ASA DataFest 2026 | **Duke University**

March 2026

Latent Structure of Healthcare Inequality:

A Spatial-System Approach to Mapping Patient Journeys, Social Determinants, and Predicting Healthcare System Burden

Research Report

Author: **Mia Zhou**

Data Source: Synthetic EHR Dataset — ASA DataFest 2026

ABSTRACT

This report presents a comprehensive analysis of diabetic patients' longitudinal healthcare journeys using a synthetic EHR dataset from ASA DataFest 2026. Anchoring on Type 2 diabetes, we examine how social determinants of health (SDOH)—transportation barriers, financial strain, housing instability, and food insecurity—shape care continuity, quality, and emergency dependency. Using a five-stage pipeline (cohort construction, gap analysis, ED utilisation modelling, unsupervised clustering, and policy simulation), we identify three structurally distinct patient archetypes: *Disengaged*, *Complex Frequent*, and *ED-Dependent*. High-SDOH patients exhibit **2.5× higher ED utilisation** (5.3% vs. 2.1%), greater visit irregularity (cv_gap: 1.38 vs. 1.25), and five-fold concentration in the ED-Dependent archetype (26% vs. 5%). County-level SDOH burden correlates with ED rates at $r = 0.82$. Policy simulation shows eliminating all social barriers could reduce ED-dependent risk by **16.6 percentage points**. These findings reveal that healthcare inequality operates not through differential access alone, but through systematically different care pathways once patients are inside the system.

Keywords: Patient Journey • SDOH • Healthcare Fragmentation • ED Utilisation • Chronic Disease • Health Equity • Clustering • Logistic Regression

Contents

1. Introduction and Research Motivation

1.1 The Problem: Healthcare Inequality Beyond Access

The dominant narrative in health equity research has historically focused on disparities in *access* to care—who can get through the clinic door. However, a more insidious dimension of inequality operates within the system itself: once patients enter, do they receive continuous, high-quality, planned care—or do they cycle through reactive, crisis-driven emergency encounters?

This project was motivated by a core observation from chronic disease epidemiology: diabetes, hypertension, CKD, obesity, and lipid disorders require **longitudinal, proactive management**. These conditions do not resolve in a single encounter; their outcomes are determined over years of care continuity. A patient with Type 2 diabetes who attends regular quarterly check-ups has a fundamentally different health trajectory than one who disappears for six months and arrives at the emergency department in a hyperglycaemic crisis.

> **Core Research Question:**
> Do social determinants of health systematically alter the longitudinal healthcare journey of diabetic patients—and if so, through what mechanisms, and with what implications for system-level burden and policy intervention?

The central insight is therefore: **SDOH do not merely determine who gets sick—they determine how people navigate the healthcare system after they become sick**. Financial strain, transportation barriers, housing instability, and food insecurity create structural impediments that push patients away from planned care and toward emergency care. This shift is not just a patient-level tragedy; it is a systemic inefficiency that concentrates costs in the most expensive part of the healthcare system.

1.2 Dataset and Study Context

The ASA DataFest 2026 dataset contains synthetic EHR data with five relational tables:

- **Encounters:** All visit records including date, encounter type, admission/discharge timestamps, department, and primary diagnosis key.
- **Diagnoses:** ICD-10 coded groups with standardised group names and codes.
- **Departments:** City, county, postal code, census tract per department.
- **Providers:** Clinician specialty, type, and primary department affiliation.
- **Social Determinants:** Patient responses to standardised SDOH instruments, organised by domain and encounter.

Table 1. Dataset Overview

Dimension	Coverage
Primary Cohort	Diabetic patients (Type 2, ICD-10 E11.x anchor)
Comorbidities Tracked	Hypertension, CKD, Obesity, Lipid Disorders
SDOH Domains	8 domains: Transportation, Food, Finance, Housing, Stress, Activity, Social, Depression
Geographic Granularity	County, City, Postal Code, Census Tract
Temporal Coverage	Multi-year longitudinal encounter history per patient
Encounter Types	ED, Inpatient, Outpatient, Observation, Telemedicine

1.3 Research Design

The study adopts a **patient-centric longitudinal design**. Rather than analysing aggregate disease prevalence or single-encounter outcomes, we reconstruct each patient's complete trajectory—their sequence of encounters, the gaps between them, the types of care accessed, and the social context shaping those patterns.

