NHIS Access-to-Care:

Forecasting Barriers and Identifying At-Risk Populations

Machine learning analysis of U.S. healthcare access, 2019–2024

Dubstech Datathon 2026 // Matrix Men

Mohriz Murad, Maaz Murad, Raunak Gupta, Shayaan Ali

https://github.com/mojipao/Datathon-2026/tree/main

Table of Contents

1 1.1, 1.2

Explanation of Process

2 2.1, 2.2, 2.3

Quality of Data Exploration

3 3.1, 3.2, 3.3, 3.4

Description of Model (Three Core Ideas)

4 4.1a, 4.1b, 4.2

Quality & Description of Model Created

5

Value of the Model Created

The Challenge

The NHIS Adult Summary Health Statistics dataset contains 26,208 survey-based estimates spanning 54 **health topics** and roughly 75 **demographic subgroups** from 2019 to 2024. The challenge is not only to measure access barriers such as delayed care or unmet needs, but to **predict how these barriers will evolve and identify which populations are most at risk of falling through the cracks**. Our goal is to use machine learning to move from retrospective description to **forward-looking insight.**

Most public health analyses stop at description. Forecasting is what enables **intervention before harm occurs**.

1

Explanation of Process

Analytical Process Overview

Our analysis follows a structured pipeline: rigorous data preprocessing, careful separation of access barriers from prevalence measures, exploratory analysis to understand dispersion and inequity, and finally **predictive modeling** with uncertainty quantification. Each step was designed to preserve the survey's **statistical meaning** while enabling robust **machine learning analysis**. Survey data is fragile; each step must preserve statistical meaning or downstream results become misleading.

1.1 Data Ingestion and Preprocessing

Silent string mismatches and malformed numerics create false subgroup inflation and bias estimates without obvious errors. To prevent this, we performed the following actions to process the data:

- Loaded 26,208 NHIS Adult Summary records spanning 2019–2024
- Retained 13 analytic columns capturing demographics, time, estimates, and uncertainty
- Normalized all string fields to prevent silent duplication of subgroup keys
- Safely coerced numeric fields to handle malformed or missing entries
- Applied a three-tier filtering process to separate raw, numeric, and fully valid rows
- Final modeling dataset contains 23,609 clean observations

1.1

Recovering Statistical Uncertainty

The dataset does not populate standard errors, but every estimate includes a **95% confidence** interval. Rather than discarding uncertainty, we reconstructed standard errors directly from these intervals. This preserves survey precision and allows downstream models to weight more **reliable estimates** more heavily, preventing noisy observations from distorting trends.

Access Measure Selection

- Performed **keyword-based classification** across topics and estimate descriptions
- Separated **access barriers** from disease **prevalence measures**
- Avoided conflating **health outcomes** with **access constraints**
- Required **consistent availability** across years and subgroups
- Prioritized measures directly tied to **affordability** and **continuity**
- Selected **four high-coverage access indicators**

1.2 Analysis Pipeline

Using multiple methods allows us to best capture trends, inequality, uncertainty, and multivariate risk simultaneously. Our methods consisted of:

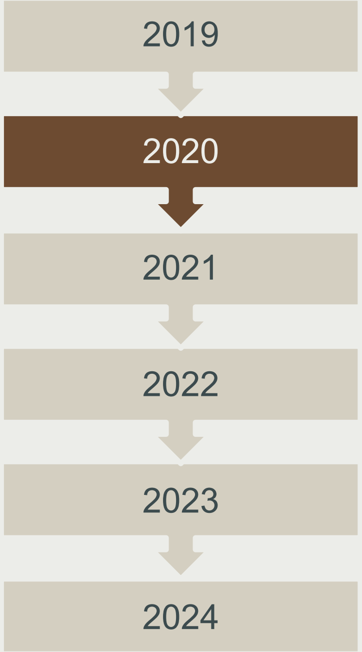
- Time-series trend analysis with confidence intervals
- Subgroup ranking using error bars to avoid false precision
- Dispersion analysis via box plots
- Composite friction score construction using z-scores
- Forecasting with weighted linear regression
- Multivariate subgroup risk identification using clustering and anomaly detection

