The Accessibility of Healthcare

Team LCSM

Background:
The "Hidden" Crisis in Healthcare Access

The fragmented landscape

- Economic Paradox: Despite the U.S. being a top-tier economy, it consistently ranks last in financial access to care among high-income nations.
- Beyond Insurance: Healthcare "cracks" are not caused by a lack of insurance alone, but by a complex intersection of barriers,

Why Traditional Analysis Falls Short

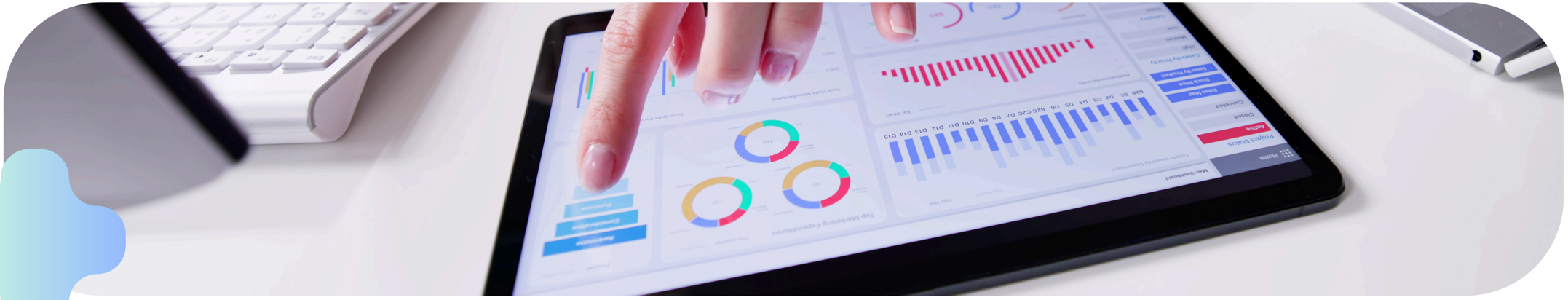
- Oversimplified Reporting: Standard demographic reporting (e.g., by age or race) often misses the multidimensional nature of vulnerability.
- Insurance vs. Access: Patients may have insurance but still face "unmet care" due to geographic or transportation barriers.
- The "Overlap" Effect: High-risk individuals often belong to multiple subgroups.

The Role of Unsupervised Learning

- Moving Beyond Categories: Identifying those "falling through the cracks" requires looking past pre-defined demographic labels.
- Clustering (K-Means): Groups individuals based on shared patterns of barriers to identify common struggle points.

Our Process

Data Understanding → Data Cleaning → Clustering → Results



Data Understanding → Data Cleaning → Clustering → Results

Data Preparation for Subgroup and Trend Analysis

Data Cleaning

- Excluded estimates flagged as unreliable to ensure statistical validity
- Removed records missing year or estimate values

Aggregation Strategy

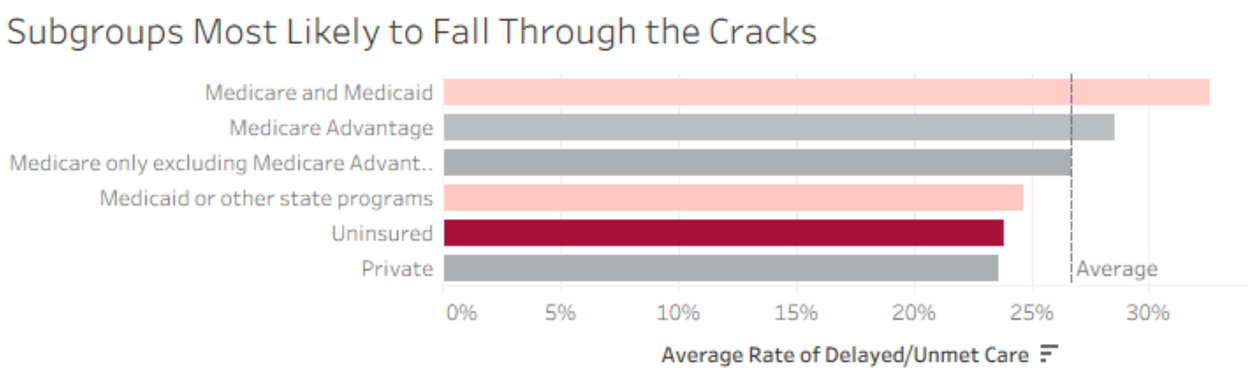
- Aggregated survey estimates by subgroup and year
- Calculated 2019–2023 averages and net change for comparison

