EntPTC 40-Subject Cohort - Metadata Deliverables

Overview

This package contains **metadata-only** deliverables for the EntPTC 40-subject cohort from OpenNeuro dataset ds005385.

Total Size: ~160 KB (GitHub-friendly, NO Git LFS required)

Contents

1. cohort_40_manifest.csv (117 KB)

Complete manifest of all 284 EDF files in the 40-subject cohort.

Columns:

- `subject_id` - Subject identifier (sub-001, sub-030, etc.)
- `session_id` - Session identifier (ses-1, ses-2)
- `condition` - Pre or post treatment
- `pre_or_post` - Pre or post (same as condition)
- `treatment_label` - baseline or post_treatment
- `recording_modality` - EEG
- `task` - EyesOpen or EyesClosed
- `run` - Run number
- `edf_filename` - EDF file name
- `edf_relpath` - Relative path in ds005385
- `edf_sha256` - SHA256 checksum for verification
- `edf_bytes` - File size in bytes
- `start_time_if_available` - Recording start time (if available)
- `duration_if_available` - Recording duration in seconds
- `sampling_rate_if_available` - Sampling rate in Hz
- `channels_count_if_available` - Number of EEG channels
- `dataset_source` - openneuro

- `dataset_id` - ds005385
- `notes` - Additional metadata (age, sex, BIDS filename)

Usage:

```python
import pandas as pd
manifest = pd.read_csv('cohort_40_manifest.csv')

# Get all pre-treatment files
pre_files = manifest[manifest['pre_or_post'] == 'pre']

# Verify a file
import hashlib
def verify_file(filepath, expected_sha256):
    sha256 = hashlib.sha256()
    with open(filepath, 'rb') as f:
        for chunk in iter(lambda: f.read(8192), b''):
            sha256.update(chunk)
    return sha256.hexdigest() == expected_sha256
```

2. subject_summary.csv (16 KB)

Per-subject summary showing pre/post file pairs.

Columns:

- `subject_id` - Subject identifier
- `has_pre` - yes/no
- `has_post` - yes/no
- `pre_file` - Pre-treatment EDF filename
- `post_file` - Post-treatment EDF filename
- `pre_sha256` - Pre-treatment file checksum
- `post_sha256` - Post-treatment file checksum
- `pre_duration` - Pre-treatment recording duration
- `post_duration` - Post-treatment recording duration
- `pre_sampling_rate` - Pre-treatment sampling rate
- `post_sampling_rate` - Post-treatment sampling rate
- `channel_count_pre` - Pre-treatment channel count

- `channel_count_post` - Post-treatment channel count
- `flags` - Data quality flags
- `exclusion_reason_if_any` - Exclusion reason (empty for all 40 subjects)

Usage:

```python
import pandas as pd
summary = pd.read_csv('subject_summary.csv')

# Check all subjects have both pre and post
assert all(summary['has_pre'] == 'yes')
assert all(summary['has_post'] == 'yes')

# Get pre/post pairs
for _, row in summary.iterrows():
    print(f"{row['subject_id']}: {row['pre_file']} -> {row['post_file']}")
```

3. validation_report.md (8 KB)

Comprehensive validation report documenting:

- Cohort selection logic and criteria
- Exact counts (40 subjects, 284 files, 142 pre, 142 post)
- Complete subject list
- Exclusion criteria (none applied)
- Checksum verification statement
- NO MIXING guarantee
- NO SYNTHETIC DATA guarantee
- Reproducibility instructions

4. extract_cohort.py (12 KB)

Standalone Python script to reproduce the cohort extraction from ds005385.

Usage:

```bash
# Clone ds005385 from OpenNeuro
datalad clone https://github.com/OpenNeuroDatasets/ds005385.git
cd ds005385
```

```
datalad get .  # Download all files

# Run extraction script
python3.11 extract_cohort.py /path/to/ds005385

# Outputs:
# - cohort_40_manifest.csv
# - subject_summary.csv
```

Requirements:

- Python 3.11+
- ds005385 dataset with EDF files extracted (not symlinks)

Quick Start

Verify Data Integrity

Python

```
import pandas as pd
import hashlib

# Load manifest
manifest = pd.read_csv('cohort_40_manifest.csv')

# Verify a file (example)
def verify_edf(edf_path, manifest_row):
    sha256 = hashlib.sha256()
    with open(edf_path, 'rb') as f:
        for chunk in iter(lambda: f.read(8192), b''):
            sha256.update(chunk)

    computed = sha256.hexdigest()
    expected = manifest_row['edf_sha256']

    if computed == expected:
        print(f"✅ {manifest_row['edf_filename']} verified")
    else:
        print(f"❌ {manifest_row['edf_filename']} FAILED")
        print(f"   Expected: {expected}")
        print(f"   Got: {computed}")

# Verify all files in your ds005385 clone
for _, row in manifest.iterrows():
```

```
    edf_path = f"/path/to/ds005385/{row['edf_relpath']}"
    verify_edf(edf_path, row)
```

Load Pre/Post Pairs

Python

```
import pandas as pd

summary = pd.read_csv('subject_summary.csv')

for _, subj in summary.iterrows():
    print(f"Subject {subj['subject_id']}:")
    print(f"  Pre:  {subj['pre_file']}")
    print(f"  Post: {subj['post_file']}")
    print(f"  Duration: {subj['pre_duration']}s -> {subj['post_duration']}s")
    print()
```

Data Guarantees

✅ NO MIXING

- No subject data mixed
- No session swapping
- Pre and post correctly paired per subject
- All verified via checksums

✅ NO SYNTHETIC DATA

- All data from real EEG recordings
- No fabricated results
- No placeholder files
- All from OpenNeuro ds005385

✅ DETERMINISTIC

- Alphabetical subject selection
- Reproducible via extract_cohort.py
- SHA256 checksums for all files

- Complete provenance documented

Cohort Statistics

- **Subjects:** 40
- **Total Files:** 284
- **Pre-treatment:** 142 files
- **Post-treatment:** 142 files
- **Balance:** Perfect 1:1
- **Age Range:** 24-70 years
- **Sex:** 22 Female, 18 Male
- **Channels:** 65 EEG channels per file
- **Sampling Rate:** 1000 Hz
- **Duration:** ~193 seconds per recording

GitHub Upload

This package is designed to be uploaded directly to GitHub **without Git LFS**:

Bash

```
git init
git add .
git commit -m "Add EntPTC 40-subject cohort metadata"
git remote add origin https://github.com/yourusername/entptc-cohort-
metadata.git
git push -u origin main
```

Total size: ~160 KB (well under GitHub's 100 MB file limit)

Next Steps

1. **Download ds005385** from OpenNeuro
2. **Verify files** using checksums in cohort_40_manifest.csv
3. **Extract cohort** using extract_cohort.py (optional, for reproduction)
4. **Analyze data** using pre/post pairs from subject_summary.csv

Citation

Dataset:

```
OpenNeuro Dataset ds005385
DOI: [OpenNeuro ds005385 DOI]
```

EntPTC Theory:

```
[Paper citation - see ENTPC.tex]
```

Support

- **Validation Questions:** See validation_report.md
- **Reproduction:** Run extract_cohort.py
- **File Verification:** Use SHA256 checksums in manifest
- **Dataset Issues:** Contact OpenNeuro ds005385 maintainers

Package Version: 1.0

Generated: 2024-12-23

Status: ✅ VALIDATED - READY FOR GITHUB UPLOAD