EEG Prognostic Data Pipeline & Characterization Report

> **Project Focus:** EEG-based prognostication in severe brain injury patients using language tracking, oddball/P300, and command-following paradigms to assess residual auditory and cognitive processing in the acute recovery window.
>
> **Institution:** Harborview Medical Center
>
> **Data Status:** Active research project (latest data: October 2025)

Executive Summary

This document provides a comprehensive characterization of the EEG prognostic data pipeline for severe brain injury patients. The project leverages multimodal data streams—EEG recordings, stimulus timing logs, audio stimuli, and clinical metadata—to identify neural markers of residual cognitive processing that may improve recovery prognostication.

Key Data Assets:

- **17 EEG recordings** (EDF format) from 10 patients
- **21 CSV stimulus timing logs** with 1,120+ trial records
- **38+ audio stimulus files** (language sentences, commands, oddball tones)
- **Multiple experimental paradigms:** language tracking, motor commands, oddball/P300, emotional voice stimuli

1. Project Context

Severe brain injury patients can retain residual auditory and cognitive processing even without overt behavioral responses. The objective is to leverage EEG markers to improve recovery prognostication in the acute window by assessing:

- **Language tracking:** Neural signatures of speech comprehension (e.g., ITPC - Inter-Trial Phase Coherence)
- **Oddball/P300 responses:** Attention and novelty detection capabilities
- **Command-following:** Motor imagery and volition via EEG spectral features (bandpower)
- **Emotional processing:** Responses to familiar voices (loved ones)

2. Data Streams in Use

2.1 EEG Recordings (Primary Data Stream)

Attribute	Details
Format	European Data Format (.EDF / .EDF+)
Location	OneDrive: `EEG Project Data/EEG/edf/`
Count	17 EDF files
Patients	CON001–CON010, TEST
Versions	Original and clipped (`*_clipped.EDF`)
Channels	16–64 EEG channels (multi-channel neurophysiological recordings)
Sampling Rate	Typically 250–2000 Hz for clinical EEG
Bit Depth	16-bit resolution

Characteristics:

- Raw EEG recordings with both original and clipped versions available
- Clipped versions trimmed to experimental session windows for focused analysis
- Multi-channel recordings capturing comprehensive brain activity

2.2 Trial/Event Logs (CSV Files)

Attribute	Details
Format	CSV (comma-separated values)
Location	OneDrive: `EEG Project Data/EEG/` and `stimuli_record/`
Count	21 CSV files
Largest File	`patient_df_043025.csv` with 1,120 rows
Timing Precision	Millisecond precision (Unix timestamps with 3+ decimal places)

Schema (Core Columns):

- `patient_id` : Patient identifier (e.g., CON008, CON009)
- `date` : Experiment date
- `trial_type` : Type of stimulus presented (see trial types below)
- `sentences` : Array of sentence IDs for language trials
- `start_time` : Unix timestamp of trial start
- `end_time` : Unix timestamp of trial end
- `duration` : Trial duration in seconds

Optional Columns (Schema Variants):

- `trial_index` : Sequential trial number
- `paradigm` : Experimental paradigm category
- `stimulus_details` : Additional stimulus information
- `notes` : Qualitative annotations

Trial Types:

Trial Type	Description	Duration
`language`	Sentence stimuli (12 sentences per trial)	~15–16 seconds

`left_command+p` / `right_command+p`	Motor command paradigms (imagery/execution)	~200–212 seconds
`oddball+p`	Oddball detection task (standard/rare tones)	~30–34 seconds
`control`	Control condition (baseline)	~2.6 seconds
`loved_one_voice`	Emotional stimulus from familiar person	~3.5 seconds