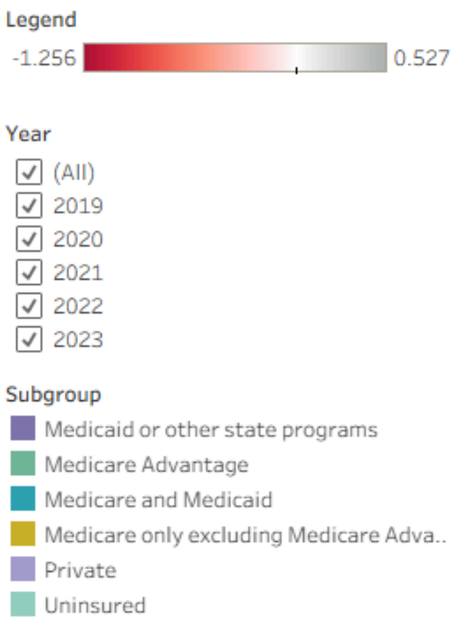
```python
# Drop rows missing essentials ('year', 'estimate')

df_clean = df.dropna(subset=['year', 'estimate'])
```

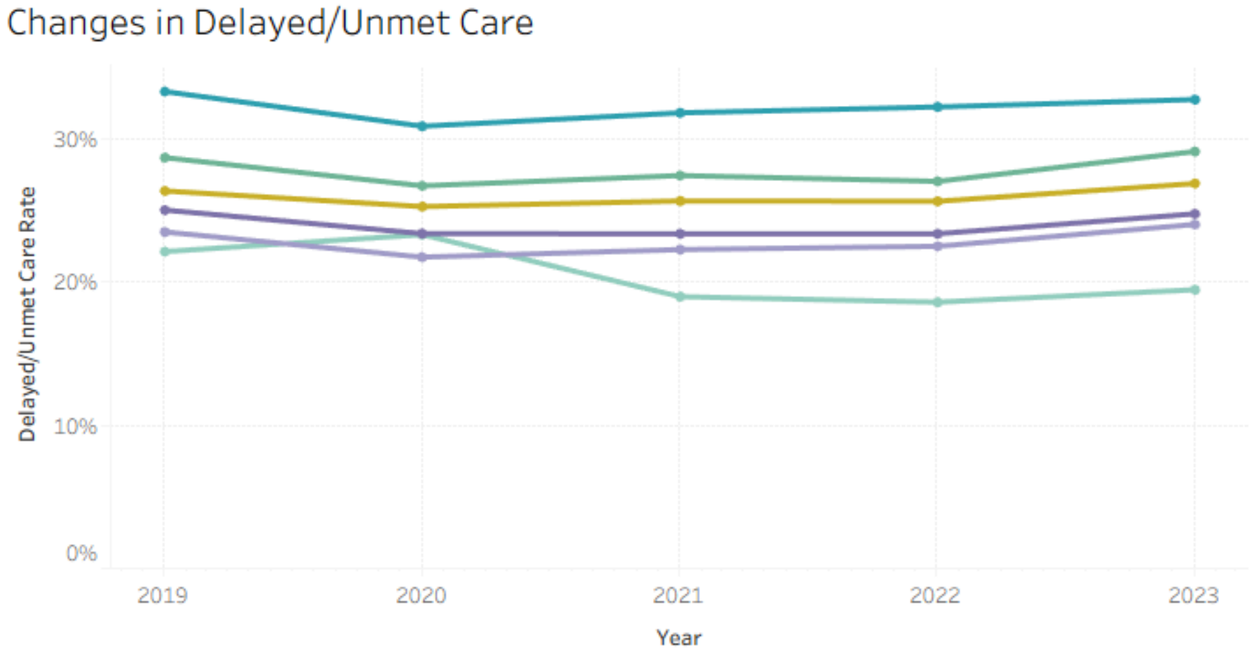
```python
# Drop rows where 'flag' is non-empty
if "flag" in df_clean.columns:
    df_clean["flag"] = df_clean["flag"].astype(str)
    df_clean = df_clean[df_clean["flag"].isin(["nan", "None", ""])]
```

Data Understanding → Data Cleaning → Clustering → Results

Who Is Falling Through the Cracks in Healthcare Access?