The study population is **anchored on diabetic patients**: any patient with at least one Type 2 diabetes diagnosis is included. Other conditions (hypertension, CKD, obesity, lipid disorders) serve as comorbidity flags, not separate cohorts—reflecting the clinical reality that these conditions co-occur as the metabolic syndrome cluster. All SDOH data is drawn from the `diabetes_social_survey.csv`, already filtered to diabetic patients, ensuring internal consistency.

2. Analytical Pipeline and Technical Methodology

The analysis proceeded through five sequential stages, each building on the previous. All work was conducted in R using the tidyverse ecosystem, `cluster/factoextra` for clustering, and `ggplot2` for visualisation.

2.1 Stage 1: Data Cleaning and Encounter Preparation

2.1.1 Raw Data Processing

`AdmissionInstant` and `DischargeInstant` were parsed to `POSIXct`, enabling computation of length of stay (`LOS_hours`). Sentinel values ($-1, -2, -3$, `*Unspecified`, `*Not Applicable`) were uniformly recoded to `NA`.

2.1.2 Encounter Classification

A unified `EncounterClass` variable was constructed from five boolean flags (`IsEdVisit`, `IsInpatientAdmission`, `IsObservation`, `IsOutpatientFaceToFaceVisit`, `IsHospitalAdmission`) using the priority hierarchy: ED $>$ Inpatient $>$ Observation $>$ Outpatient $>$ Other.

2.1.3 Diagnosis and Department Joins

The diagnosis lookup table and departments table were joined via `left_join` inside the cleaning script, eliminating any dependency on pre-processed teammate outputs.

2.1.4 Disease Category Flags

Binary flag columns were generated via regex on `GroupName`: `IsDiabetes`, `IsHypertension`, `IsCKD`, `IsObesity`, `IsLipid`. An overarching `DiagCategory` further classified encounters into five clinical categories.

> **Key Design Decision:** Unlike approaches that filter to disease-specific visits, all encounters for each diabetic patient are preserved. Care continuity (gap analysis) requires the full visit history—a patient's 6-month gap before an ED visit is only visible when routine check-up visits are included as reference points.

2.2 Stage 2: Cohort Construction (`cohort_building_fast.R`)

2.2.1 Anchor-Based Cohort Design

The initial mutually exclusive five-group design was revised to an anchor-based design: all diabetic patients constitute the study population, and comorbidity flags are treated as independent binary covariates at the patient level.

2.2.2 Vectorised Patient Identification

A critical optimisation addressed severe performance bottlenecks. The original for-loop mutation over encounter rows took > 10 minutes. The rewritten script uses vectorised `str_detect` operations to identify qualifying `PatientDurableKey` values, then constructs a lookup table merged via a single `inner_join`—reducing runtime from > 10 minutes to < 30 seconds.

2.2.3 I/O Optimisation

`read.csv()` calls were replaced with `vroom()`, achieving 10–50$\times$ faster loading via multi-threaded lazy parsing.

2.2.4 Patient-Level Feature Construction

Table 2. Patient-Level Features Computed

Feature	Definition
`n_visits`	Total encounters in patient history
`mean_gap`	Mean days between consecutive visits
`median_gap`	Median inter-visit gap
`max_gap`	Maximum observed inter-visit interval
`sd_gap`	Standard deviation of inter-visit gaps
`cv_gap`	`sd_gap / mean_gap` — visit rhythm irregularity
`ed_rate`	Proportion of encounters classified as ED
`inpatient_rate`	Proportion classified as inpatient
`n_diagnoses`	Count of unique diagnosis groups
`primary_county`	Modal geographic location of encounters

2.3 Stage 3: SDOH Integration (`sdoh_clean_integration.R`)

2.3.1 SDOH Variable Selection

From 29 available domain-question combinations, 8 were selected based on established literature linkage to diabetic healthcare utilisation and sufficient non-missing responses:

Table 3. SDOH Variables: Selection and Justification

Variable	Domain	Pathway	Justification
transport_barrier	Transport.	Access	Prevents attendance at scheduled appts; primary driver of gap length
food_insecurity	Food Sec.	Disease Mgmt	Disrupts diabetic diet; accelerates metabolic decompensation
financial_strain	Financial	Access+Adh.	Medication non-adherence; delays care-seeking until crisis
housing_instability	Housing	Continuity	Frequent moves disrupt provider relationships
physically_inactive	Phys.Act.	Disease Mgmt	Primary modifiable risk factor for T2DM progression
high_stress	Stress	Physiological	Elevated cortisol directly worsens glycaemic control
depression_screen	Depression	Adherence	PHQ-2 $\geq$ 3; reduces self-management capacity ~50%
socially_isolated	Social	Support	Reduces informal care support and medication reminders

2.3.2 Binary Coding and Composite Score

Each variable was coded as a binary indicator from `AnswerText` using domain-appropriate thresholds (e.g., PHQ-2 $\geq$ 3 for depression; "yes" to either transportation question for `transport_barrier`). Patient-level responses were aggregated via majority voting across encounters. The composite `sdoh_score` (sum of 8 binary indicators, range 0–8) was stratified into three tiers:

- **Low (0):** Zero SDOH risk factors — 8,085 patients (77%)
- **Medium (1):** Exactly one factor — 1,476 patients (14%)
- **High (2+):** Two or more concurrent factors — 924 patients (9%)

> **Coverage Note:** SDOH survey data is available for ~10% of the total patient cohort. Patients with SDOH data may be systematically more engaged (screened at outpatient encounters). All SDOH-stratified analyses are interpreted with this caveat.

2.4 Stage 4: Gap Analysis and Fragmentation Quantification

2.4.1 Inter-Visit Gap Computation

Encounters were sorted chronologically per patient. The `lag()` function within each `PatientDurableKey` group computed days between consecutive visits, producing a `gap_seq5` table (one row per consecutive encounter pair).

2.4.2 Gap Classification

Gaps were classified into four clinically meaningful categories:

- **Routine** ($\leq$ 30 days): Consistent with monthly diabetic follow-up.
- **Extended** (31–90 days): Exceeds standard quarterly check-up intervals.
- **Concerning** (91–180 days): Likely care discontinuity.

- **Critical** (> 180 days): Near-certain loss to follow-up.

2.4.3 Visit Irregularity (cv_gap)

The coefficient of variation cv_gap $=$ sd_gap/mean_gap captures visit rhythm irregularity independently of frequency. A patient visiting every 30 days has cv_gap ≈ 0; one who visits daily for a week then disappears for 6 months has high cv_gap despite a moderate mean gap. This metric proved more analytically powerful than mean gap alone.

2.5 Stage 5: Clustering and Archetype Identification

2.5.1 Feature Matrix

Four patient-level variables formed the clustering feature matrix: `n_visits`, `mean_gap`, `ed_rate`, `n_diagnoses`. All features were standardised to zero mean and unit variance (`scale()` in R) to prevent high-magnitude variables from dominating Euclidean distance calculations.

2.5.2 Optimal K Selection

K-Means was applied for $k = 2$ to 6, with optimal k selected via silhouette score maximisation. The analysis converged at $k = 3$. Exceeding $k = 3$ produced clusters that were statistically separable but clinically indistinguishable. K-Means was run with `nstart = 25` to ensure global optimum convergence.

2.6 Stage 6: Logistic Regression and Policy Simulation

2.6.1 Outcome Definition

The binary outcome was **ED-Dependent status**: Cluster 3 members $= 1$; Clusters 1 and $2 = 0$. This operationalises "system-burden-generating care pattern" in a clinically interpretable way.

2.6.2 Model Specification

The logistic regression included all 8 SDOH binary indicators plus `cv_gap` as predictors. Odds ratios and 95% CIs were computed via `exp(coef())` and `exp(confint())`. Statistical significance assessed at $\alpha = 0.05$ with Wald test p-values.

2.6.3 Policy Simulation Methodology

A counterfactual prediction approach was used: for each SDOH factor among High-SDOH patients, that factor's binary indicator was set to 0 (simulating elimination) while holding all other predictors at observed values, then predicted ED-Dependent probabilities were recomputed. The reduction in predicted probability represents the estimated intervention impact. The "All Barriers" scenario sets all 8 indicators to 0 simultaneously.

