#healthydata

From Data to Evidence

Intelligent Machine Learning Anomaly Detection in Clinical Research

Powered by the CLADE-IS Electronic Data Capture Platform

The Quest for #HealthyData

Adherence to Evidence-Based Medicine relies heavily on data quality. Real-World Data (RWD) collected in clinical registries provides vital answers unaddressed by Randomized Controlled Trials.

Carelessness

Incomplete fields, mistyped inputs

Systematic Error

Investigator misinterpretation, structural biases

Fraud

Intentional fabrication or falsification

Biased evidence generates harmful health decisions. Assuring the highest data quality is a systemic necessity.

The Limits of Traditional Monitoring

Conventional electronic case report form (eCRF) edit checks prevent basic invalid data entry. However, as clinical complexity scales, validation procedures designed by data managers become overly complicated and prone to error.

Alert Fatigue

Complex edit checks generate unintelligible alert messages for clinical investigators.

Statistical Blindspots

Standard statistical monitoring often fails on small sample sizes or rare anomalies.

Cost & Scale

Exhaustive source data verification and manual on-site visits are no longer economically viable.

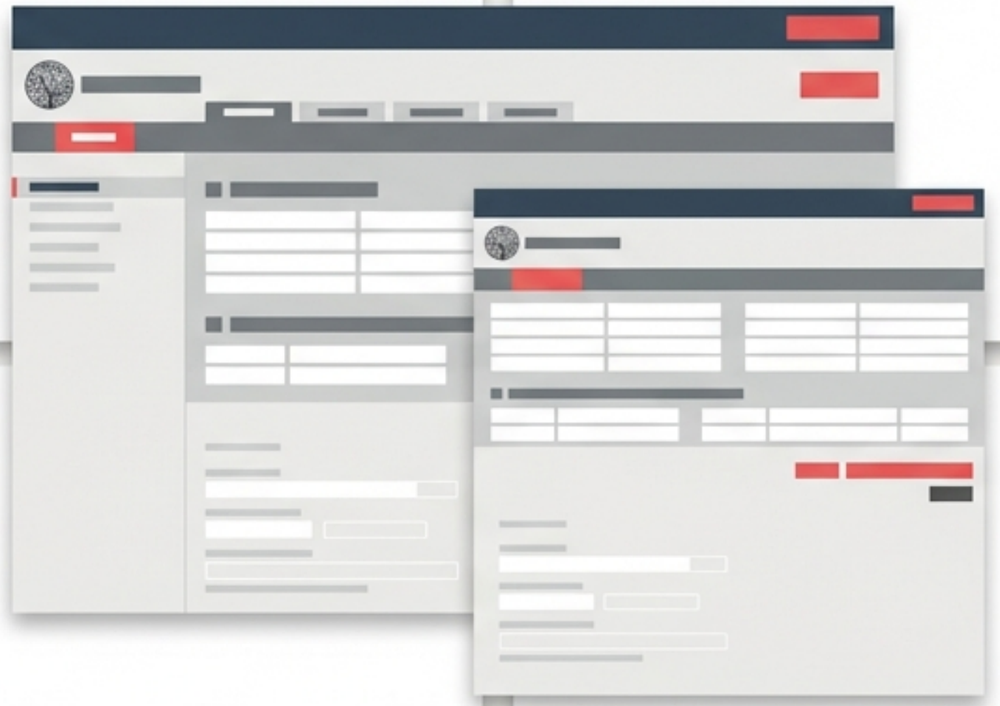
Enter CLADE-IS: Built by a CRO, for CROs

CLADE-IS
CLINICAL DATAWAREHOUSE

A robust, modular, web-based Clinical Data Warehousing Information System serving numerous clinical specialties. Designed for secure, intuitive data management.

Designer

Robust eCRF builder with skip logic and custom cross-form validation.



Adminer

Flexible data access management with unlimited hierarchy of sites.