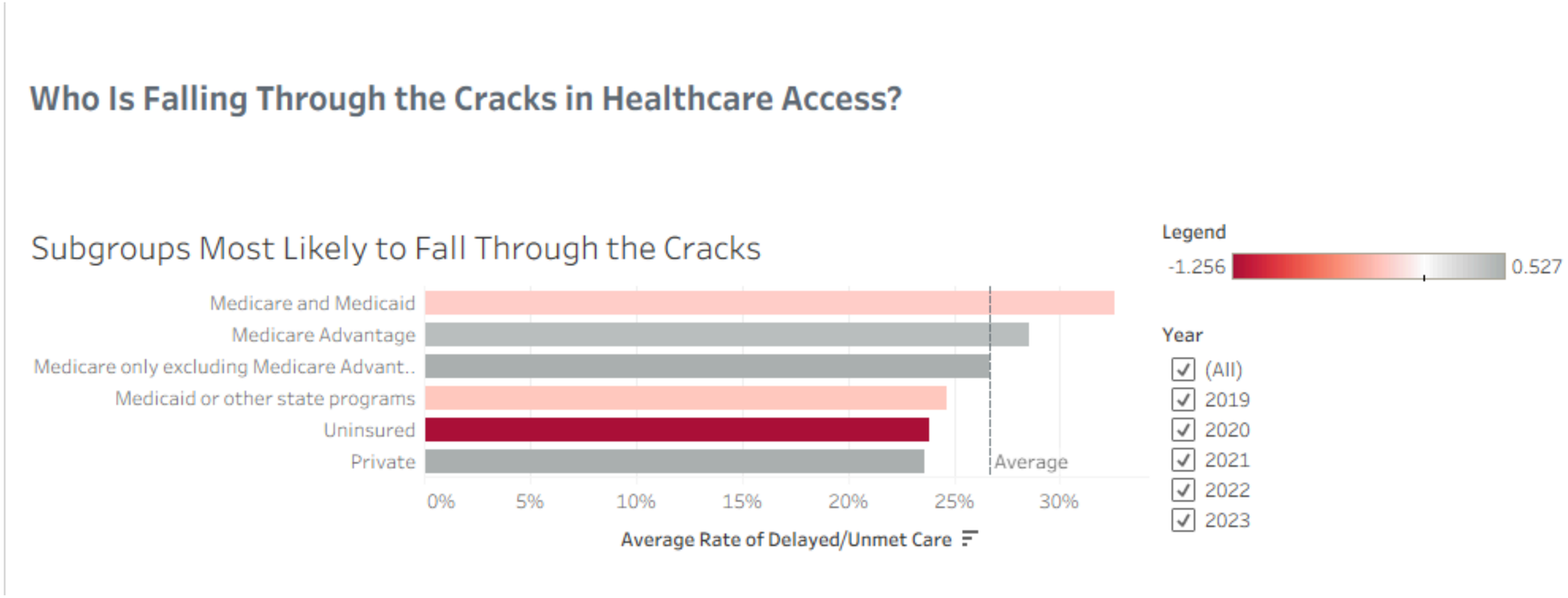




- Focused on insurance-based groups
- Goal is to capture financial and system-level barriers
- Insurance status is a key determinant of affordability, continuity, and care-seeking behavior
- Allows clear comparison across groups with different levels of coverage and support