2

Quality of Data Exploration

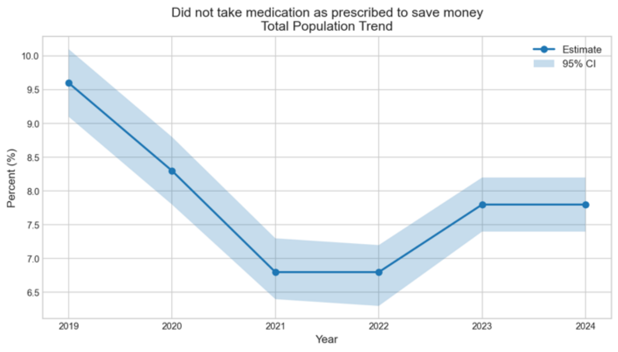
2.1 Scope of Exploration

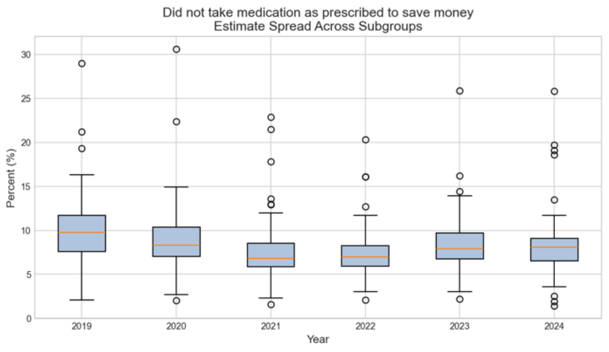
- Six-year span covering pre-pandemic, pandemic, and post-pandemic periods
- Approximately 75 demographic subgroups
- National and subgroup-level perspectives
- Explicit treatment of uncertainty throughout
- Focus on inequity, not just averages
- Designed to surface structural patterns

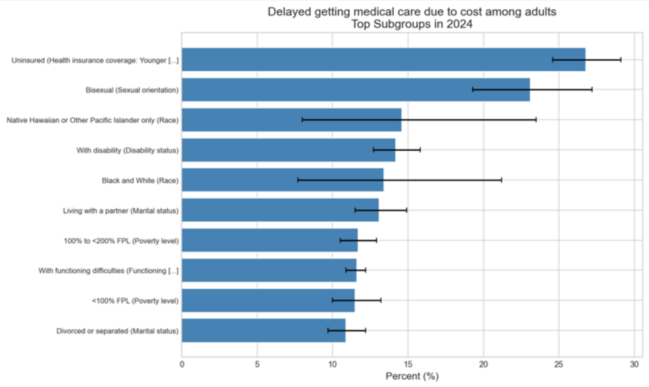


2.2 Exploratory Visualizations

- Trend plots with confidence bands for statistical significance
- Subgroup bar charts with error bars
- Dispersion plots to capture inequality spread