Studies

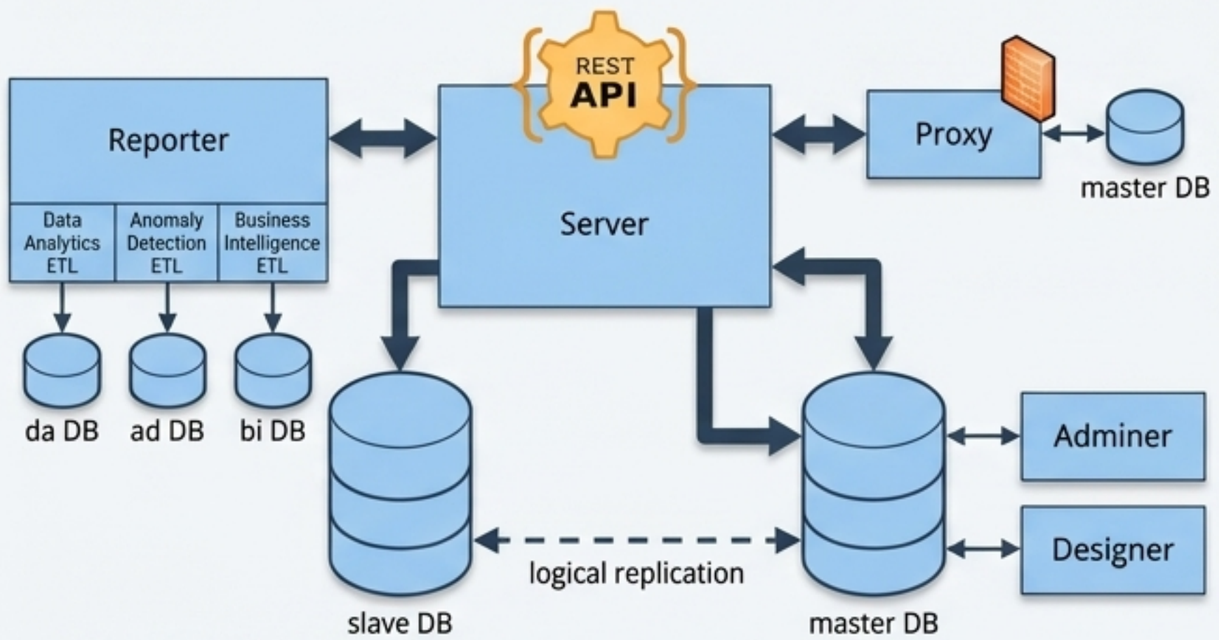
Web-responsive dashboards, audit trails, and SAE management.

Reporter

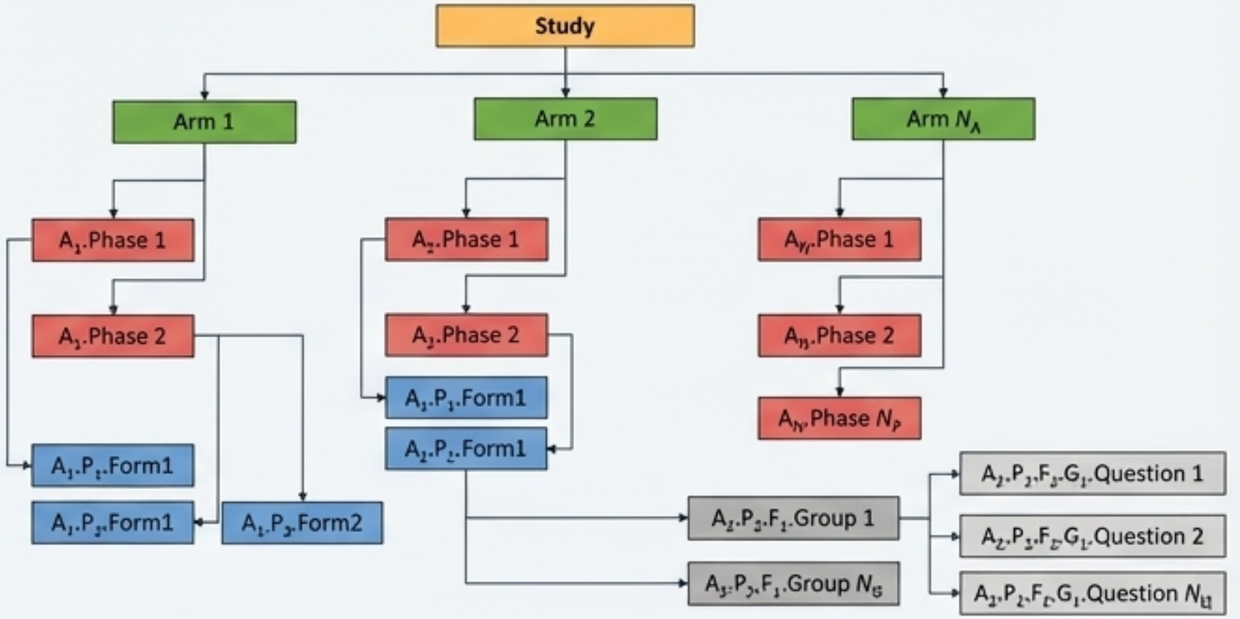
Continuous data reporting, complete custom extracts, and smart filtering.