Data Understanding → Data Cleaning → Clustering → Results

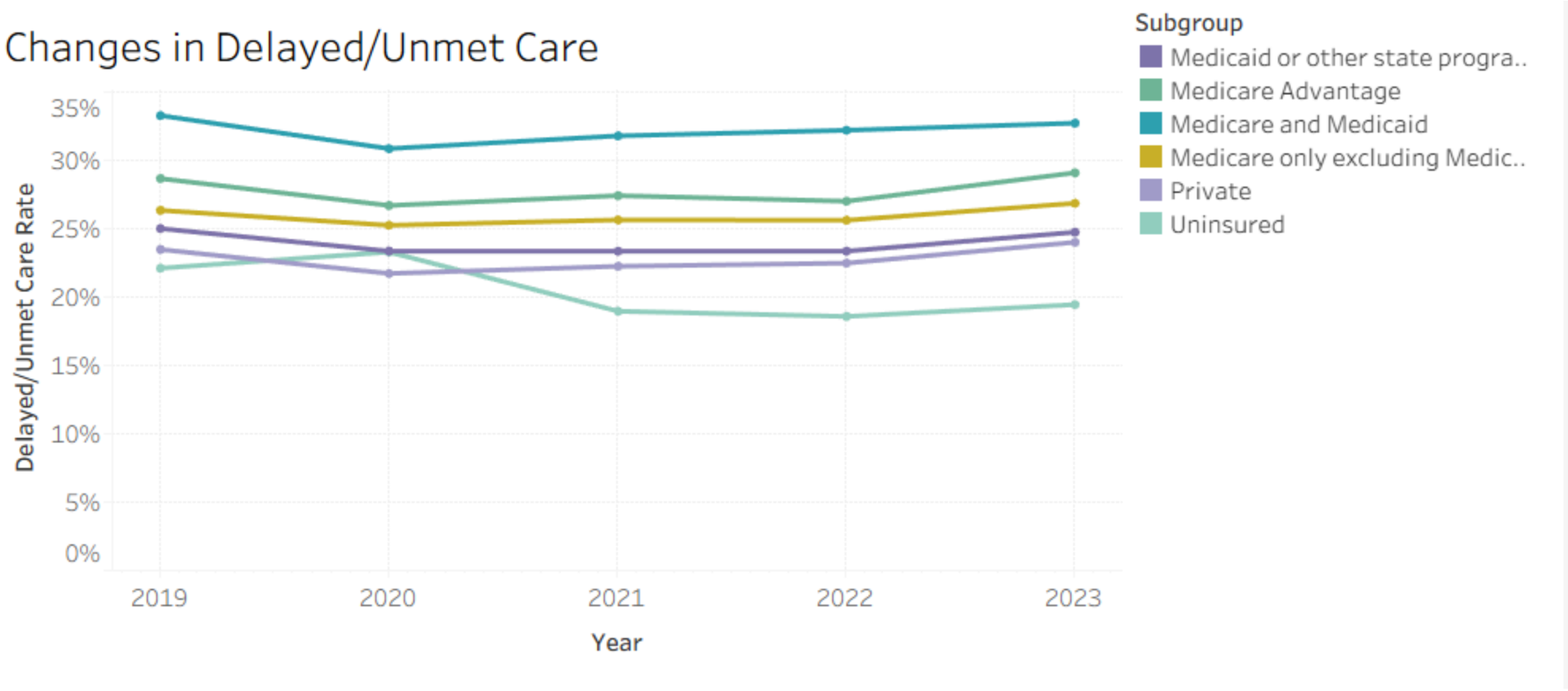
Which subgroups experience the highest rates of delayed or unmet care?



- **Publicly Insured Adults (Medicare/Medicaid)** - highest rates of delayed or unmet care
- **Uninsured Adults** - rates generally exceed the overall average
- **Private Insurance Holders** - lowest rates reported

Data Understanding → Data Cleaning → Clustering → Results

Are there trends over time (2019–2023) in delayed care or health outcomes?



- Delayed/Unmet Care peaked **around 2020** - followed by partial recoery among most subgroups
- **Publicly Insured & Uninsured Adults** - consistently higher through 2019-2023
- **Gaps between insurance subgroups persist** - even when improvement is shown, showing uneven recovery over time

After exploring the **financial barriers** in the first phase, we would like to increase the dimensions of barriers that may lead to medicare crackers. Therefore, we adopted two approaches to build the machine learning models:

- **Full-Variable Categorization and Clustering**
- **Vulnerability-Focused Metric Modeling**

While the two approachers adopted the same workflow but differ in their feature engineering strategy.

The full-variable approach incorporates a wide range of topics and categorizes all available variables prior to clustering, allowing vulnerability to emerge from the interaction of multiple factors. In contrast, the vulnerability-focused approach explicitly decomposes vulnerability into three conceptual dimensions.