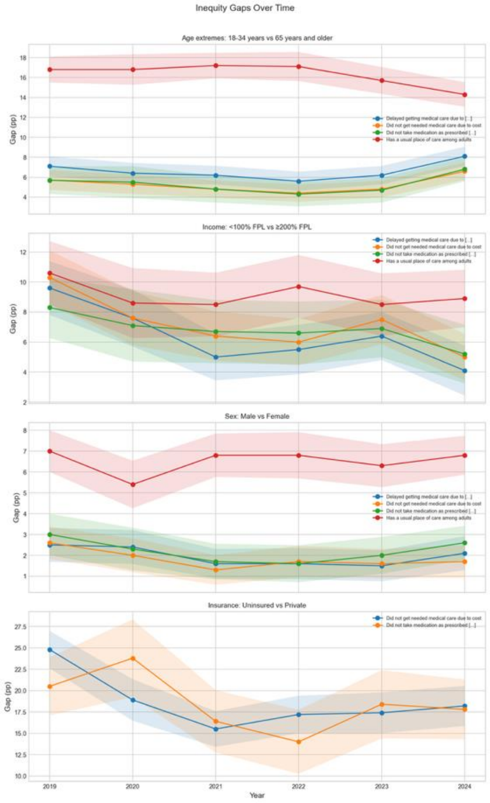


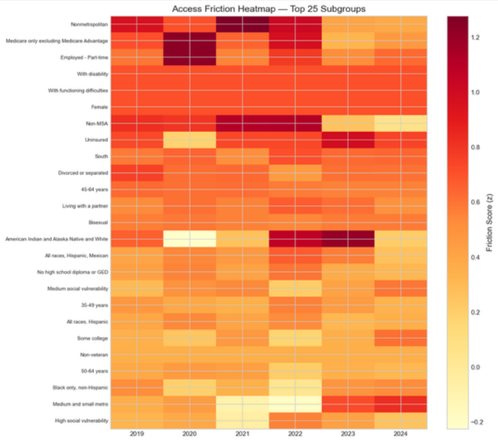


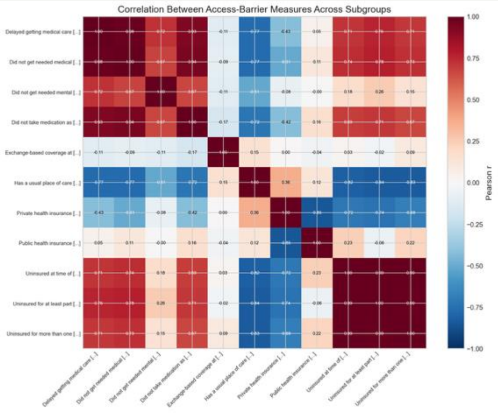


2.2 Exploratory Visualizations

- Heatmaps for composite friction patterns
- Correlation matrices to detect structural relationships
- Diagnostic plots to support modeling assumptions

