NPIS: Characterizing What Scientific Data Can Support

Nonparametric Information Sufficiency

M. Antony Ewing, PhD
Conquer Medical Health
ORCID: 0009-0005-1550-1718

Technical Report v1.0 – 4 September 2026

Abstract

Nonparametric Information Sufficiency (NPIS) is the information-characterization layer of CMH Scientific. It asks which distinctions a declared data substrate contains for a declared task before limitations are attributed to model architecture. The September 2026 public release separates a callable interface for arbitrary user data from frozen validation evidence. The canonical direct five-cohort analysis contains 56,199 individuals and 408,014 eligible longitudinal states across ADNI, PPMI, PRO-ACT ALS, HRS, and Add Health. At the primary $K = 25$ support definition, NI20 ranges from 64.9% to 98.7%, while materially opposed future-direction support ranges from 59.4% to 81.9%. This report defines the information-relative formulation, the public call contract, step-by-step use, validation, and the boundary between finite empirical falsification and unsupported global identifiability claims.

Public-release note. This revision aligns the technical report with the September 2026 two-layer software release: a generic callable public interface for new analyses and a separate frozen validation layer for previously reported evidence. Protected operational implementations are not distributed.

1 Purpose and information-relative formulation

NPIS begins upstream of model selection. Let $\mathcal{G} = \sigma(X, A)$ denote the information generated by the observed state X and any declared exposure or intervention variables A. A target T is identifiable relative to $\mathcal{G}$ when the target is measurable with respect to that information set (up to null sets when a probability measure is present). Probability can quantify unresolved branch mass; the containment question itself is logically prior.

For the direct local analysis, let p_{id} be the empirical mass assigned to future-direction branch d within the declared neighborhood of state i. Direct branch mass is

$$B_i = 1 - \max_d p_{id}.$$

NI20 is the fraction of assessed states for which $B_i \geq 0.20$. The opposed-direction statistic is stricter: it requires material local support for both negative and positive realized future directions. These quantities are kept separate from deeper branching/separation diagnostics, temporal benefit, future-added-information, deformation, reversal, and residual-direction objects.