Data Understanding → **Data Cleaning** → **Clustering** → Results

K-Means Modeling Workflow

Choosing the Number of Clusters → **Most Vulnerable Cluster** → **Most Vulnerable Subgroup**

- Elbow method used as a diagnostic tool (k = 2–8)
- Larger k values yield marginal gains but complicate interpretation
- k = 5 provides meaningful differentiation without sacrificing clarity

- For each cluster, calculate the centroid, representing the mean value of each feature across all observations in the cluster.
- For each feature, clusters are ranked based on their centroid values after aligning directionality (so higher values indicate greater vulnerability).
- Identify the cluster that consistently ranks worst across features as the most vulnerable.

- Focus on the most vulnerable cluster identified in the previous step
- Use the average feature value as a composite risk score
- Within each group, identify the subgroup with the highest risk score
- Find the most vulnerable subgroups overall

First Approach:
Full-Variable Categorization and Clustering

What are the determinants of limited access to healthcare?

behavioral factors
Current cigarette smoking
Current electronic cigarette or vaping product use
Obesity in adults, self-reported
cost
Delayed getting medical care due to cost among adults
Did not get needed medical care due to cost
Did not get needed mental health care due to cost
Did not take medication as prescribed to save money
disability
Any difficulty communicating
Any difficulty hearing
Any difficulty remembering or concentrating
Any difficulty seeing
Any difficulty walking or climbing steps
Any difficulty with self care
Disability status (composite) among adults
Functioning difficulties status (composite) among adults
Regularly experienced chronic pain
Six or more workdays missed due to illness, injury, or disability
insurance
Uninsured at time of interview among adults
Uninsured for at least part of the past year among adults
Uninsured for more than one year among adults
mental health
Counseled by a mental health professional
Regularly had feelings of depression
Regularly had feelings of worry, nervousness, or anxiety
Taking prescription medication for feelings of depression
Taking prescription medication for feelings of worry, nervousness, or anxiety

→ **Current smokers are less compliant with recommended preventive care and medication use than nonsmokers (Sherman & Lynch, 2014).**

→ **The United States ranks last of financial access to care among 11 high-income countries (Schoen et al., 2013).**

→ **Providers might be unable to communicate with patients because of language issues, or their offices might not be accessible to people with disabilities (MACPAC, 2016; NCHS, 2016).**

→ **Uninsured adults were:**

- **4–5 times more likely than those who had private coverage**
- **1.5–3.0 times more likely than those who had Medicaid to report difficulty in medical care and prescription access (NASEM, 2018)**

→ **Prevalence of approachability, availability, and affordability was higher among people with MHC compared to those without. (Coombs et al., 2021).**