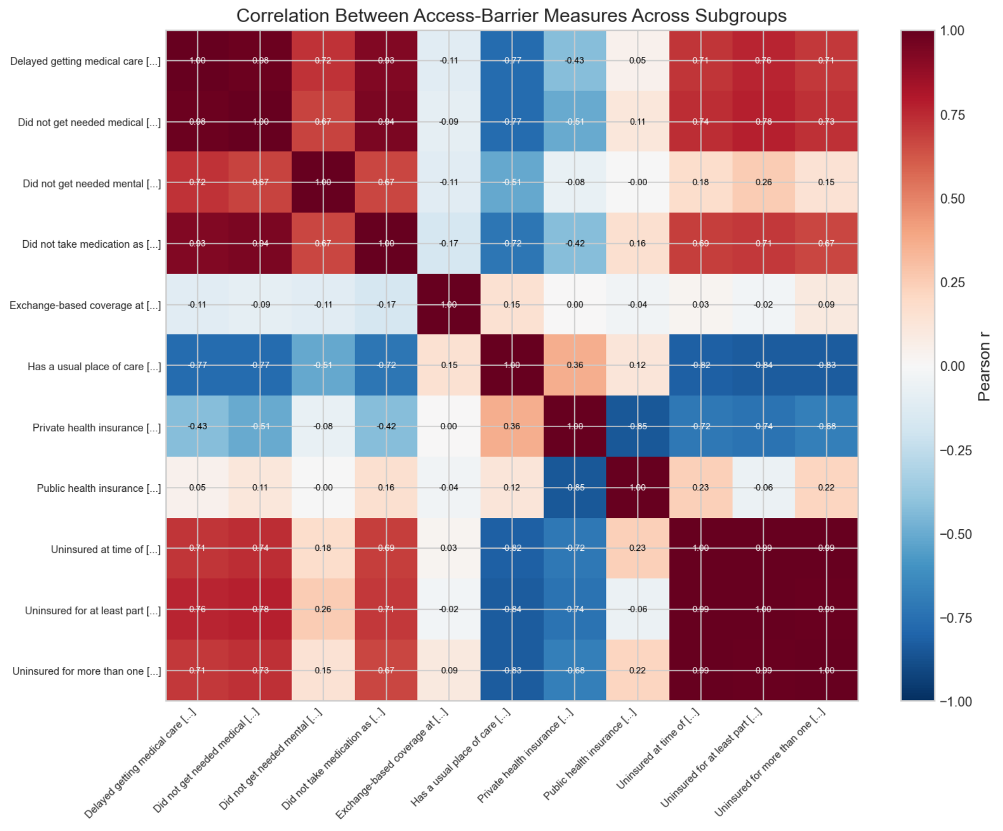






2.3 Key Exploratory Findings

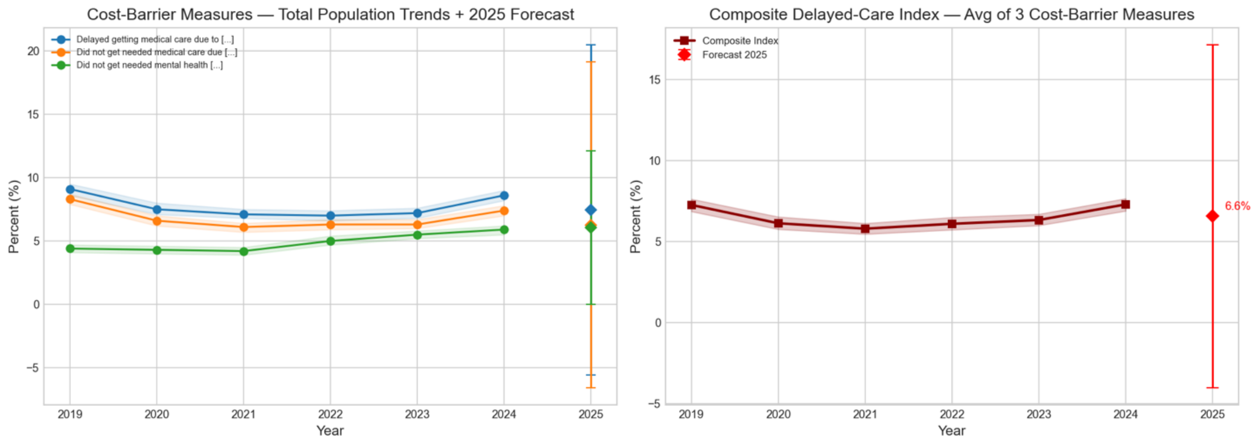
1) Cost-related access barriers are tightly correlated, with Pearson correlations exceeding 0.94 across delayed care, unmet care, and medication non-adherence.
2) These outcomes are not independent problems, but manifestations of a single affordability-driven access syndrome.



2.3 Key Exploratory Findings

Forecasting COVID-19 Distortion Impacts

- Cost-related access barriers dropped 1.3–1.7 percentage points in 2020
- Decline driven by reduced healthcare-seeking during lockdowns
- Fewer encounters led to fewer reported cost barriers
- Barriers rebounded after 2021 at ~0.4 pp per year
- 2025 projections exceed pre-pandemic levels
- COVID acts as a quantified natural experiment



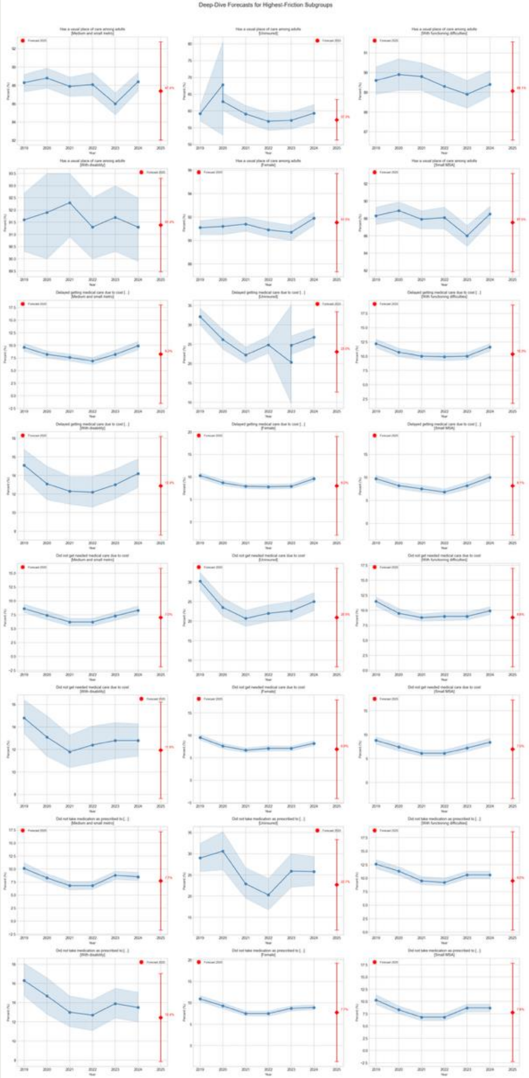
3

Description of Model (Three Core Ideas)