The Architecture of #HealthyData

System Architecture Diagram



Data Model Hierarchy



Callout 1 (The Ecosystem)

5 mutually communicating components (Proxy, Server, Adminer, Designer, Reporter) utilizing a REST API and distributed Master/Slave databases to ensure data integrity and GDPR compliance.

Callout 2 (The Entity-Attribute-Value Model)

Efficiently encodes sparse clinical registry data.

Callout 3 (JSON Data Structure)

Data is securely stored and easily parsed (e.g., "Q10":{"value":63,"state":"done"}).

Beyond Standard Monitoring: Intelligent Anomaly Detection

Integrated directly into the CLADE-IS Reporter module, an automated machine learning algorithm identifies anomalous patterns caused by systematic error, carelessness, or fabrication.

1. Translating multi-dimensional clinical forms into analyzable vectors.

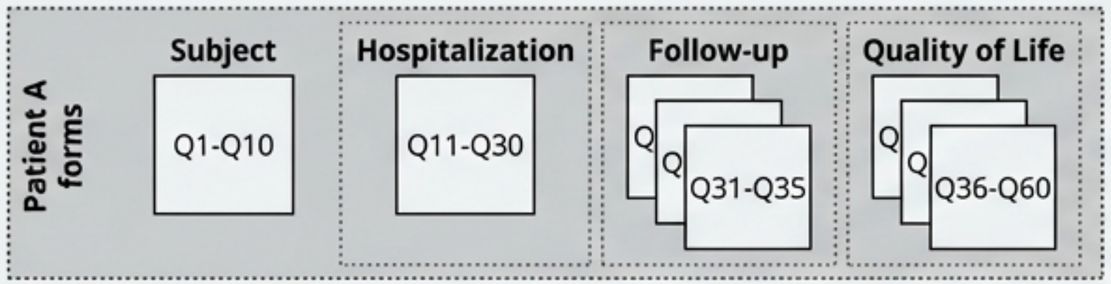
2. Unsupervised machine learning measuring distance from normal clinical clustering.

3. Automatic post-hoc univariate testing generating targeted queries for data managers.

Step 1: Solving the Dimensionality Problem

The algorithm requires merging all eCRF questions into one flat-wide table. But patients fill out repeating 1:N forms (e.g., multiple Follow-ups), which would misalign variables in standard merging.

The Innovation: Semi-Flattened Tables



Prefix Table
(Single-instance forms like Subject Demographics)

Repeating Instance
(e.g., Follow-up Form 1, Follow-up Form 2)

	Q1	Q2	...	Q10	Q11	Q12	...	Q30	Q31	Q32	...	Q35
Patient A	$V_{1,1}$	$V_{1,2}$	...	V_{10}	$V_{1,1}$	$V_{1,2}$	...	V_{30}	V_{31}	V_{32}	...	V_{35}
Patient B	$V_{1,1}$	$V_{1,2}$	...	V_{10}	$V_{1,1}$	$V_{1,2}$	...	V_{30}	V_{S1}	V_{32}	...	V_{35}