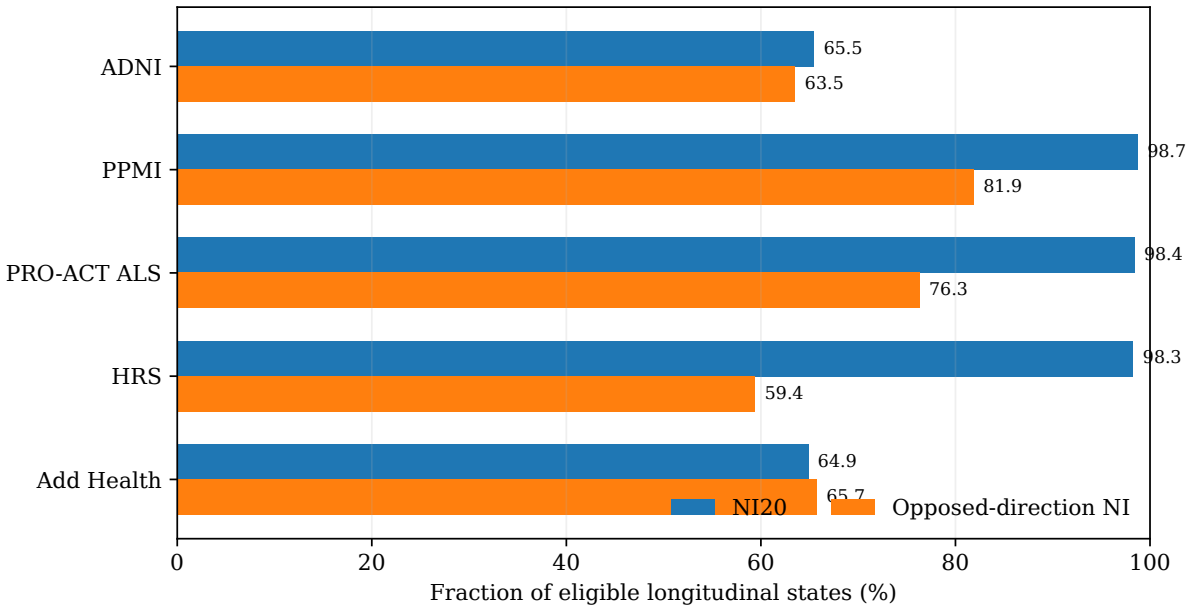

Figure 1: Canonical five-cohort direct validation at $K = 25$. The revised public validation fixture reports the frozen cohort-specific values rather than only a range.

2 Two-layer public release

Layer 1 – generic callable interface. The public `npis/` package accepts a new user's data, declared task, schema, horizon, and requested outputs. It sends the specification through an authorized transport and returns a standardized `NPISResult`. Frozen cohort values are not embedded in the generic calculation path.

Layer 2 – frozen validation. `validation/frozen\_five\_cohort.json` preserves the canonical five-cohort results as an auditable regression/validation artifact. It is evidence about the released analyses, not a substitute for computation on new data.

3 Step-by-step use

1. **Declare the scientific claim before choosing a model.** State the target, horizon, independence unit, and whether the question concerns population ranking, individual direction, temporal refinement, or intervention-sensitive behavior.
2. **Map the data schema.** Identify subject, time, state, exposure/intervention, and target columns. Do not silently infer an endpoint or time horizon.
3. **Request information diagnostics.** Ask for direct branching, opposed futures, and any deeper temporal/deformation diagnostics required by the scientific claim.
4. **Run through an authorized engine transport.** The public client handles the request/response contract; the protected estimator performs the computation.
5. **Interpret the claim map.** Separate supported population/subgroup statements from unsupported individual or causal claims. If the desired claim remains unresolved, hand tested candidate measurements to Acquire rather than assuming that more of the same data will solve the problem.

Python public call

```python
from npis import NPISClient

client = NPISClient(transport=my_authorized_transport)

result = client.assess(
    data=my_longitudinal_data,
    task={"id": "future_direction",
          "horizon": "next_observation"},
    schema={
        "subject_col": "patient_id",
        "time_col": "visit_date",
        "state_cols": ["marker_a", "marker_b", "age"],
        "value_col": "marker_a",
    },
    options={"requested_outputs":
             ["branching", "opposed_futures"]},
)

print(result.information_statement)
print(result.metrics)
print(result.claim_map)
```

Example prompt

Characterize the information in these longitudinal records for individual 36-month trajectory prediction. Report population ranking and individual future direction separately, quantify unresolved/opposed futures, and do not infer causal mechanism from observational structure.

Structured result

RESULT: The represented substrate can contain useful population-level directional information while leaving a material subset of individual future directions unresolved.
SUPPORTED: claims stable under the declared information set and validation design.
NOT SUPPORTED: universal individual trajectory identification when opposed futures remain in local support; causal mechanism without a causal design.
NEXT ACTION: if the requested claim remains unresolved, test richer history, another modality, or a revised target definition and pass admissible candidates to Acquire.

4 Canonical direct validation

Cohort	Individuals	Eligible states	NI20	Opposed NI
ADNI	2,257	10,013	65.45%	63.53%
PPMI	3,852	14,634	98.73%	81.88%
PRO-ACT ALS	5,400	44,927	98.38%	76.30%
HRS	38,375	238,726	98.27%	59.41%
Add Health	6,315	99,714	64.87%	65.73%
Total	56,199	408,014	–	–

Table 1: Frozen canonical cross-patient direct analysis in the public validation layer.

The direct analysis excludes same-unit neighbors and uses primary $K = 25$. The key scientific point is not that the state contains no information; rather, the same represented state can remain compatible with materially different realized futures. Direct nonidentifiability and opposed-future support therefore quantify a limitation in the represented substrate-task pair, not model incompetence.