3.1 Forcasting Model (Prompt 1) Predict Delayed Care & Unmet Needs for Specific Subgroups

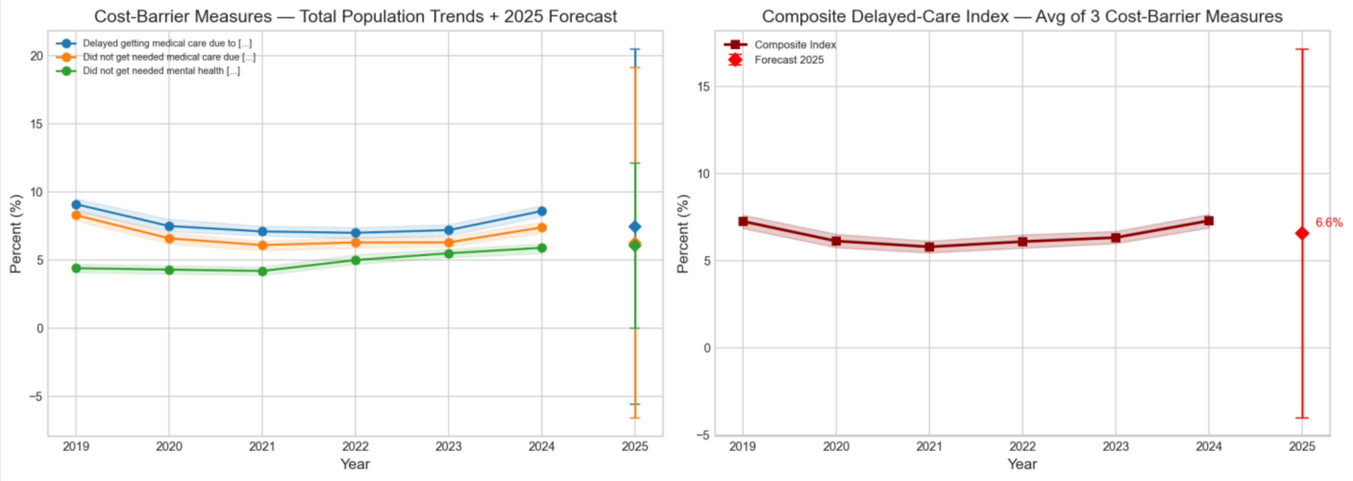
We generated subgroup-level forecasts for delayed care and unmet medical needs using uncertainty-aware models to understand how access barriers are likely to evolve for different populations.

This approach highlights not only which groups currently experience high barriers, but which are projected to worsen, enabling early identification of populations at greatest future risk rather than reacting after disparities widen.



3.2 Anomaly Detection Idea 2: Predict Dataset-Wide Trends

By aggregating access barriers across all subgroups, we identified national-level trends that reveal how healthcare affordability and access are changing over time. Composite indices reduce noise from individual measures and show that cost-related barriers move together and continue to rise post-pandemic, indicating systemic rather than isolated access challenges.