- **Categorized factors into 5 categories:**

Mental health

Behavioral factor

Cost

Insurance

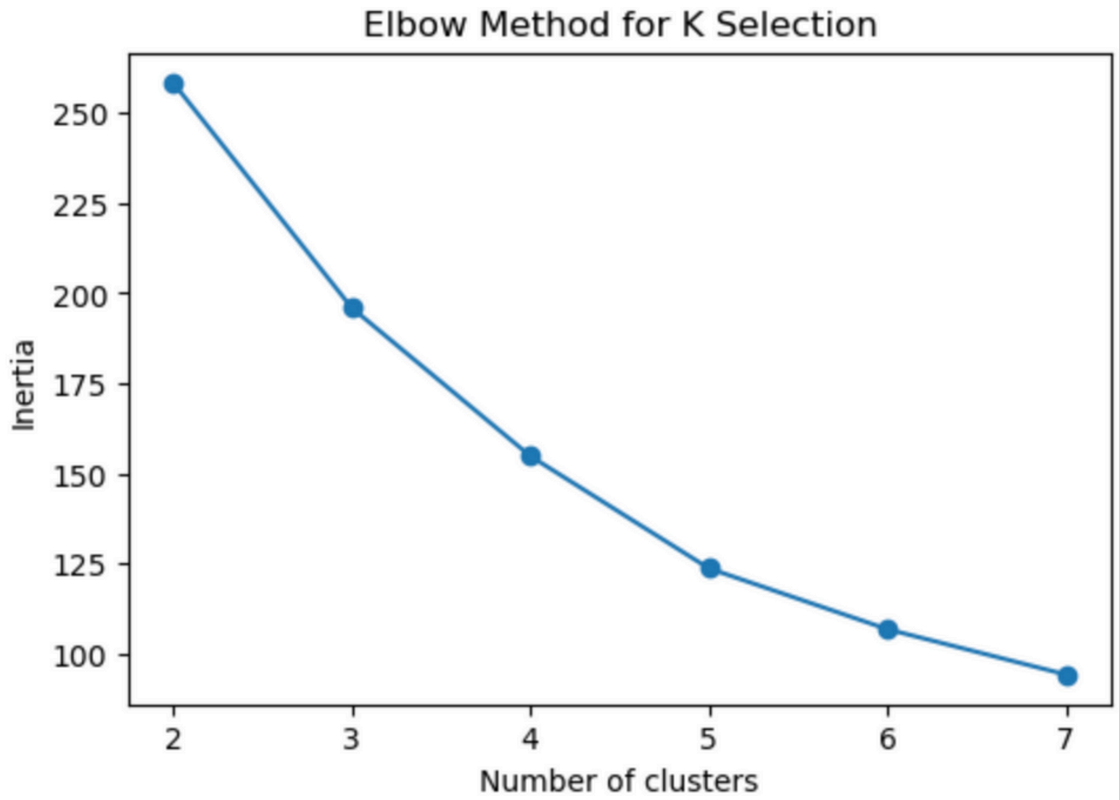
Disability

- **Group/Subgroup filters:**

  - Removed “Age groups with 75 years and older” due to repetitiveness
  - Removed “Total”
  - Droped “Estimate” that are NaN

Choosing Number of Clusters

To confirm the most appropriate k value, we conducted **elbow method testing**, the result of the testing is below:



Based on the result, we decided to cluster the target dataset into **5** groups.

First Approach

The clustering result of the first phase is below, as cluster 0 and cluster 3 get the same rank score, we would like to closely examine both clusters and see which subgroups belong to these two clusters.

TopicCategory cluster	behavioral factors	cost	disability	insurance	mental health	total_rank_score
0	1.0	1.0	4.0	3.0	2.0	11.0
3	2.0	2.0	2.0	2.0	3.0	11.0
4	3.0	5.0	1.0	5.0	1.0	15.0
1	4.0	3.0	3.0	4.0	4.0	18.0
2	5.0	4.0	5.0	1.0	5.0	20.0

Second Approach:
Vulnerability-Focused Metric Modeling

Second Approach

Defining Vulnerability: Who Falls Through the Cracks?

Barrier

Measures whether individuals are unable to obtain timely or adequate care.

- Did not get needed medical care due to cost (mean)
- Delayed getting medical care due to cost (mean)
- Did not take medication as prescribed to save money (mean)
- Any difficulty communicating with providers (mean)
- Has a usual place of care (mean, reverse-coded)

Persistence

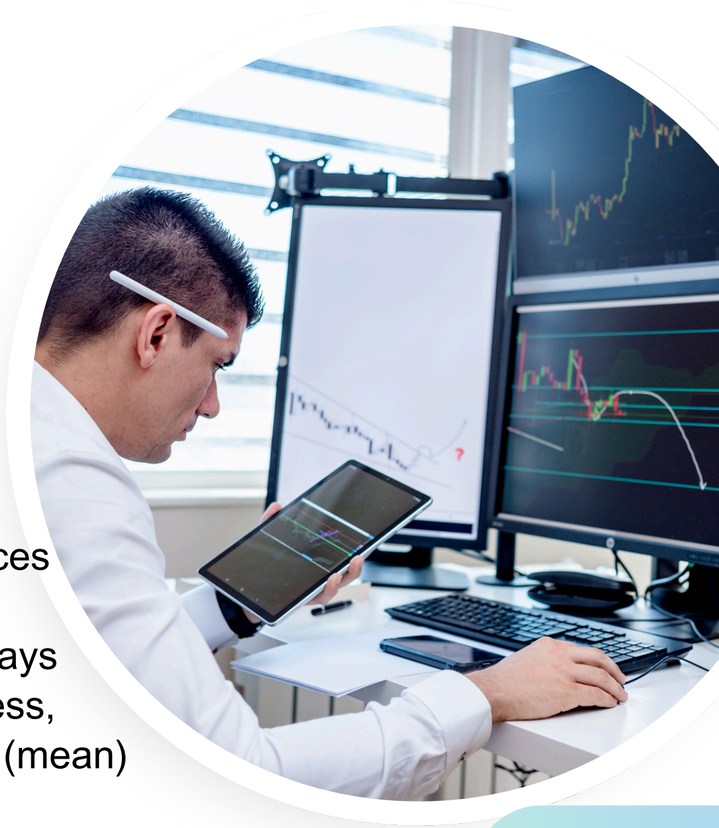
Captures ongoing or unstable health and care challenges.