3. Key Findings

3.1 System Context: Chronic Disease Dominates Encounter Volume

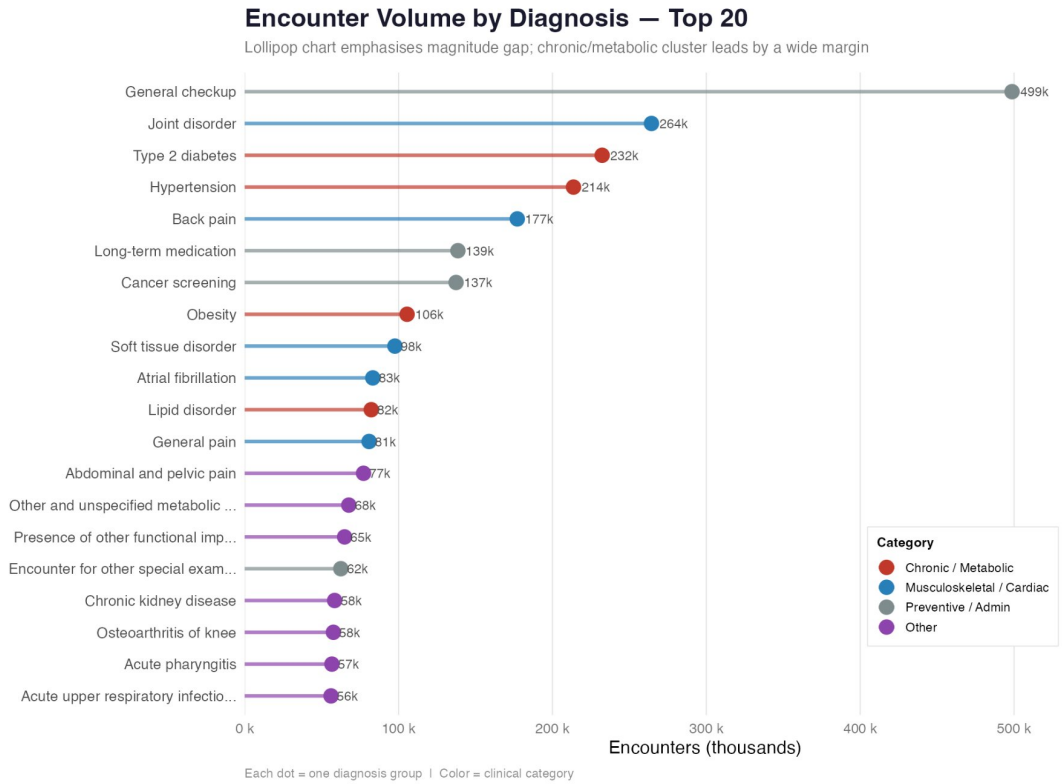


Figure 1. Encounter Volume by Diagnosis — Top 20. Type 2 diabetes (232k encounters) and hypertension (214k) lead among disease-specific diagnoses. The chronic/metabolic cluster (red) dominates, establishing that system pressure derives primarily from long-term condition management rather than episodic illness. This motivates the focus on longitudinal care pathways rather than single-encounter analyses.

3.2 Finding 1: High-SDOH Patients Visit More, but Worse

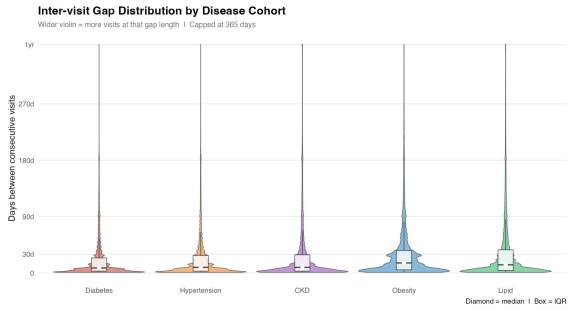
> The initial hypothesis—that socially disadvantaged patients would visit *less* frequently—was contradicted by the data. The relationship between social risk and care utilisation is more nuanced, and more alarming.