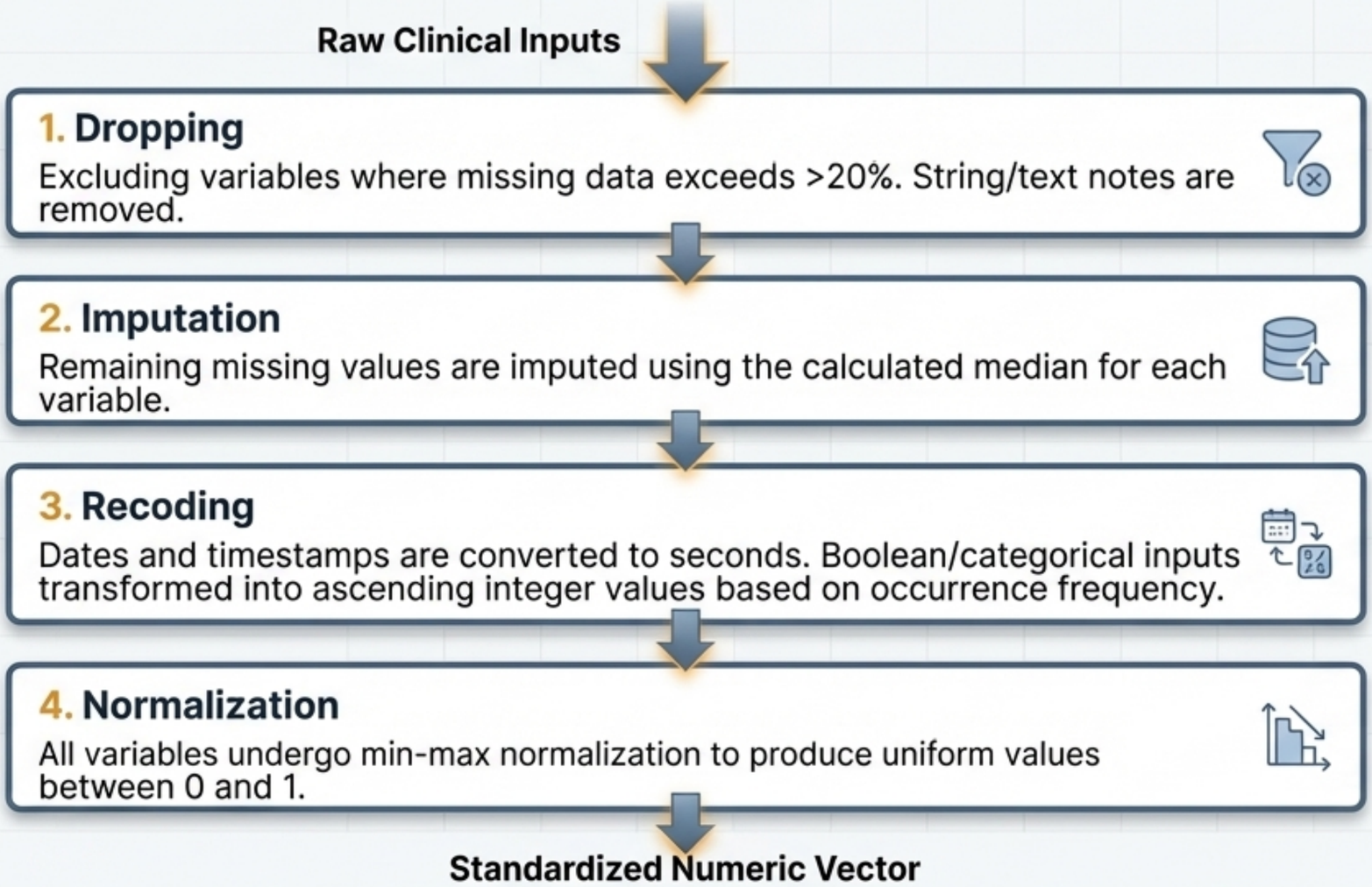
$N_{sw} = 2$

	Q1	Q2	...	Q10	Q11	Q12	...	Q30	Q36	Q37	...	Q60
Patient B	$V_{1,1}$	$V_{1,2}$	...	V_{10}	V_{11}	V_{12}	...	V_{30}	V_{36}	V_{37}	...	V_{60}
Patient B	$V_{1,1}$	$V_{1,2}$	...	V_{10}	V_{11}	V_{12}	...	V_{30}	V_{36}	V_{37}	...	V_{60}

Result: A system that runs the algorithm independently across multiple meaningful combinations, ensuring precise data alignment.

Step 2: The Preprocessing Pipeline

Raw clinical inputs are automatically transformed into normalized feature vectors, discarding unusable text while preserving vital numeric and categorical context.



