically, structural MRI (sMRI) is used to depict the structure of the brain, while functional MRI (fMRI) reveals metabolic activity in the brain during specific tasks(e.g., sensory, motor, cognitive functions, ect.). Each modality provides unique insights into different aspects of a patient's condition. Furthermore, one of the most notable features of AD is hippocampal atrophy, however, in early mild cognitive impairment, structural changes are often not apparent and are typically inferred through functional assessments. In our future work, we plan to incorporate more modalities, such as functional MRI, combined with EHRs for comprehensive longitudinal analysis of brain diseases.

References

2022. *WWW '22: Proceedings of the ACM Web Conference 2022*. Association for Computing Machinery, New York, NY, USA.

Deepa Anand, Vanika Singhal, Dattesh D Shanbhag, Shriram KS, Uday Patil, Chitresh Bhushan, Kavitha Manickam, Dawei Gui, Rakesh Mullick, Avinash Gopal, et al. 2023. One-shot localization and segmentation of medical images with foundation models. *arXiv preprint arXiv:2310.18642*.

Ewa Andrejczuk, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yasemin Altun. 2022. Table-to-text generation and pre-training with tabt5. *arXiv preprint arXiv:2210.09162*.

Saikumar Aramadaka, Raam Mannam, Rajagopal Sankara Narayanan, Arpit Bansal, Vishnu R Yanamaladoddi, Sai Suseel Sarvepalli, and Shree Laya Vemula. 2023. Neuroimaging in alzheimer's disease for early diagnosis: a comprehensive review. *Cureus*, 15(5).

Sanoojan Baliah, Fadillah A Maani, Santosh Sanjeev, and Muhammad Haris Khan. 2023. Exploring the transfer learning capabilities of clip in domain generalization for diabetic retinopathy. In *International*

Workshop on Machine Learning in Medical Imaging, pages 444–453. Springer.

Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2022. Biomedlm: a domain-specific large language model for biomedical text. *Stanford CRFM Blog*.

Anna L Buczak, Steven Babin, and Linda Moniz. 2010. Data-driven approach for creating synthetic electronic medical records. *BMC medical informatics and decision making*, 10:1–28.

Hongmin Cai, Zhixuan Zhou, Defu Yang, Guorong Wu, and Jiazhou Chen. 2023. Discovering brain network dysfunction in alzheimer's disease using brain hypergraph neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–240. Springer.

Maria C Carrillo and Eliezer Masliah. 2023. Nia-aa revised clinical criteria for alzheimer's disease. In *Alzheimer's Association International Conference*. ALZ.

Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. 2023. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. *arXiv preprint arXiv:2305.10799*.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR.

Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. 2023. Fairvlm: Mitigating bias in pre-trained vision-language models. In *The Twelfth International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

Björn Engelmann, Timo Breuer, and Philipp Schaer. 2023. Simulating users in interactive web table retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 3875–3879, New York, NY, USA. Association for Computing Machinery.

Yingjie Feng, Wei Chen, Xianfeng Gu, Xiaoyin Xu, and Min Zhang. 2023. Multi-modal semi-supervised evidential recycle framework for alzheimer's disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–140. Springer.

Giovanni B Frisoni, Nick C Fox, Clifford R Jack Jr, Philip Scheltens, and Paul M Thompson. 2010. The clinical use of structural mri in alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77.

Sebastian Gehrmann, Falcon Z Dai, Henry Elder, and Alexander M Rush. 2018. End-to-end content and plan selection for data-to-text generation. *arXiv preprint arXiv:1810.04700*.

Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. 2020. Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1978–1988.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2018. Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. 2010. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. 2008. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Mindy J Katz, Richard B Lipton, Charles B Hall, Molly E Zimmerman, Amy E Sanders, Joe Verghese, Dennis W Dickson, and Carol A Derby. 2012. Age-specific and sex-specific prevalence and incidence of mild cognitive impairment, dementia, and alzheimer dementia in blacks and whites: a report from the einstein aging study. *Alzheimer Disease & Associated Disorders*, 26(4):335–343.