2.3 Audio Stimuli Files (Supporting Data)

Category	Location	Count	Details
Sentence Stimuli	`Audio/sentences/`	35 files	`lang0.wav` through `lang34.wav` (missing `lang28.wav`)
Prompt Audio	`Audio/prompts/`	2 files	`motorcommandprompt.wav` , `oddballprompt.wav`
Static Commands	`Audio/static/`	5 files	`left_keep.mp3` , `left_stop.mp3` , `right_keep.mp3` , `right_stop.mp3` , `sample_beep.mp3`
Control Voices	`Audio/Voice/Trimmed/`	2 files	`ControlStatement_female.wav` , `ControlStatement_male.wav`

Audio Technical Specifications:

- **Format:** WAV (uncompressed), MP3 (compressed)
- **Sample Rate:** 44.1 kHz or 48 kHz (standard for .WAV)
- **Bit Depth:** Typically 16-bit

- **Channels:** Mono or stereo
- **Duration:**
  - Sentences: ~1.3s each (15.5s total / 12 sentences)
  - Commands: Variable (200s sessions)
  - Voices: ~3.5s each

2.4 Session Metadata & Patient Information

File Type	Examples	Content
Patient DataFrames	`patient_df*.csv` (versions: 043025, 052225, etc.)	Patient clinical information, trial-level data, multiple versions with date suffixes
Patient History	`patient_history*.csv`	Visit dates, session history
Patient Notes	`patient_notes*.csv`	Qualitative observations (e.g., "interrupted by monitor briefly", "audio on left side")

Known Metadata Examples:

- **CON001a:** "audio on left side, planning second experiment on same patient, same day" (2024-09-17)
- **CON002:** "interrupted by monitor briefly" (2024-09-24)
- **CON003:** "instructed not to follow commands" (2025-01-14)

3. Data Volume & Characterization

3.1 Overall Data Inventory

Data Type	Count	Storage
EEG Files (.EDF)	17 files	~MB–GB range (binary format)

Stimulus Records (.CSV)	21 files	Lightweight text files
Audio Files (.WAV/.MP3)	38+ files	~MB range per file
Patient Metadata (.CSV)	~6 files	Lightweight text files

3.2 Trial Volume by File

File	Row Count	Notes
`patient_df_043025.csv`	1,120 rows	Largest aggregated trial log
Other `patient_df` variants	5–672 rows	Multiple versions with varying completeness
Stimulus result files	24–184 rows	Session-specific logs (e.g., CON009: 184 trials)
Legacy "old stimulus software" logs	~84 rows each	Historical data from earlier protocol versions

3.3 Experimental Session Examples

CON008 (August 14, 2025):

- Total trials: **84**
- Language trials: 60+ (~15.5s each)
- Motor command trials: 4+ (~210s each)
- Oddball trials: 3+ (~32s each)
- Total session duration: ~2+ hours

CON009 (August 26, 2025):

- Total trials: **184**
- Language trials: 100+ (~15.5s each)
- Motor command trials: 4+ (~212s each)

- Oddball trials: 5+ (~33s each)
- Control trials: 30+ (~2.6s each)
- Loved one voice trials: 40+ (~3.5s each)
- Total session duration: ~3+ hours

3.4 Typical Trial-Type Distribution per Session

Trial Type	Count	Total Duration
Language	~72 trials	~18–19 minutes
Control	~50 trials	~2 minutes
Loved one voice	~50 trials	~3 minutes
Command blocks (left/right)	~3 each (6 total)	~20 minutes
Oddball/beep	~4 trials	~2 minutes

3.5 Patient Cohort

Active Patients Analyzed:

- **CON008** (August 14, 2025): 84 trials
- **CON009** (August 26, 2025): 184 trials
- **CON010** (October 31, 2025): Data available
- **Historical data:** CON001–CON007, TEST

Total Patient Count: n = 10 (CON001–CON010)

4. Data Storage & Access

4.1 Storage Location

Primary Repository

OneDrive (cloud-based storage)

Local Working Directory

```
/Users/arnavdixit/Downloads/OneDrive_2025-12-09/
```

Organizational Structure

The data is organized in a hierarchical directory structure as follows:

- **OneDrive_2025-12-09/**
  - **EEG Project Data/**
    - **Audio/** *(Stimulus audio files)*
      - **sentences/** - Language stimuli (35 files)
      - **prompts/** - Task instructions
      - **static/** - Motor command audio
      - **Voice/**
        - **Trimmed/** - Control and emotional voice stimuli
    - **EEG/**
      - **edf/** - EEG recordings (17 files)
        - **old stimulus software/** - Legacy recordings
      - ***.csv** - Patient metadata
      - **stimuli_record/** - Experimental timing logs (21 files)
        - **old stimulus software/** - Legacy logs
  - **Slides/** - Research presentations

4.2 Software & Access Methods

Data Type	Primary Software	Key Libraries/Tools

EEG Data Processing	Python with MNE-Python	• MNE (EDF loading, epoching, filtering) • NumPy (array operations) • SciPy (signal processing, PSD, FFT)
Stimulus Timing Analysis	Python with pandas	• pandas (dataframe operations) • NumPy (numerical operations)
Audio Stimulus Management	Python audio libraries	• librosa (audio feature extraction) • soundfile (WAV I/O) • pydub (audio manipulation)
Exploratory Data Analysis	Jupyter Notebooks	• pandas (tabular EDA) • matplotlib / seaborn (visualization)

4.3 Typical Analysis Pipeline

1. **Load EDF files** using MNE
2. **Synchronize** with stimulus timing CSV files
3. **Epoch data** based on trial timing
4. **Apply signal processing:** filtering, artifact rejection (ICA)
5. **Extract features:**
    - Event-related potentials (ERPs)
    - Power spectral density (PSD) / bandpower
    - Inter-trial phase coherence (ITPC) for language tracking
6. **Statistical analysis** and modeling

5. Data Pipeline Workflow

Phase 1: Data Generation (Experimental Collection)

```
Patient Setup → EEG Recording Start → Stimulus Presentation → Real-time Logging → EEG Recording Stop → Data Storage (OneDrive)
```

Input:

- Patient consent and setup
- Audio stimuli (pre-loaded)
- Experimental protocol

Output:

- Raw EDF file (continuous EEG data)
- CSV stimulus log (precise timing of each trial)

Phase 2: Data Preprocessing

```
Load EDF → Filter/Clean → Clip to Experimental Window → Save Clipped EDF → Manual QC
```

Input: Raw EDF files

Output: `*_clipped.EDF` files (trimmed to experimental session)

Phase 3: Data Analysis (Current & Planned)

```
Load Clipped EDF → Load Stimulus CSV → Epoch by Trial Type → Signal Processing → Feature Extraction → Statistical Analysis → Results
```

Planned Feature Extraction:

- **PSD / Bandpower:** For command-following paradigms (motor imagery)
- **ITPC (Inter-Trial Phase Coherence):** For language-tracking comprehension markers
- **ERP (Event-Related Potentials):** For oddball/P300 attention responses
- **Time-frequency analysis:** Spectrograms, wavelet transforms

6. Planned Data Products

6.1 Processed Data Outputs

Data Product	Format	Description
Aligned EEG Epochs	HDF5 / NumPy arrays	Trial-locked EEG segments synchronized with event logs (language, command, oddball)
Feature Tables	Parquet / CSV	Tabular features per trial: PSD, bandpower, ITPC, ERP amplitudes, latencies
Behavioral Event Tables	CSV	Trial timing, stimulus labels, QC flags (artifact contamination, completeness)
Visualizations	PNG / HTML (Plotly)	ERP waveforms, spectrograms, topographic maps, statistical summaries
Data Dictionary	Markdown / PDF	Comprehensive variable definitions, schema documentation, versioning

6.2 Organized Output Structure (Proposed)

To maintain reproducibility and organization, all processed outputs will be stored in a structured directory hierarchy. The proposed organization includes the following subdirectories:

processed/

- **epochs/** - Aligned EEG epochs (HDF5, NumPy)
- **features/** - Extracted feature tables (Parquet, CSV)
- **qc/** - Quality control logs and flags
- **plots/** - Visualizations (PNG, HTML)
- **models/** - Trained models (if applicable)
- **README.md** - Data dictionary and schema documentation