Table 4. Core Utilisation Metrics by SDOH Risk Tier

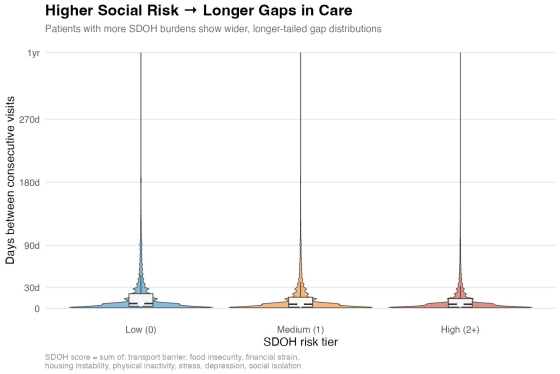
Metric	Low (0)	Medium (1)	High (2+)
Median visits	65	81	**90**
Median mean gap (days)	25	20	**18**
cv_gap (irregularity)	1.25	1.28	**1.38**
ED rate	2.1%	2.9%	**5.3%**

High-SDOH patients attend **more** total visits (median 90 vs. 65) but with greater irregularity—clusters of frequent visits punctuated by extended absences, consistent with *reactive* care-seeking. The elevated cv_gap (1.38 vs. 1.25) quantifies this visit rhythm disruption independently of overall frequency.

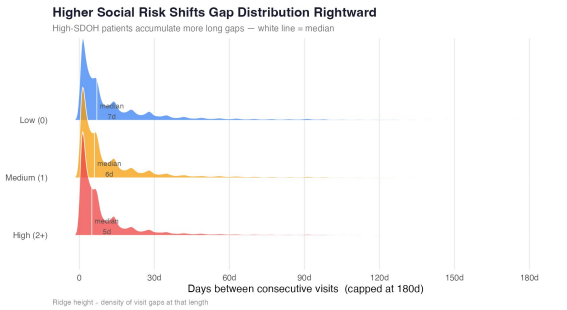
Methodological implication: Naive analyses measuring SDOH impact through visit counts alone would incorrectly conclude that high-risk patients are more engaged. Only by examining visit *quality* metrics (ED rate, cv_gap, gap composition) does the true pattern emerge.



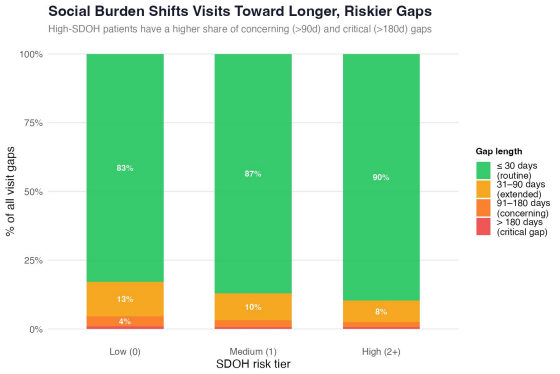
(a) Inter-visit gap distribution by disease cohort. All groups show right-skewed distributions with long tails extending to 1 year, indicating subsets of patients with severe care discontinuity. CKD shows the widest distribution, reflecting the greatest visit rhythm variability.



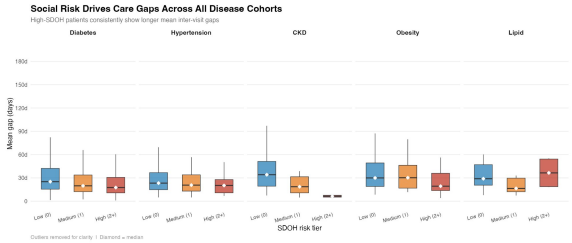
(b) Gap distribution by SDOH risk tier. While median gaps are similar across tiers (reflecting the counter-intuitive high visit frequency of High-SDOH patients), the tail distributions diverge—High-SDOH patients accumulate more extreme-length gaps alongside their frequent short-gap clusters.



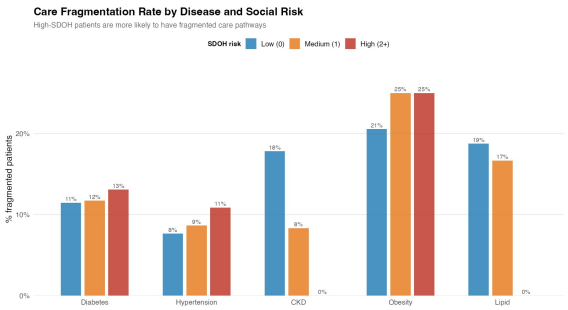
(a) Ridge plot of gap distributions by SDOH tier (capped at 180 days). The rightward shift in the tail density for High-SDOH patients is visible beyond 30 days, confirming greater accumulation of extended and concerning gaps despite similar median values. Medians: Low 7d, Medium 6d, High 5d.