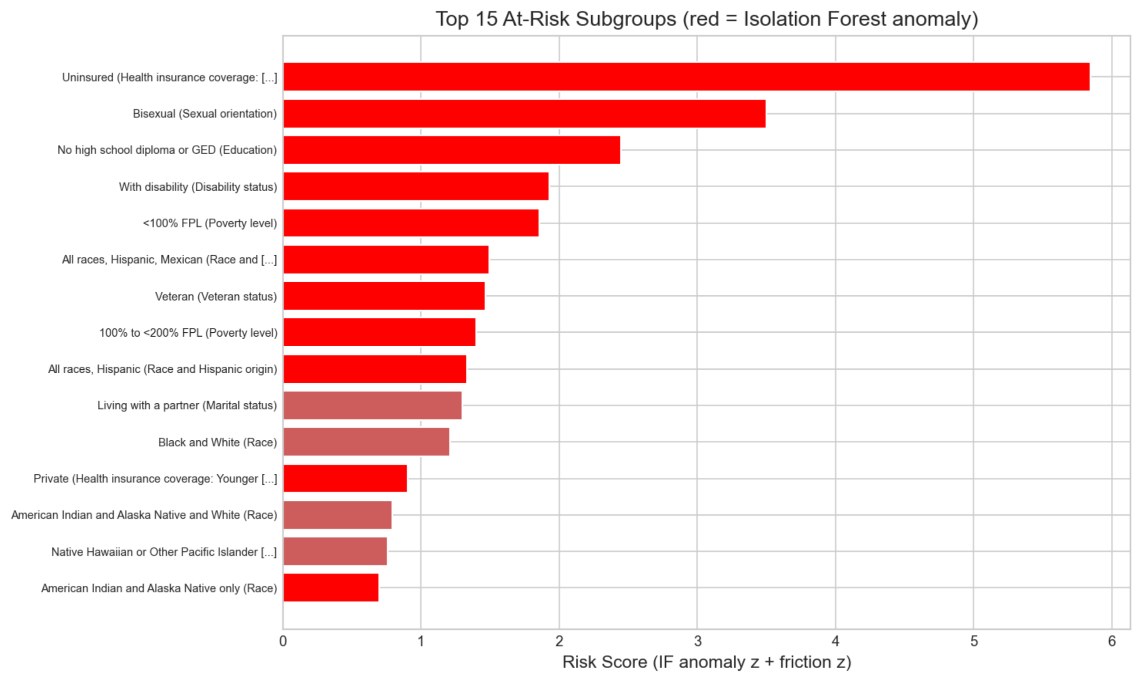


3.3 Idea 3: Use Clustering & Anomaly Detection to Identify Subgroups at Risk

Rank	Subgroup	Risk Score	Why Anomalous
1	Uninsured (under 65)	5.84	Extreme outlier on all cost barriers
2	Bisexual	3.50	Disproportionate barriers vs peers
3	No HS diploma	2.45	Compounding low literacy + low income
4	With disability	1.93	High barriers despite insurance
5	<100% FPL	1.86	Poverty concentrates all barriers

3.4 Who's Falling Through the Cracks?

- **Uninsured adults under 65** are dominant outliers
- **Low-education groups** show compounding disadvantage
- **Disabled populations** face barriers despite coverage
- **Near-poverty groups** affected by lack of subsidies
- Risk is multivariate, not single-metric
- Patterns persist across years



4

Quality & Description of Model Created

We avoided complex models due to short time series and instead prioritized interpretability and calibrated uncertainty. This is because, with only six data points, complex models overfit and provide false confidence.

4.1a Features of our Model (1)

Precision Weighing

- Predicts trends for 298 unique health metric subgroups using time as the predictor.
- Accounts for differences in survey reliability: each year's NHIS estimate has a different effective sample size, encoded in confidence intervals.
- Derives standard errors (SE) from CIs and applies precision weighting (weights = $1/SE^2$) so more reliable years influence the trend proportionally.
- Prevents distortion: without weighting, a highly precise year (SE = 0.5) and a noisy year (SE = 3.0) would contribute equally, potentially misrepresenting trends.
- Produces t-distribution-based prediction intervals for realistic uncertainty (97.3% coverage), rather than overconfident z-intervals.
- Captures trend direction and magnitude, enabling extrapolation to 2025.
- Provides reliable planning bounds even when point estimates are noisy.

4.1b Features of our Model (2)