Yilin Leng, Wenju Cui, Chen Bai, Zirui Chen, Yanyan Zheng, and Jian Zheng. 2023. Dynamic structural brain network construction by hierarchical prototype embedding gcn using t1-mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 120–130. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, et al. 2021. Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*.

Wei Liang, Kai Zhang, Peng Cao, Pengfei Zhao, Xiaoli Liu, Jinzhu Yang, and Osmar R Zaiane. 2023. Modeling alzheimers' disease progression from multi-task and self-supervised learning perspective with brain networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 310–319. Springer.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. 2024. Fairclip: Harnessing fairness in vision-language learning. *arXiv preprint arXiv:2403.19949*.

Yanjun Lyu, Xiaowei Yu, Dajiang Zhu, and Lu Zhang. 2022. Classification of alzheimer's disease via vision transformer: Classification of alzheimer's disease via vision transformer. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 463–468.

Scott McLachlan, Kudakwashe Dube, and Thomas Gallagher. 2016. Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 439–448. IEEE.

Michael B Miller, August Yue Huang, Junho Kim, Zinan Zhou, Samantha L Kirkham, Eduardo A Maury, Jennifer S Ziegenfuss, Hannah C Reed, Jennifer E Neil, Lariza Rento, et al. 2022. Somatic genomic changes in single alzheimer's disease neurons. *Nature*, 604(7907):714–722.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chantal Pellegrini, Matthias Keicher, Ege Özsoy, Petra Jiraskova, Rickmer Braren, and Nassir Navab. 2023. Xplainer: From x-ray observations to explainable zero-shot diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. Springer.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. 2010. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Constantin Seibold, Simon Reiß, M Saquib Sarfraz, Rainer Stiefelhagen, and Jens Kleesiek. 2022. Breaking with fixed set pathology recognition through

report-guided contrastive training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 690–700. Springer.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2023. Evaluating and enhancing structural understanding capabilities of large language models on tables via input designs. *arXiv preprint arXiv:2305.13062*.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Junyang Wang, Yi Zhang, and Jitao Sang. 2022a. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. *arXiv preprint arXiv:2210.14562*.

Zifeng Wang, Chufan Gao, Cao Xiao, and Jimeng Sun. 2023. Meditab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022b. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Yongjian Wu, Yang Zhou, Jiya Saiyin, Bingzheng Wei, Maode Lai, Jianzhong Shou, Yubo Fan, and Yan Xu. 2023. Zero-shot nuclei detection via visual-language pre-trained models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 693–703. Springer.

Xiaowei Yu, Lu Zhang, Yanjun Lyu, Tianming Liu, and Dajiang Zhu. 2023. Supervised deep tree in alzheimer's disease. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Lu Zhang, Saiyang Na, Tianming Liu, Dajiang Zhu, and Junzhou Huang. 2023. Multimodal deep fusion in hyperbolic space for mild cognitive impairment study. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 674–684. Springer.

Lu Zhang, Li Wang, Jean Gao, Shannon L Risacher, Jingwen Yan, Gang Li, Tianming Liu, Dajiang Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. 2021. Deep fusion of brain structure-function in mild cognitive impairment. *Medical image analysis*, 72:102082.

Ziqi Zhang, Chao Yan, Diego A Mesa, Jimeng Sun, and Bradley A Malin. 2020. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association*, 27(1):99–108.

Rong Zhou, Houliang Zhou, Brian Y Chen, Li Shen, Yu Zhang, and Lifang He. 2023a. Attentive deep canonical correlation analysis for diagnosing alzheimer's disease using multimodal imaging genetics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 681–691. Springer.

Yanjie Zhou, Youhao Li, Feng Zhou, Yong Liu, and Liyun Tu. 2023b. Learning with domain-knowledge for generalizable prediction of alzheimer's disease from multi-site structural mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 452–461. Springer.

appendix

A Related Work

EHRs Generation EHRs are a longitudinal record that encompass a patient's health information. Early works on generating EHRs were predominantly ruled-based(Buczak et al., 2010; McLachlan et al., 2016). However, powered by neural networks, recent advancements have seen the adoption of deep learning models and artificial intelligence (AI), which have significantly enhanced this process. One pioneering innovation was MedGAN proposed by(Choi et al., 2017), aiming to generating multi-label discrete patient records using generative adversarial networks (GANs) models (Goodfellow et al., 2020). Following this development, numerous studies have focused on improving medical text generation(Guan et al., 2018; Baowaly et al., 2019; Zhang et al., 2020). However, most EHRs only focus on static, single-modal EHRs and overlook the representations of imaging information. To the best of our knowledge, we are the first to focus on building a multimodal and longitudinal EHRs dataset for brain disease analysis.

