yi_julia_progress_report1

November 13, 2025

1 Final Project Progress Report — HBN Cognition × Temporal Discounting

Student: Julia Yi • **Repo:** https://github.com/beaverjuly/hbn_project/ • **Date:** Nov 13, 2025

1.1 Scope update

I will test whether **clinical diagnoses are associated with specific cognition–temporal-discounting patterns** in the Healthy Brain Network (HBN). Features include temporal discounting parameters (k, log-k, ED50) and NIH Toolbox cognition (Flanker, Processing Speed, List Sorting), with age/sex covariates.
Why change now: My lab **gained access to HBN yesterday**, which aligns directly with my focus on transdiagnostic mechanisms and provides sufficient sample size for robust ML.

1.2 Data sources & access

- **HBN Phenotype (public pheno CSV):** Accessed via **HTTP programmatically** (`requests` → `pandas.read_csv`) to satisfy the API/web requirement.

- **HBN Clinician-Consensus Diagnosis:** Downloaded CSV from the HBN portal; merged locally (not committed).

- **HBN NIH Toolbox & Temporal Discounting tables:** Local CSVs; merged to an interim master, then cleaned into **processed** analysis views.

API proof: `tests.py` performs an HTTP GET of the public pheno CSV and asserts load/shape/ID columns; it also runs my pipeline and verifies processed outputs (row count, required TD columns, missingness thresholds).

1.3 What I implemented

- **Pipeline script (`hbn_data_processing_pipeline.py`):**
  - Ingest (API + local), ID normalization, joins.

  - Feature engineering: `k_mean`, `logk_mean`, `ed50_mean`, `k_abs_diff`.

- Missingness policy: drop 90%-missing columns; build **core** (20% missing) and **extended** views; require 90% row completeness in core.
- Export to `data/processed/` (CSV), plus validation metadata.

- **Tests (`tests.py`):** API fetch test; pipeline run test; processed-file assertions (TD features present, N 1000, 10% missing in key TD).

1.4 Planned analysis (1-month, laptop-feasible)

1. **Unsupervised subtypes:** KMeans/Agglomerative (k=3–5) on z-scored TD+NIH; select k via silhouette/Calinski–Harabasz; bootstrap stability.

2. **Diagnosis enrichment:** ^{2}/Fisher across clusters; Cramér's V; effect-size profiles of TD/NIH by cluster; sensitivity with age/sex adjustment.

3. **Supervised confirmation (lightweight):** L1-logistic (balanced) predicting selected diagnoses; 5-fold AUROC/AUPRC; permutation importance; calibration.

4. **Reporting:** Heatmaps of cluster centroids, PCA/UMAP visuals, concise enrichment tables and model metrics.

1.5 Risks & mitigations

- **DUA timing / portal reliability:** API requirement is already satisfied by public pheno HTTP fetch; analysis proceeds with locally joined tables (not committed).

- **Severe missingness (Picture Sequence):** Excluded from core view; focus on Flanker and Processing Speed with low missingness.

- **ID heterogeneity (EID vs Identifiers):** Normalized to `_EID` prior to merges.

1.6 Reproducibility

- `data/` and `results/` are **gitignored**; only code, small samples (if any), and this report are committed.

-

1.7 `environment.yml` documents dependencies; `README.md` explains how to run `tests.py`.

1.8 Data sources

Data source #	Name / short description	Source URL	Type	List of fields (key)	Format	Tried Python?	Estimated data size / points to use
1	**HBN Phenotype (R1–R11)** — demographics & screening variables. Pulled programmatically.	http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/File/_ (e.g., HBN_R11_Pheno.csv)	**API** (HTTP download)	EID, Sex, Age, selected screening totals (e.g., EHQ_Total)	CSV	**Yes** (requests → pandas)	2,500–3,000 rows available; plan to use ~2,100 matched
2	**Clinician Consensus Diagnosis** — DSM-style best-estimate diagnoses.	Local export: data/raw/Diagnosis_ClinicianConsensus.csv	**file**	EID, DX_01, DX_01_Cat, DX_02..., flags/notes	CSV	**Yes** (pandas read locally)	4,700 rows; plan to use ~2,100 matched
3	**NIH Toolbox (Cognition)** — attention/inhibitory control, cognitive flexibility, processing speed, working memory, vocabulary.	Local export: data/raw/NIH_final.csv	**file**	Flanker, DCCS, Pattern Comparison, List Sort, Picture Vocabulary scores (raw/standard), admin flags	CSV	**Yes** (pandas)	3,000 rows; plan to use ~2,100 matched

Data source #	Name / short description	Source URL	Type	List of fields (key)	Format	Tried Python?	Estimated data size / points to use
4	**Temporal Discounting** — delay-reward task features.	Local export: `data/raw/Temp_Disc_Final.csv`	**file**	k (discount rate), AUC, model fit stats, task/visit metadata	CSV	**Yes** (`pandas`)	2,400 rows; plan to use ~2,100 matched

All sources exceed 300 records; source #1 satisfies the "API/web scraping" requirement.