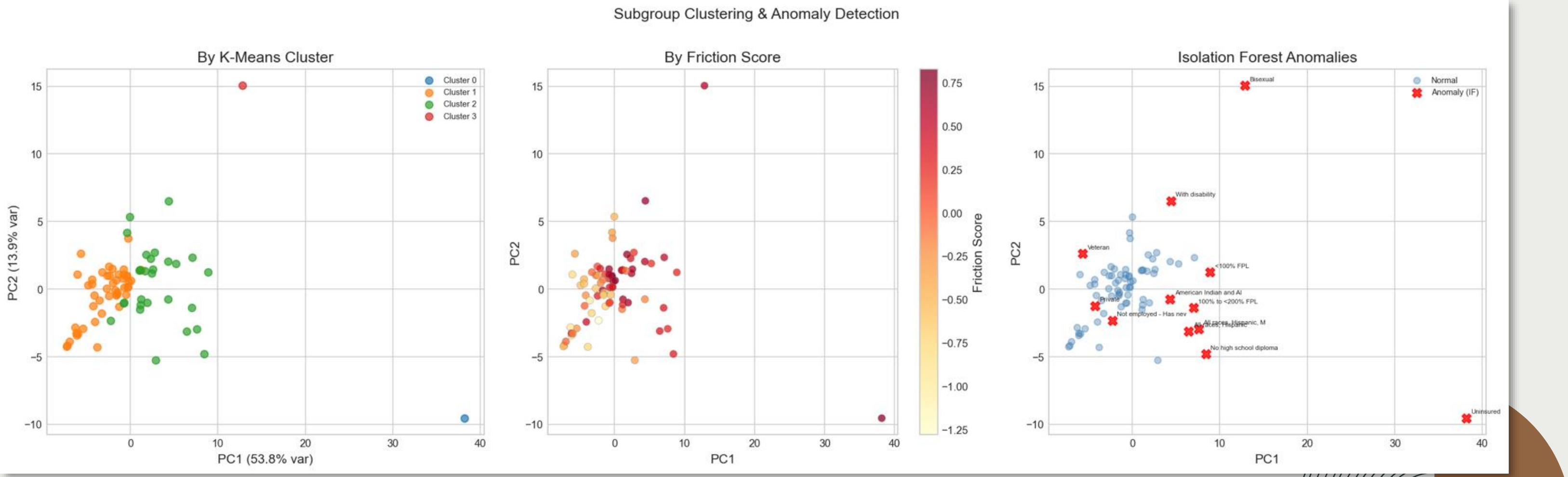
T – Distribution & Leverage Correction

Each series in our analysis has only 5–6 data points, making the standard normal approximation (z = 1.96) unreliable for estimating uncertainty. To account for this, we use t-distribution critical values based on the appropriate degrees of freedom (df = n - 2), which ensures wider and more accurate prediction intervals (e.g., t = 3.182 for df = 3; t = 2.571 for df = 5).

Additionally, we apply leverage correction to account for extrapolation, so predictions for 2025, which lie farther from the observed mean, have wider intervals to reflect increased uncertainty. This approach avoids the overconfidence of a basic z-based interval and provides more reliable forecasts for decision making.

4.2 Model Validation

Using centroid distance from K-Means clusters as an anomaly score creates circular logic. High-friction subgroups pull cluster centroids toward themselves, giving them low distances while moderate-friction subgroups between clusters get high distances. Isolation Forest operates independently of cluster assignments — it builds random isolation trees and scores each subgroup by how easily it can be separated from the rest. This is a principled anomaly detection method that avoids the double-counting problem.



Model Design Philosophy

The forecasting and anomaly detection components are designed to work together. Forecasting identifies where access barriers are trending, while anomaly detection highlights populations with globally unusual access profiles. Together, they surface both emerging risk and persistent disadvantage.

5

Value of the Model Created

Policy-Relevant Insights

Because cost barriers form a tightly coupled syndrome, interventions targeting affordability or coverage can simultaneously reduce delayed care, unmet needs, and medication non-adherence. Forecast intervals allow policymakers to prioritize action where trends are worsening with high confidence.

Limitations

- Short time series limits trend certainty
- Standard errors derived from confidence intervals
- Linear trend assumption may not hold under policy shifts
- No intersectional subgroup data
- Clusters represent continua, not sharp categories
- Results specific to NHIS Adult Summary data

Conclusion

Access barriers are structurally linked and have worsened in the post-pandemic period, particularly for cost-related measures. Machine learning adds value not by maximizing point accuracy, but by calibrating uncertainty and enabling forward-looking forecasts that support proactive intervention. The use of anomaly detection surfaces vulnerable populations that would be missed by single-metric analysis, and the resulting insights provide a data-driven foundation for more equitable healthcare policy design.

Thanks!

Mohriz Murad, Maaz Murad,
Raunak Gupta, Shayaan Ali