Vision-Language Models The integration of vision and language in deep learning, exemplified by models such as CLIP, trained on a large scale of paired image-text multimodal data, represents

a significant advancement in aligning visual information with textual descriptions in AI. Bootstrapping Language-Image Pre-training (BLIP)(Li et al., 2022) introduces a novel multimodal, unified encoder-decoder framework to learn from noisy image-text pairs. Further enhancing this work, BLIP-2 (Li et al., 2023) proposed a lightweight querying transformer, achieving state-of-the-art performance on various vision-language tasks with considerably fewer trainable parameters.

These success of CLIP-like models have achieved great success in many downstream computer vision applications. Hence, it is intuitive to understanding complex medical imaging using VLP models, which has also led to rapid advancements in various medical domain (e.g., Chest X-ray, multi-organ CT, Brest Histology). For instance, Xplainer (Pellegrini et al., 2023) leverages the CLIP to align the of X-Ray scans and clinical radiology reports representations close in latent space for zero-shot diagnosing pathologies. MedCLIP(Wang et al., 2022b) employs inter-report semantical correlation as the soft optimization target for the alignment between X-Ray medical image and text. CoOPLVT (Baliah et al., 2023) investigate CLIP's transfer learning capabilities and its potential for cross-domain generalization in diabetic retinopathy (DR) classification. Wu et al.(Wu et al., 2023) proposed a zero-shot nuclei detection framework based on VLP models by directly using automatic text prompts.

Above research demonstrate that CLIP's image encoder, initially pre-trained on nature images, also achieves impressive performance in the medical domain. However, most existing medical VLP models are designed based on 2D images, for 3D MRI images, (Anand et al., 2023) taking them into 2D slices. The brain, with its complex features, especially for brain disease diagnosis, is sensitive to even minor tissue changes, and slicing may compromise these crucial features. MedBLIP (Chen et al., 2023) extracts and fuses 3D medical volume images, aligning them with text features in a common space using BioMedLM (Bolton et al., 2022), and then fine-tunes the alignment using LoRA(Hu et al., 2021). However, they only utilizes patient age and a limited number of cognitive test results as textual descriptions, missing out on a wealth of valuable information contained in ADNI EHRs.

One major challenge in applying VLP for brain analysis lies in the scarcity of available image-text pair datasets. Therefore, this paper aims to develop a multimodal and longitudinal EHRs that includes 3D brain image volumes and corresponding clinical notes for brain disease studies. We expect that this dataset not only enhances the quality but also broadens the horizon for research in the brain disease domain.

B Data Descriptions

ADNI provides a dictionary that clarifies the meaning of each variable. To supplement the details, Table 5 presents information about the variables along with their descriptions. We also include an example in this table to illustrate their application. Table 5

C Data Distribution

To supplement the details for our neuroimaging EHR in the main paper, we present additional analyses. Figures 5a and 5b illustrate the distribution of gender and marital status, respectively. Figure 5c depicts the distribution of initial and final diagnoses, while Figures 5d and 5e show the distributions of racial and ethnic backgrounds. Figure 5f presents the distribution of the APOE4 gene allele carriers.