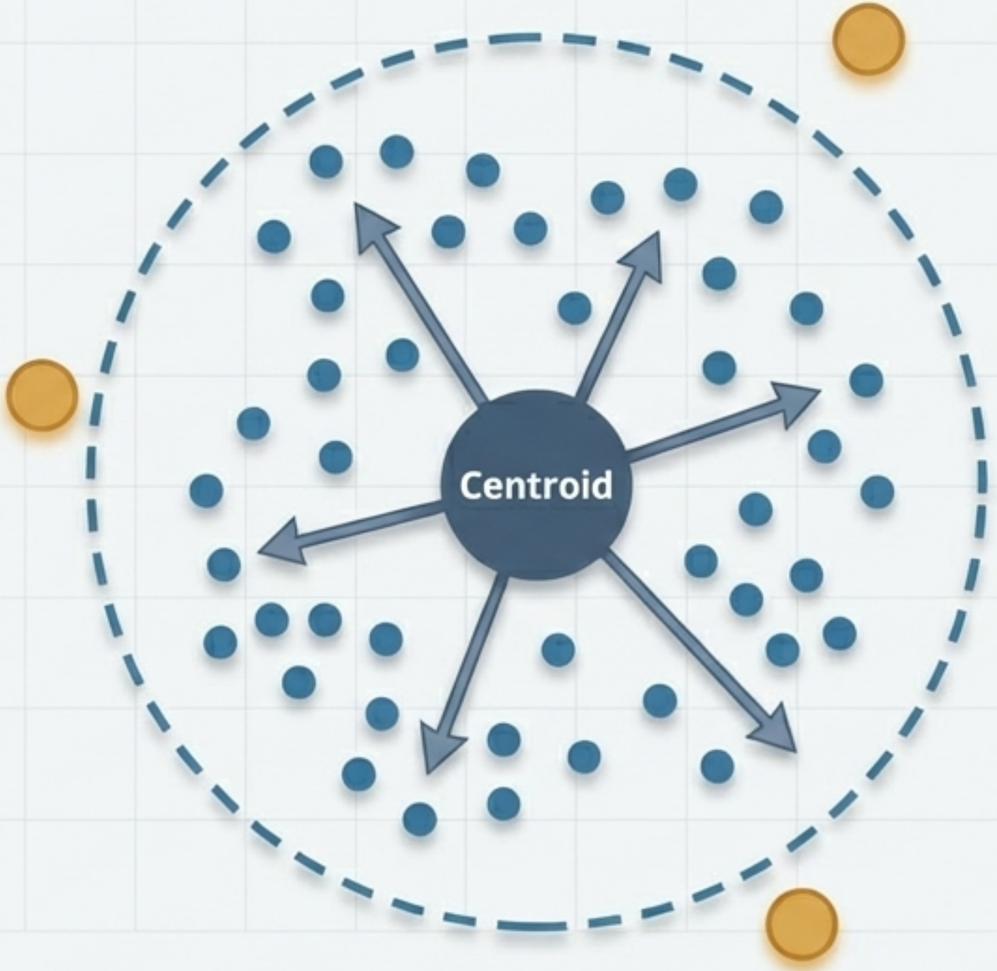
Step 3: Distance Metrics & Thresholding

The algorithm treats all patients as a single cluster. Anomaly detection identifies objects whose feature vectors fall too far from the cluster centroid.

The 7 Distance Metrics Tested:
Canberra, Chebyshev, Cosine, Euclidean, Manhattan, Mahalanobis, and Minkowski.

The Threshold Rule:
Thresholds are dynamically calculated using preset percentiles and the Interquartile Range (IQR) rule.

The Output: The strength of an anomaly is determined by the number of metrics that flag it, automatically triggering investigative queries in CLADE-IS.



Peer-Reviewed Validation: The Simulation Experiment

To establish ground-truth knowledge, the algorithm was evaluated against authentic Real-World Data from five diverse neuroscience clinical registries running on CLADE-IS.

The Methodology

Simulated Anomalies

A randomized 1% of all preprocessed data cells were intentionally manipulated.

Transformation

Normal distributions were shifted to a mean of 6σ. Non-normal distributions were swapped for extreme values (frequency <10%).

Real-World Triggers

Manipulated data still had to pass the built-in CLADE-IS automatic edit checks, simulating a human investigator creatively fabricating or mistyping plausible data.