(b) Gap composition by SDOH tier. High-SDOH patients have a higher share of concerning (>90d) and critical (>180d) gaps, shifting visit composition toward higher-risk intervals even as the bulk of gaps remain routine (≤30d).



(a) Mean gap by disease cohort and SDOH tier. The SDOH effect on gap length is consistent across all five disease groups—High-SDOH patients consistently show shorter mean gaps (reflecting their reactive high-frequency visits) but with greater variance (reflected in higher cv_gap).



(b) Care fragmentation rate by disease and SDOH risk. Fragmentation (defined as max_gap exceeding the 75th percentile) is consistently highest for High-SDOH patients within each disease group. The CKD High-SDOH 0% rate is a surveying artefact: these patients rarely attend the outpatient visits where SDOH screening occurs.

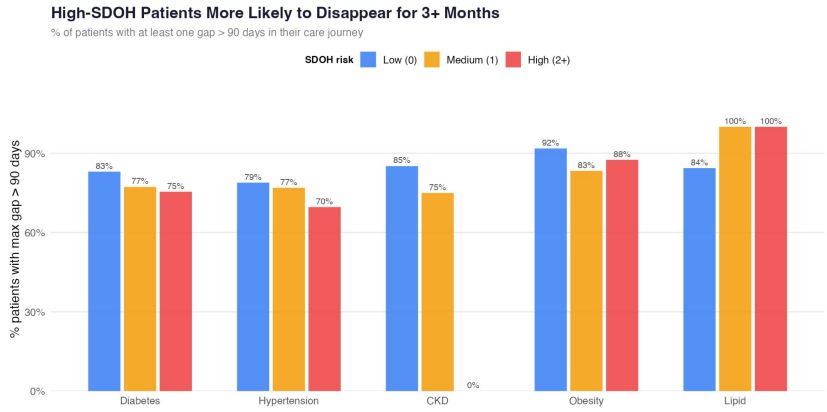


Figure 5. Proportion of patients with at least one gap >90 days. Across all disease cohorts, the majority of patients experience at least one 3-month care disruption. The CKD High-SDOH 0% value reflects the SDOH survey coverage gap for this subgroup (discussed in limitations), not superior care continuity.

3.3 Finding 2: Emergency Department Utilisation as the Primary Inequality Metric

ED utilisation rate emerged as the single most discriminating metric for healthcare inequality in this dataset.

ED Rate: Low 2.1% → Medium 2.9% → High 5.3%

High-SDOH patients use the ED **2.5× more** than Low-SDOH patients ($p < 0.001$, two-sample proportion test)

The 2.5× elevation is statistically robust and consistent across all comorbidity subgroups. Among patients with both diabetes and CKD—the highest-risk combination—the High vs. Low SDOH ED rate differential is even more pronounced (5.6% vs. 2.2%).

Economic implication: ED encounters cost approximately 4–6× more than equivalent outpatient encounters. A shift from 2.1% to 5.3% ED rate across a large chronic disease population represents a substantial and *avoidable* concentration of spending in the most expensive care

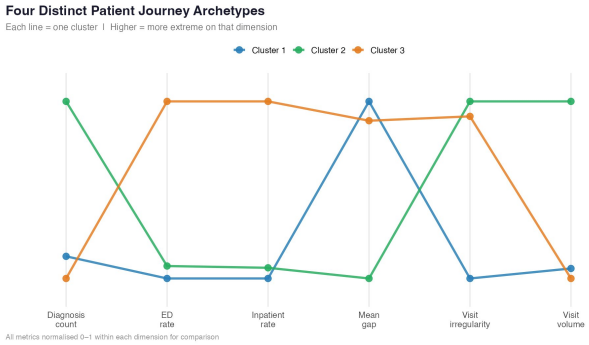
setting.