Directory Descriptions

- **epochs/** - Contains trial-locked EEG segments synchronized with event logs, stored in HDF5 or NumPy array format for efficient access
- **features/** - Extracted features per trial (PSD, bandpower, ITPC, ERP metrics) in Parquet or CSV format for easy integration with analysis pipelines
- **qc/** - Quality control logs documenting artifact rejection, data completeness, and timing validation flags
- **plots/** - Visualization outputs including ERP waveforms, spectrograms, topographic maps, and statistical summaries
- **models/** - Trained machine learning models and associated metadata (hyperparameters, performance metrics)
- **README.md** - Comprehensive data dictionary with variable definitions, processing parameters, and version information

7. Data Weaknesses, Limitations & Risks

> **Critical Challenges:** The following limitations must be addressed to ensure robust, reproducible analyses.

7.1 Sample Size Limitations

Issue:

- Small patient cohort: **n = 10** total
- Only ~3 patients fully processed to date (CON008, CON009, CON010)

Impact:

- Limited statistical power for group-level analyses
- Risk of overfitting in machine learning models
- Difficult to generalize findings

Mitigation:

- Carefully designed within-subject analyses
- Conservative statistical approaches (e.g., non-parametric tests, cross-validation)
- Focus on single-case or small-N designs

7.2 Missing Data Points

Issue:

- **CON006.EDF** noted as not downloaded in error log
- Sentence file **lang28.wav** missing from sequence (lang0–lang34)
- Inconsistent trial counts across patients (84 vs. 184 trials)

Impact:

- Incomplete dataset
- Potential gaps in stimulus coverage
- May affect counterbalancing of stimuli

Mitigation:

- Document all missing data explicitly
- Analyze only complete cases
- Request missing files from OneDrive sync

7.3 Data Heterogeneity & Protocol Drift

Issue:

- Different experimental protocols across patients:
  - CON009 includes `loved_one_voice` and `control` trials not present in CON008
  - "old stimulus software" folder suggests protocol evolution
- Variable trial counts per session (24–184 trials)

Impact:

- Difficult to pool data across subjects
- Protocol changes over time complicate longitudinal analyses
- Requires cohort-specific analyses

Mitigation:

- Document protocol versions meticulously
- Analyze by cohort (legacy vs. current protocol)

- Focus on paradigms common to all patients (language, oddball)

7.4 Schema Drift Across CSV Variants

Issue:

- Multiple `patient_df` versions with inconsistent schemas:
  - Optional columns: `trial_index`, `paradigm`, `stimulus_details`, `notes`
  - Date suffixes: `043025`, `052225`
  - "(1)" and "(3)" suffixes suggest duplicates or outdated files
- No clear indication of which version is "current"

Impact:

- Risk of analyzing outdated or incorrect metadata
- Requires manual schema harmonization
- Potential duplication/overlap across trial logs

Mitigation:

- Reconcile and deduplicate all `patient_df` versions
- Lock a unified schema for all downstream analyses
- Implement proper version control (Git)
- Date-stamp final analysis datasets

7.5 Temporal Alignment Challenges

Issue:

- EEG files and stimulus CSV files stored separately
- Requires precise timestamp synchronization:
  - Unix timestamps in CSV (millisecond precision)
  - EDF internal timing (may use different clock)

Impact:

- Risk of misalignment between neural data and stimulus events
- Could invalidate event-locked analyses (ERPs, ITPC)

Mitigation:

- Careful validation of timing synchronization
- Use of trigger markers embedded in EEG (if available)
- Cross-check event timing with audio stimulus durations

7.6 Class Imbalance

Issue:

- Far fewer command trials (~6 per session) compared to language trials (~72 per session)
- Some sessions are small (e.g., 24 rows in legacy files)

Impact:

- Unbalanced datasets for classification tasks
- May bias models toward majority class (language)

Mitigation:

- Use stratified sampling for train/test splits
- Apply class weighting in models
- Evaluate with balanced metrics (F1, AUC-ROC, not just accuracy)

7.7 Sparse Clinical Metadata

Issue:

- No demographics (age, sex) in current CSVs
- No outcome measures (e.g., Glasgow Outcome Scale, CRS-R scores)
- Minimal clinical notes (`patient_notes.csv` has only 6 entries)

Impact:

- Cannot assess prognostic value without outcome labels
- Difficult to control for confounds (age, injury severity)

Mitigation:

- Integrate clinical outcome data from medical records
- Maintain separate secure database for PHI-compliant metadata

7.8 Unlabeled Stimulus Content

Issue:

- Sentence content (what `lang0.wav` actually says) not documented in data
- No semantic labeling of sentences (e.g., topic, complexity, word count)

Impact:

- Difficulty interpreting language-specific results
- Cannot assess stimulus-level effects (e.g., sentence difficulty)

Mitigation:

- Create stimulus manifest linking sentence IDs to text/audio content
- Annotate linguistic features (word count, semantic category)

7.9 Artifact Contamination Risk

Issue:

- Clinical EEG recordings prone to:
  - Movement artifacts
  - Electrical interference (ICU equipment)
  - Eye movements/blinks
  - Muscle activity

Impact:

- Reduced signal quality
- Potential false positives in analysis

Mitigation:

- Robust artifact rejection pipelines
- Independent component analysis (ICA) for artifact removal
- Manual QC of epochs with flagging system

7.10 Missing Validation Data Structure

Issue:

- No obvious held-out test set or cross-validation structure
- Risk of data leakage in model training

Impact:

- Risk of overfitting to training data
- Optimistic performance estimates

Mitigation:

- Implement proper train/validation/test splits
- Use leave-one-subject-out cross-validation (LOSO-CV)

7.11 Documentation Gaps

Issue:

- Experimental notes minimal (`patient_notes.csv` : 6 entries)
- No data dictionary for CSV column meanings
- Unclear what triggered creation of `*_clipped` files

Impact:

- Difficulty reproducing analyses
- Unclear data provenance

Mitigation:

- Create comprehensive README and data dictionary
- Log all preprocessing steps with parameters
- Maintain analysis notebooks with detailed annotations

8. Data Quality Indicators

Strengths

- **Precise Timing:** Sub-second (millisecond) precision in stimulus timing logs
- **Standardized Format:** EDF is industry-standard for EEG; CSV for tabular data
- **Multiple Modalities:** Combined neural, behavioral, and stimulus data
- **Longitudinal Structure:** Multiple sessions per patient over time
- **Controlled Stimuli:** Standardized audio files for reproducibility
- **Rich Paradigms:** Language, motor commands, oddball, emotional stimuli

Areas for Improvement

- **Metadata Completeness:** Enhance patient and session documentation (demographics, outcomes)
- **Data Centralization:** Consolidate versioned files into single authoritative sources
- **Stimulus Documentation:** Create manifest linking sentence IDs to actual text/audio content
- **Quality Control Logs:** Implement systematic QC checkpoints with pass/fail criteria
- **Schema Standardization:** Lock unified schema across all trial logs

9. Technical Specifications Summary

EEG (EDF) Specifications

- **Standard:** EDF/EDF+ format (European Data Format)
- **Channels:** 16–64 EEG channels (specific count TBD from file inspection)
- **Sampling Rate:** Typically 250–2000 Hz for clinical EEG
- **Bit Depth:** 16-bit resolution

Audio Stimulus Specifications

- **Sample Rate:** 44.1 kHz or 48 kHz (standard for .WAV)
- **Bit Depth:** Typically 16-bit
- **Channels:** Mono or stereo
- **Duration:**

  - Sentences: ~1.3s each (15.5s total / 12 sentences)
  - Commands: Variable (200s sessions)
  - Voices: ~3.5s each