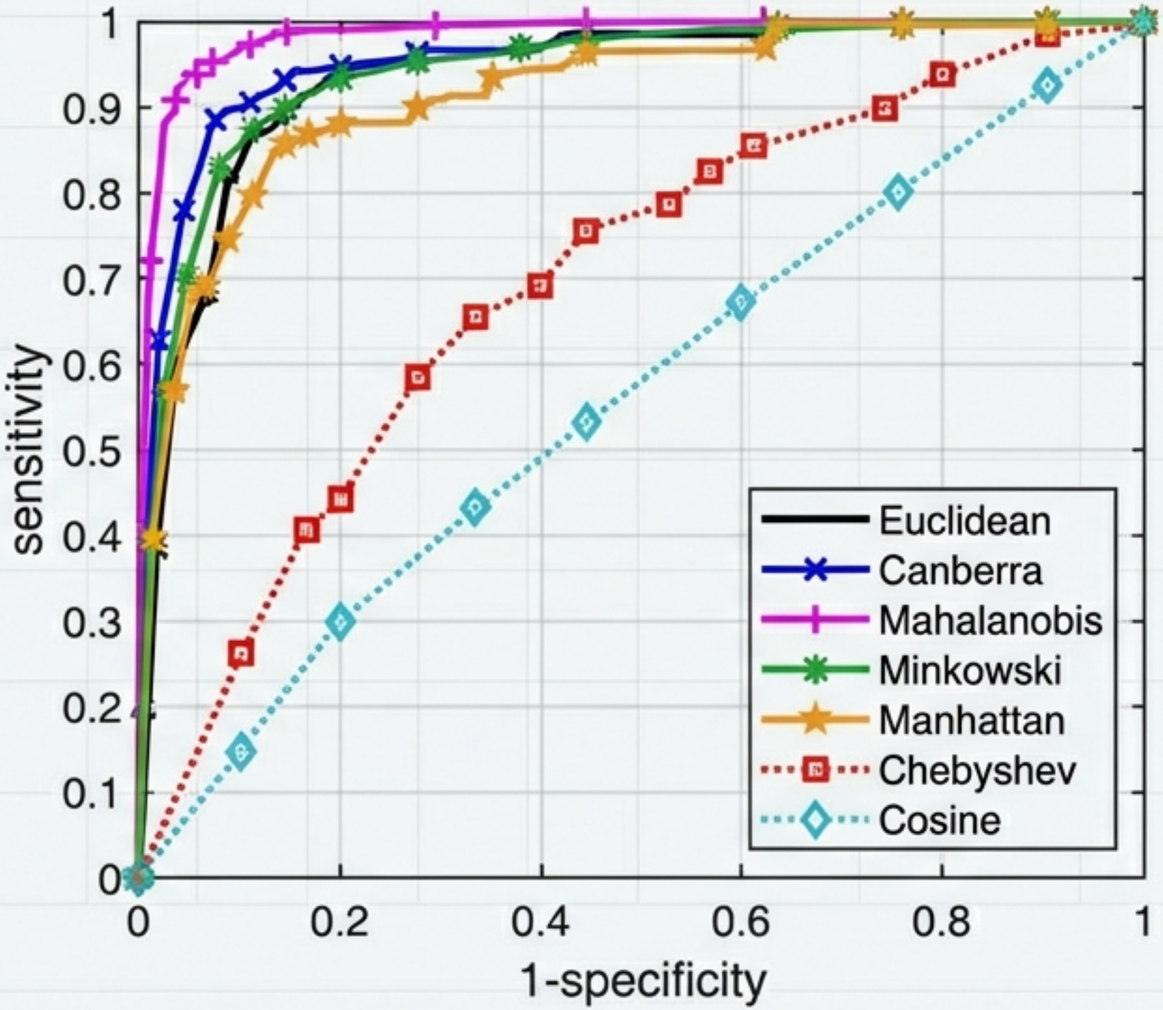
Finding the Optimal Detection Formula

Extensive ROC curve analysis tested all single distance metrics across 81 percentile thresholds to maximize accuracy and the Youden index.

The Results

Eliminated:
Cosine and Chebyshev metrics were discarded due to poor independent performance.

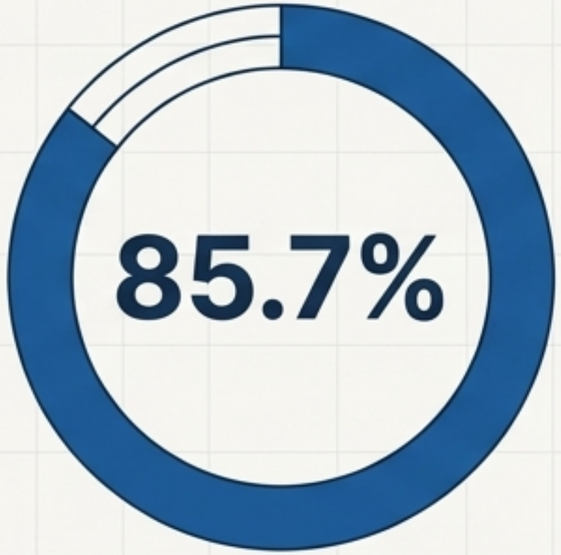
The Winning Ensemble:
Combining the Mahalanobis, Manhattan, and Canberra metrics generated the highest multi-metric detection performance, outperforming any single metric alone.