- Functioning difficulties status (mean)
- Fair or poor health status (mean)
- Did not get needed medical care due to cost (standard deviation)
- Did not take medication as prescribed to save money (standard deviation)

Spillover

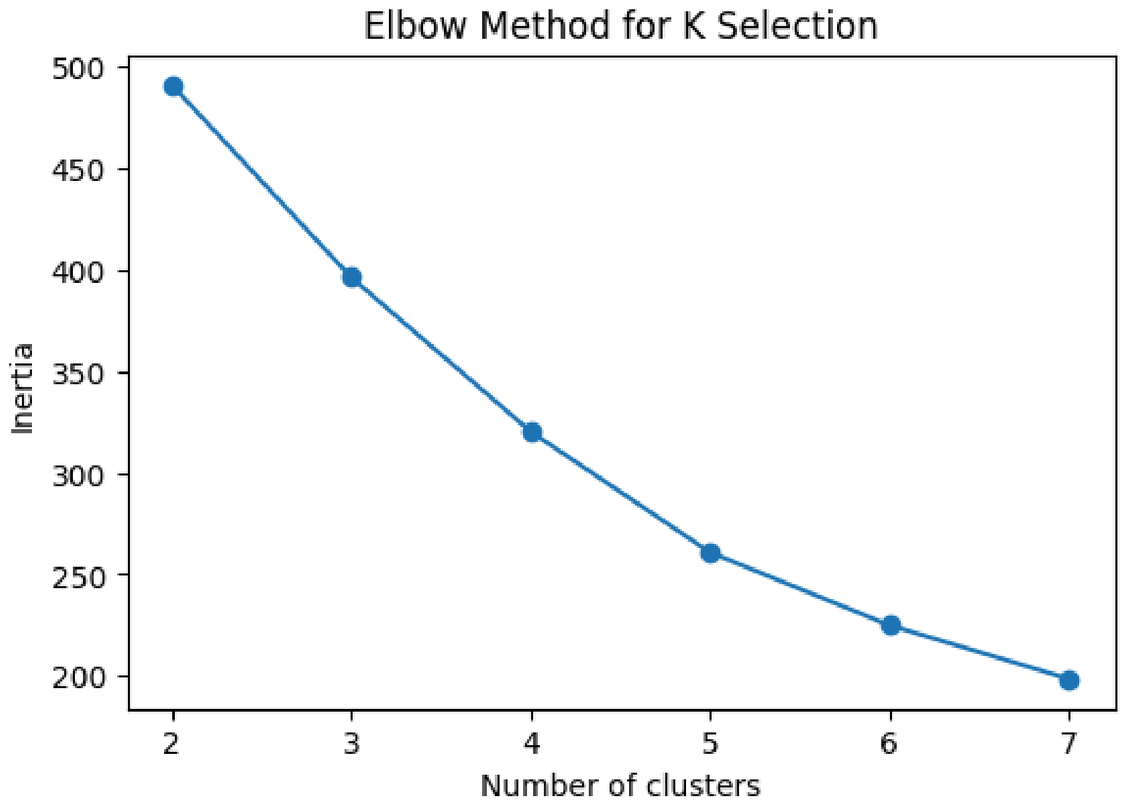
Measures consequences beyond healthcare.

- Six or more workdays missed due to illness, injury, or disability (mean)



Choosing Number of Clusters

To confirm the most appropriate k value, we conducted **elbow method testing**, the result of the testing is below:



Based on the result, we decided to cluster the target dataset into **5** groups.

