Dataset

Step one: Combine subject and assessment and plan for original **long dialogue**

- **Dataset:** Data/ACIBENCH_train_processed, ACIBENCH_validation_processed and dialogue-processed

Step Two: Data cleaning and data split into train/val/test

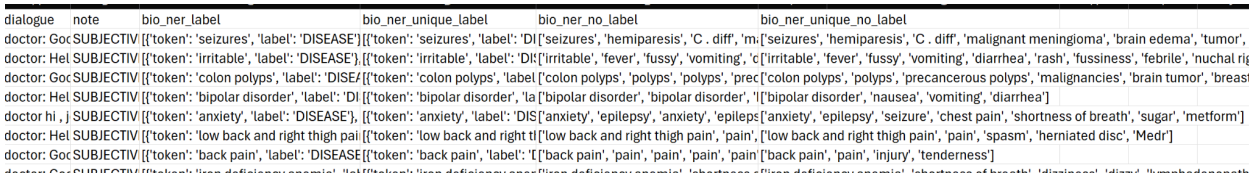
- **Dataset:** Data/Long_dialogue/train_long.csv, test_long.csv, val_long.csv
- **Dataset:** Data/Short_dialogue/train_short.csv, test_short.csv, val_short.csv
- **Notebook:** Notebook/Data_preprocessing and analysis/Data_Split
- **Note:** These are the dataset to run the **baseline model**

Step Three: BioBert/Key Bert/Clinical NER extraction

- **Dataset:** all the data in folder **long_dialogue_NER_extraction** and **short_dialogue_NER_extraction**
- **Note:** Each NER model has it's own test/train/val dataset,
- **Note:** output for NER model are tokens, further merge into word and phrase per IBO tagging

Step Four: BioBert/Key Bert/Clinical NER cleaning

- **Dataset:** all the data in folder **long_dialogue_NER_cleaning** and **short_dialogue_NER_cleaning**
- **Notebook:** Notebook/NER/ALL_NER_Cleaning
- NER_cleaning
    - some NER appears several times and decided to just keep **unique** entities and keep **all entities**
    - keep the NER with it's **label** and **without it's label**
    - Below is a example for the for BioBert



dialogue	note	bio_ner_label	bio_ner_unique_label	bio_ner_no_label	bio_ner_unique_no_label
doctor: Goc	SUBJECTIV	[{'token': 'seizures', 'label': 'DISEASE'}	[{'token': 'seizures', 'label': 'DI	['seizures', 'hemiparesis', 'C . diff', 'm	['seizures', 'hemiparesis', 'C . diff', 'malignant meningioma', 'brain edema', 'tumor',
doctor: Hel	SUBJECTIV	[{'token': 'irritable', 'label': 'DISEASE'},	[{'token': 'irritable', 'label': 'DI	['irritable', 'fever', 'fussy', 'vomiting', 'c	['irritable', 'fever', 'fussy', 'vomiting', 'diarrhea', 'rash', 'fussiness', 'febrile', 'nuchal rig
doctor: Goc	SUBJECTIV	[{'token': 'colon polyps', 'label': 'DISEA	[{'token': 'colon polyps', 'label	['colon polyps', 'polyps', 'polyps', 'prec	['colon polyps', 'polyps', 'precancerous polyps', 'malignancies', 'brain tumor', 'breast
doctor: Hel	SUBJECTIV	[{'token': 'bipolar disorder', 'label': 'DI	[{'token': 'bipolar disorder', 'la	['bipolar disorder', 'bipolar disorder', 'I	['bipolar disorder', 'nausea', 'vomiting', 'diarrhea']
doctor hi , j	SUBJECTIV	[{'token': 'anxiety', 'label': 'DISEASE'},	[{'token': 'anxiety', 'label': 'DIS	['anxiety', 'epilepsy', 'anxiety', 'epileps	['anxiety', 'epilepsy', 'seizure', 'chest pain', 'shortness of breath', 'sugar', 'metform']
doctor: Hel	SUBJECTIV	[{'token': 'low back and right thigh pai	[{'token': 'low back and right t	['low back and right thigh pain', 'pain',	['low back and right thigh pain', 'pain', 'spasm', 'herniated disc', 'Medr']
doctor: Goc	SUBJECTIV	[{'token': 'back pain', 'label': 'DISEASE	[{'token': 'back pain', 'label': '[	['back pain', 'pain', 'pain', 'pain', 'pain	['back pain', 'pain', 'injury', 'tenderness']

Bio_ner_label: this is with all entities and their label (list of dictionary)
Bio_ner_unqiue_label: this is the unique and their label (list of dictionary)
Bio_ner_no_label: This is all entities without label (list of string)
Bio_ner_unique_no_label: This is unique entities without label (list of string)

- **Note:** These datasets are used for improved model