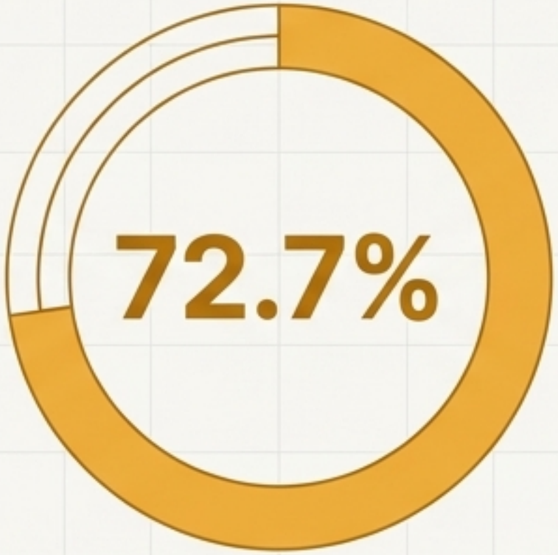
Real-World Results: Uncovering the Invisible

Validated against an independent 5-year non-interventional schizophrenia registry dataset, the fine-tuned ensemble algorithm delivered outstanding identification of anomalous records.

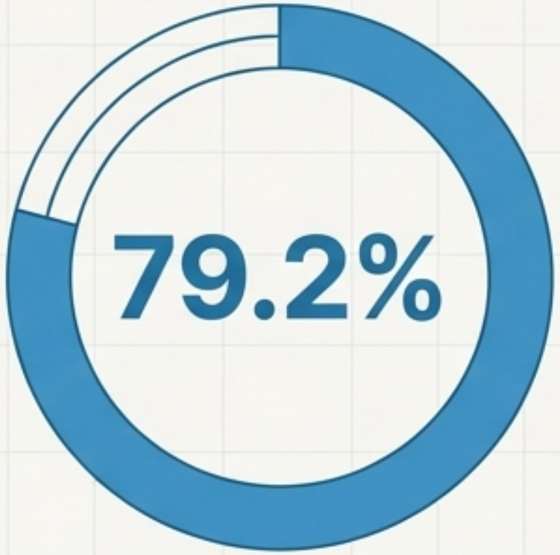
Key Metrics (Mahalanobis + Manhattan + Canberra Ensemble)