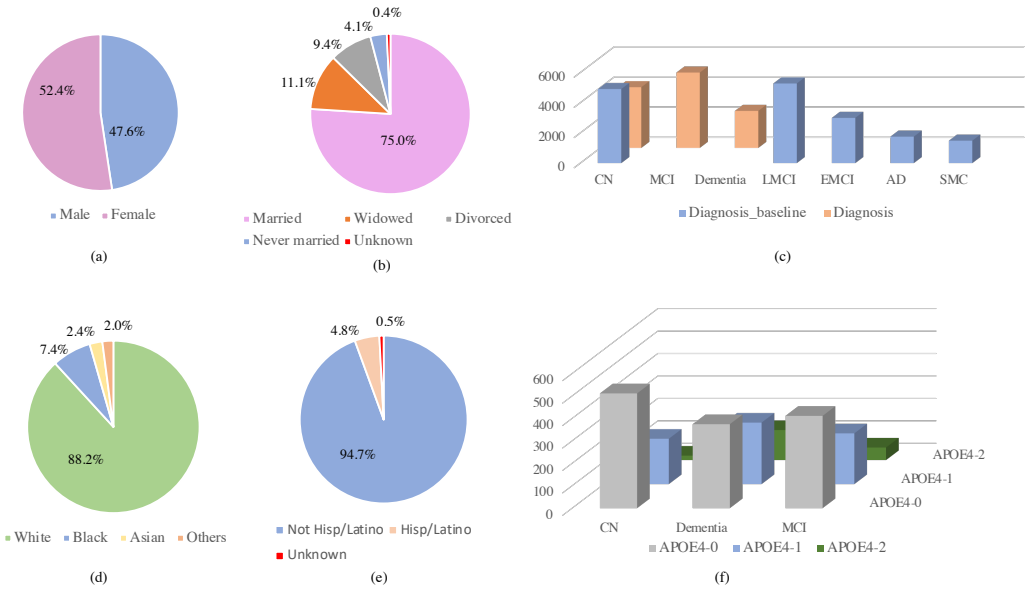


Figure 5: (a) distribution of gender, (b) distribution of marital status, (c) distribution of initial and final diagnoses, (d) distributions of racial, (e) distributions of ethnic, (f) distribution of the APOE4 gene allele carriers

Information	Information Description	Example
PTID	Original study protocol	011_S_0002
Age	Age	74.3
PTGENDER	Sex	Male
PTEDUCAT	Education	16
PTETHCAT	Ethnicity	Not Hisp/Latino
PTRACCAT	Race	White
PTMARRY	Marital	Married
EXAMDATE	Date	09/08/2005
DX_bl	Baseline DX	CN
DX	DX	CN
APOE4	ApoE4	0
FDG	Average FDG-PET of angular, temporal, and posterior cingulate.	1.33615
ABETA	CSF ABETA	741.5
TAU	CSF TAU	239.7
PTAU	CSF PTAU	22.83
CDRSB	CDR	0
ADAS11	ADAS 11	10.67
ADAS13	ADAS 13	18.67
ADASQ4	ADAS Delayed Word Recall	5
MMSE	MMSE	28
RAVLT_immediate	RAVLT Immediate (sum of 5 trials)	44
RAVLT_learning	RAVLT Learning (trial 5 - trial 1)	4
RAVLT_forgetting	RAVLT Forgetting (trial 5 - delayed)	6
RAVLT_perc_forgetting	RAVLT Percent Forgetting	54.5455
LDELTOTAL	Logical Memory - Delayed Recall	10
DIGITSCOR	Digit Symbol Substitution	34
TRABSCOR	Trails B	112
mPACCdigit	ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Digit Symbol Substitution	-4.31028
mPACCtrailsB	ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Trails B	-4.11443
FAQ	FAQ	0
FLDSTRENG	MRI Field Strength	1.5 Tesla MRI
FSVERSION	FSVERSION	Cross-Sectional FreeSurfer (FreeSurfer Version 4.3)
Ventricles	UCSF Ventricles	118233
Hippocampus	UCSF Hippocampus	8336
WholeBrain	UCSF WholeBrain	1229740
Entorhina	UCSF Entorhinal	4117
Fusiform	UCSF Fusiform	16559
MidTemp	UCSF Med Temo	27936
ICV	UCSF ICV	1984660

Table 5: This table presents information from the ADNIMERGE table and its corresponding descriptions from official ADNIMERGE-DICTIONARY, detailing patient data in the ADNI study. Each category of information is highlighted with a specific color: 7 pieces of basic personal and 3 pieces of diagnosis information, 5 biomarker measurements, 15 cognitive and neurofunctional test results, 7 imaging data points, and 2 additional related indicators(MRI Field Strength and FS VERSION). Each entry in the table includes an example for clearer understanding.