5 Boundaries and reproducibility

NPIS is task-, representation-, support-, and finite-sample-specific. A positive incompatibility witness can falsify uniqueness on tested support. Failure to observe branching in finite data cannot prove global identifiability. Neighborhood choice approximates observational equivalence and requires sensitivity analysis. NPIS does not recover a latent biological law or substitute for causal design.

Public repository: github.com/ConquerMedical/NPIS. The public package includes the callable client, result schema, examples, frozen validation fixture, and automated interface tests on Python 3.10–3.12. Protected estimators remain behind the public interface.

6 Installation, validation check, and result contract

A fresh public clone is intended to verify the interface boundary before a user connects an authorized computational transport.

Python public call

```
git clone https://github.com/ConquerMedical/NPIS.git
cd NPIS
python -m venv .venv
source .venv/bin/activate
pip install -e ".[test]"
pytest -q
python examples/generic_call.py
python examples/frozen_validation.py
```

The release tests the public client on Python 3.10, 3.11, and 3.12 and explicitly verifies that the original neighborhood-estimator modules are absent from the public package.

Result field	Meaning
task_id	Caller-declared scientific task identifier.
state	Engine state for the declared substrate-task pair.
information_statement	Human-readable statement of what the represented information supports.
metrics	Task-specific information diagnostics returned by the authorized engine.
claim_map	Explicit supported/not-supported scientific claims.
tested_domain	Representation, support, horizon, and validation scope to which the result applies.
engine_version	Provenance for the protected computation that generated the response.

Table 2: Public NPIS result contract.

7 Validation scope beyond the frozen five-cohort fixture

The broader NPIS research program reported in the source white paper spans 13 human substrates and approximately 694,095 states or observations. The public v0.4 frozen fixture deliberately narrows the regression target to the canonical direct five-cohort analysis because those cohort-specific values are the most stable, directly interpretable public validation objects. Additional neighborhood, temporal, and deformation sensitivity analyses remain scientifically relevant but should not be conflated with the frozen direct metric table.

Intellectual property, competing interests, funding, and use

Relevant methods and computational systems are the subject of previously filed U.S. provisional patent applications. This report discloses scientific objects, public interfaces, empirical validation logic, and results needed to evaluate and use the public research architecture. Proprietary operational implementations, calibration internals, candidate-generation procedures, search/ranking policies, deployment orchestration, and other system components not necessary to evaluate the public scientific claims remain outside the scope of disclosure. Publication does not grant a patent or commercial software license.

The author is Founder of Conquer Medical Health and has financial and intellectual-property interests in the technologies described. No external funding supported preparation of this technical report.

Use of AI-assisted tools

AI-assisted tools were used for editorial and software-documentation support. The author determined the scientific claims, supplied the underlying analyses and validation artifacts, and reviewed the numerical content and final report.

References

[1] Ewing, M. A. *Widespread Nonidentifiability in Patient Data Limits AI Prediction Regardless of Model Sophistication*. 2026.

[2] Ewing, M. A. *TopolAI: Information-First Scientific AI*. Conquer Medical Health, 2026.

[3] Ewing, M. A. *OptiCeil: Task-Conditioned Recoverability*. Conquer Medical Health, 2026.

[4] Ewing, M. A. *Acquire: Inverse Measurement Design for Scientific AI*. Conquer Medical Health, 2026.

[5] Ewing, M. A. *AI on Alzheimer Disease: Common-Substrate Reconstruction and Adjudication*. 2026.

[6] Cover, T. M. and Hart, P. E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1), 21–27 (1967).

[7] U.S. Department of Health and Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule.

[8] Kaggle. Notebooks documentation.

[9] Hugging Face. Hub repositories, licenses, Paper Pages, Collections, and Spaces documentation.

[10] Modal. Documentation for serverless compute, GPU workloads, and secrets.