Sensitivity
True Positive Rate



Specificity
True Negative Rate



Balanced Accuracy

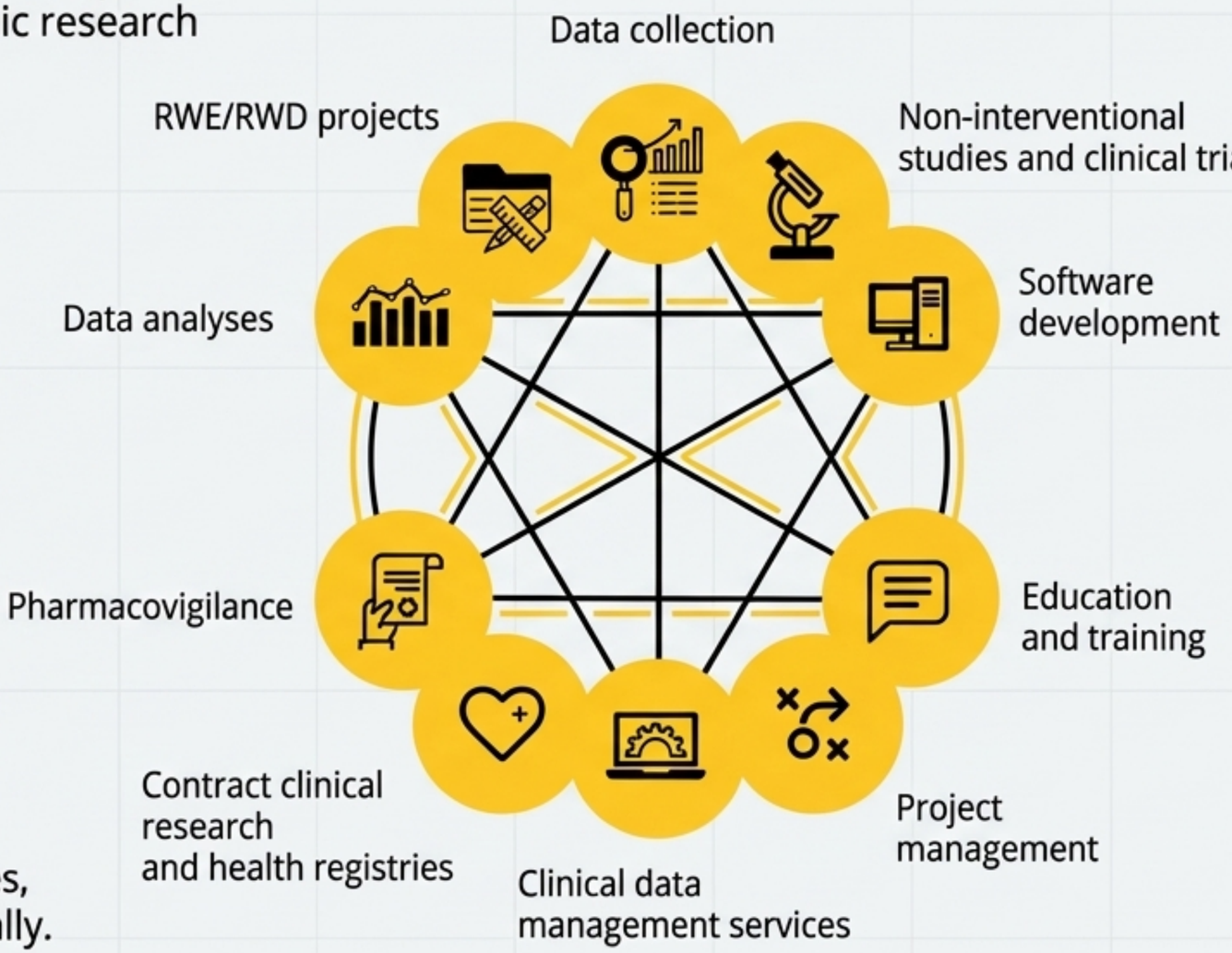
Takeaway: The experimental results demonstrate that the algorithm is universal in nature... capable of anomalous data detection with a sensitivity exceeding 85%. (JMIR Med Inform 2021;9(5):e27172)

CRO Excellence Powered by Own Tech

The Institute of Biostatistics and Analyses (IBA) is a Masaryk University spin-off transferring 15+ years of academic research into industry-leading clinical data management.

The IBA Advantage

- ecrin ECRIN-Certified Data Center
- Clinical Data Management & Biostatistics
- RWE/RWD Projects & Health Registries
- Pharmacovigilance & Software Development



We are an essential partner for medical expert societies, pharmaceutical companies, and academic teams globally.

Flexible Licensing for Every Scale

Tier 1: CLADE-IS Light (Academic)

For independent small/mid-sized academic teams.

- 1 Study, up to 5 sites / 10 users.
- Unlimited arms, forms, data points, multicentric collaboration.

Tier 2: CLADE-IS Pro (Sponsor)

For large research teams supported by sponsors.

- 1 Study, up to 99 sites / 200 users.
- Adds SAE management, advanced dashboarding, and custom reports.

Tier 3: CLADE-IS Premium (CROs & Institutions)

For Contract Research Organizations.

- Unlimited studies, sites, and users.
- Adds RESTful API, integration support, and custom feature tailoring.

Your Partner for #HealthyData

> As we have thought about the design of these studies in our clinic for so long and thoroughly, we happily avoided collecting data on paper or in Excel – a variant which in the past led to many errors... Now, we are enthusiastic about the possibilities of CLADE-IS.

Prof. Tomáš Kašpárek, MD, PhD
Head of Psychiatric Clinic & Vice-Dean for Science, Masaryk University.

Institute of Biostatistics and Analyses Ltd.

Website: www.biostatistika.cz

Email: info@biostatistika.cz

Phone: (+420) 515 915 101

#healthydata