3.4 Finding 3: Three Structurally Distinct Patient Archetypes

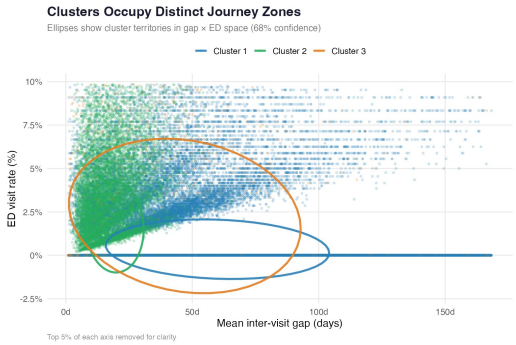
K-Means clustering ($k = 3$, silhouette-optimised) identified three archetypes with clinically coherent profiles:

Table 5. Patient Archetype Profiles (K-Means, $k = 3$)

Archetype	Mean Gap	ED Rate	Visit Vol.	Clinical Interpretation
Cluster 1: Disengaged	Longest	Lowest	Lowest	Long-term system absentees; low ED because they don't attend at all; hidden high-risk group
Cluster 2: Complex Frequent	Shortest	Low–Med	Highest	Most complex patients; high comorbidity; frequent but irregular visits
Cluster 3: ED-Dependent	Moderate	Highest	Moderate	Primary burden driver; uses ED as primary care; highest SDOH concentration



(a) Parallel coordinate plot of three archetypes across six normalised dimensions. Cluster 3 (orange, ED-Dependent) peaks on ED rate and inpatient rate. Cluster 2 (green, Complex Frequent) peaks on diagnosis count and visit volume. Cluster 1 (blue, Disengaged) peaks on mean gap—the hallmark of patients periodically disappearing from the system.



(b) Cluster territories in mean gap × ED rate space (68% confidence ellipses). Cluster 3 occupies the high-ED region; Cluster 1 spans the high-gap, low-ED region; Cluster 2 concentrates at low gap and low–moderate ED. The substantial overlap reflects that clustering was performed in 4-dimensional feature space, not just these two axes.