Experimental Timing Precision

- **Resolution:** Millisecond precision (Unix timestamps with 3+ decimal places)
- **Jitter:** Low jitter in trial timing (<100ms variation within trial types)

10. Recommendations for Pipeline Enhancement

1. **Sync Raw EDF/WAV from OneDrive**
   - Lock a unified schema for all trial logs
   - Ensure complete data synchronization
2. **Reconcile CSV Variants**
   - Deduplicate `patient_df` versions
   - Identify and document "current" authoritative version
   - Harmonize schemas across all files
3. **Align Events to Signals**
   - Validate timing synchronization between EEG and CSV logs
   - Add QC flags for artifacts, completeness, timing drift
4. **Generate Processed Feature Tables**
   - PSD/bandpower for command paradigms
   - ITPC for language-tracking markers
   - ERP amplitudes/latencies for oddball paradigms
5. **Implement Data Dictionary & Versioning**
   - Document all variables, schemas, and preprocessing steps
   - Maintain version-controlled data dictionary
6. **Organize Processed Outputs**
   - Create structured directories: `processed/` , `features/` , `plots/`
   - Use standard formats: HDF5 (epochs), Parquet (features), PNG/HTML (plots)

7. **Automated QC Pipeline**
    - Develop scripts to validate data integrity automatically
    - Flag trials with excessive artifacts, timing errors, or missing data
8. **Standardize Naming Conventions**
    - Establish consistent file naming across all data types
    - Use semantic versioning for datasets
9. **Backup Strategy**
    - Implement redundant backups (currently single OneDrive location)
    - Consider institutional data repository for long-term archival
10. **Integrate Clinical Outcomes**
    - Link EEG features to patient outcomes (Glasgow Outcome Scale, CRS-R)
    - Maintain secure PHI-compliant metadata database

11. Next Actions (Immediate Priorities)

> 1. **Sync EDF/WAV from OneDrive** and lock a unified schema for all trial logs.
> 2. **Reconcile duplicate/overlapping `patient_df` versions;** deduplicate sessions.
> 3. **Align events to signals;** add QC flags (artifacts, completeness, timing).
> 4. **Generate processed/feature tables** (PSD for command paradigms; language-tracking markers) with a clear data dictionary and versioning.
> 5. **Document storage layout** for processed outputs (`processed/` , `features/` , `plots/`) to keep analyses reproducible.

12. Data Access Example (Python)

```
import mne import pandas as pd # Load EEG data eeg_file = 'EEG
Project Data/EEG/edf/CON008_clipped.EDF' raw =
mne.io.read_raw_edf(eeg_file, preload=True) # Load stimulus timing
stim_file = 'EEG Project Data/EEG/CON008_2025-08-
```

```
14_stimulus_results.csv' stim_df = pd.read_csv(stim_file) # Filter
for language trials language_trials = stim_df[stim_df['trial_type']
== 'language'] # Create events array from stimulus timing events =
mne.make_fixed_length_events(raw, start=0, duration=1.0) # Epoch data
based on trial timing event_id = {'language': 1, 'oddball': 2,
'command': 3} epochs = mne.Epochs(raw, events, event_id, tmin=-0.2,
tmax=1.0, baseline=(None, 0)) # Compute power spectral density psds,
freqs = epochs.compute_psd(fmin=1,
fmax=40).get_data(return_freqs=True)
```

13. Data Governance & Privacy

Project Location: Harborview Medical Center (noted in `patient_notes.csv`)

Data Status: Active research project (latest data: October 2025)

Privacy: Patient data—handle according to **HIPAA/IRB protocols**

- All patient identifiers must remain de-identified in shared datasets
- Clinical metadata (demographics, outcomes) should be stored separately in secure, PHI-compliant database
- OneDrive access should be restricted to authorized research personnel only

Document Created: *December 10, 2025*

Version: *1.0*

Author: *AwakenAI Capstone Team*