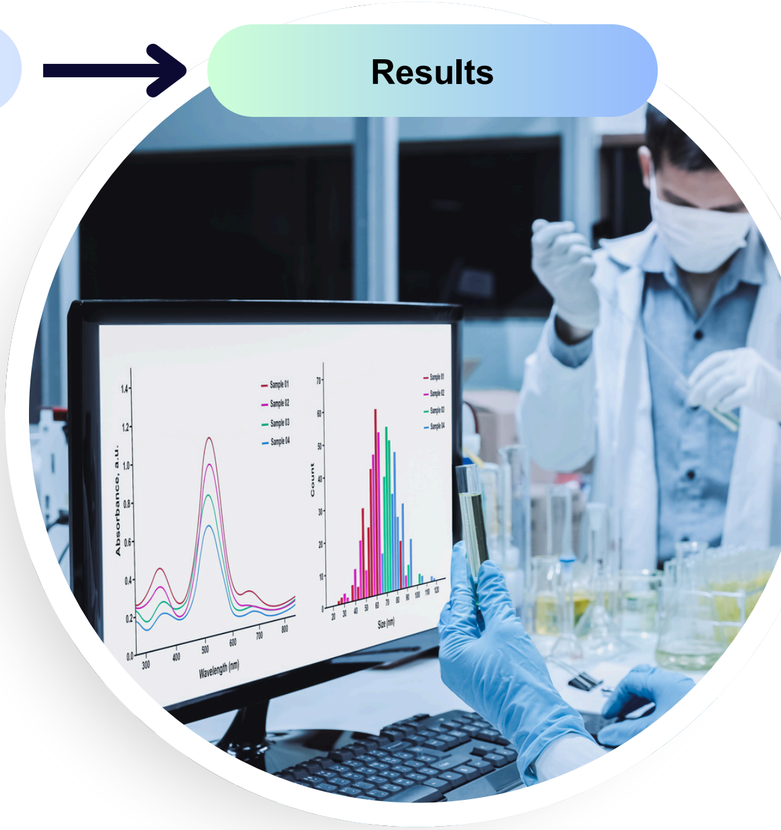
Second Approach

The clustering result of the first phase is shown below. Cluster 3 received the lowest rank score, so we identified Cluster 3 as the highest-risk group.

TOPIC cluster	Any difficulty communicating	Delayed getting medical care due to cost among adults	Did not get needed medical care due to cost	Did not take medication as prescribed to save money	Fair or poor health status in adults	Functioning difficulties status (composite) among adults	Has a usual place of care among adults	Six or more workdays missed due to illness, injury, or disability	Did not get needed medical care due to cost (std)	Did not take medication as prescribed to save money (std)	total_rank_score
3	1.0	2.0	2.0	2.0	1.0	1.0	2.0	1.0	3.0	2.0	17.0
0	4.0	1.0	1.0	1.0	4.0	4.0	1.0	2.0	1.0	1.0	20.0
1	3.0	3.0	3.0	3.0	3.0	3.0	4.0	3.0	4.0	4.0	33.0
2	2.0	5.0	5.0	5.0	2.0	2.0	5.0	4.0	5.0	5.0	40.0
4	5.0	4.0	4.0	4.0	5.0	5.0	3.0	5.0	2.0	3.0	40.0

Data Understanding → Data Cleaning → Clustering → **Results**

Result Analysis



First Approach

Based on the composite risk score in the most vulnerable cluster, the highest-risk subgroups are:

Combination 1:

1. Older adults (65+) covered by both Medicare and Medicaid
2. Previously employed but currently not employed individuals
3. Widowed individuals

Combination 2:

1. Hispanic, Mexican
2. People with a high school diploma or GED
3. People living in the South

→ **The two methods surface different types of vulnerability!**

Second Approach

Based on the composite risk score in the most vulnerable cluster, the highest-risk subgroups are:

1. Individuals with a high school diploma or GED
2. Residents of nonmetropolitan areas
3. Individuals living outside Metropolitan Statistical Areas (Non-MSA)

Why the Results Differ

Full-variable clustering captures emergent vulnerability

Unlike the vulnerability-metric model, full-variable clustering does not restrict attention to pre-selected risk indicators. Instead, it allows all variables to jointly shape cluster structure.
As a result, this approach surfaces **compound risk profiles** that reflect individuals' current living conditions, where vulnerability emerges from the interaction of **health status, insurance arrangements, and social circumstances**.

Vulnerability-metric modeling identifies upstream, structural barriers

By anchoring the analysis in well-established vulnerability dimensions, this approach prioritizes structural conditions that shape healthcare access before individuals interact with the healthcare system.
As a result, vulnerability-metric modeling highlights **systematically underserved** populations whose elevated risk arises from upstream barriers.