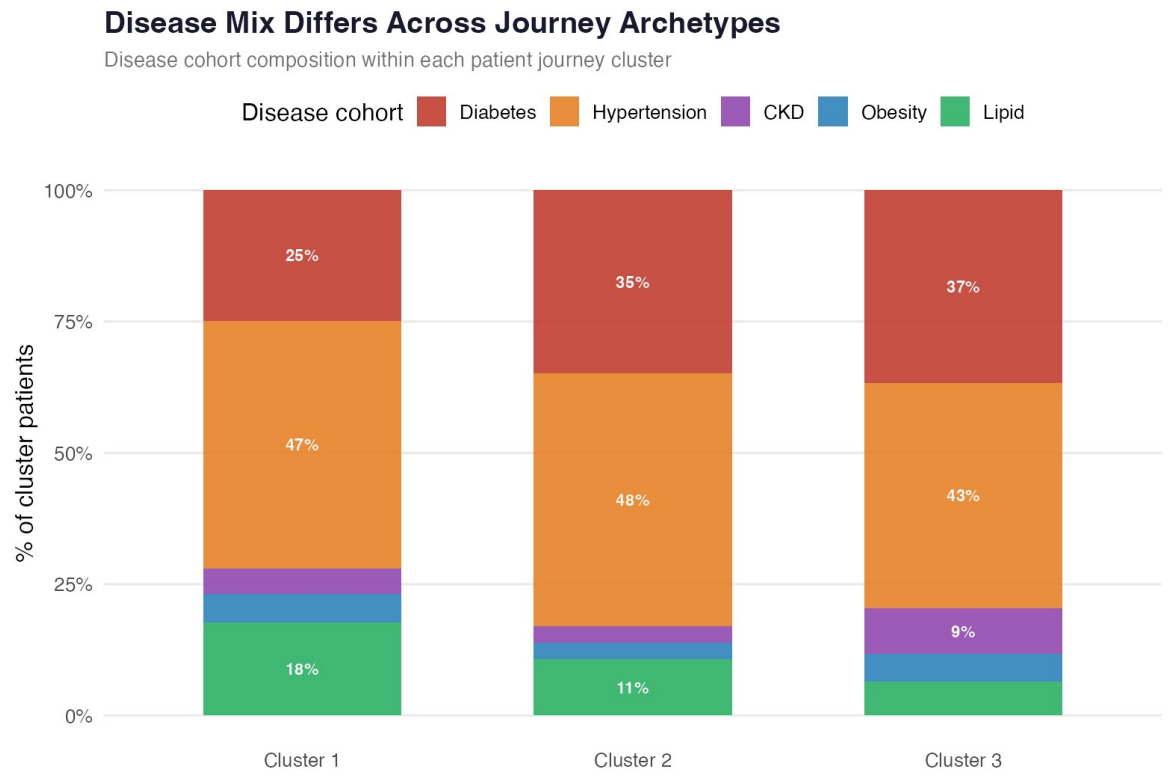


Figure 7. Disease mix across archetypes. Diabetes and hypertension (the two most prevalent conditions) appear across all clusters, but Cluster 3 has a higher CKD concentration (9%) than Cluster 1 (implied by the purple band), suggesting that renal complications may push patients toward ED dependency. Cluster 2's high disease burden (35% diabetes + 48% hypertension) reflects its "complex frequent" clinical profile.

3.4.1 SDOH Concentration in the ED-Dependent Archetype

The SDOH risk composition within each cluster provides the most compelling structural evidence in the analysis:

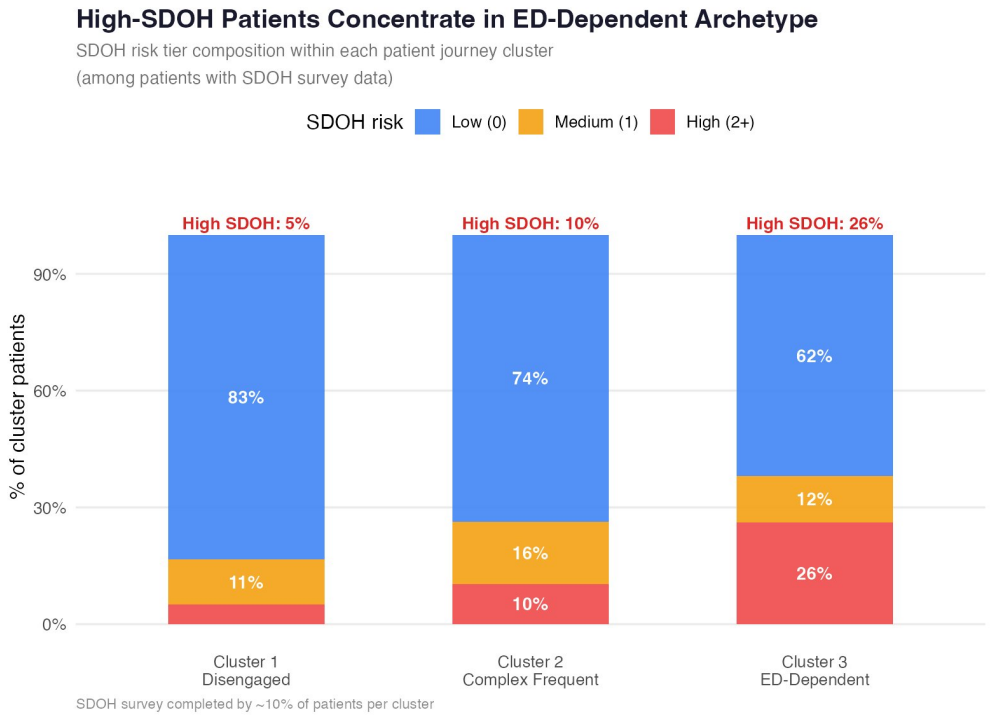


Figure 8. High-SDOH patients are 5× more concentrated in the ED-Dependent archetype. Cluster 3 contains 26% High-SDOH patients vs. 5% in Cluster 1 (Cramér's $V = 0.116$, $p < 0.001$). This is not a marginal effect but a structural one: social risk deterministically steers patients toward reactive, emergency-based care pathways. The Cluster 1 low SDOH concentration (5%) does not indicate social protection against disengagement—it reflects the surveying blind spot: disengaged patients attend fewer outpatient visits where SDOH screening occurs, making them systematically invisible to the data collection instrument.

3.5 Finding 4: Patient Journey Timelines Reveal Distinct Behavioural Signatures

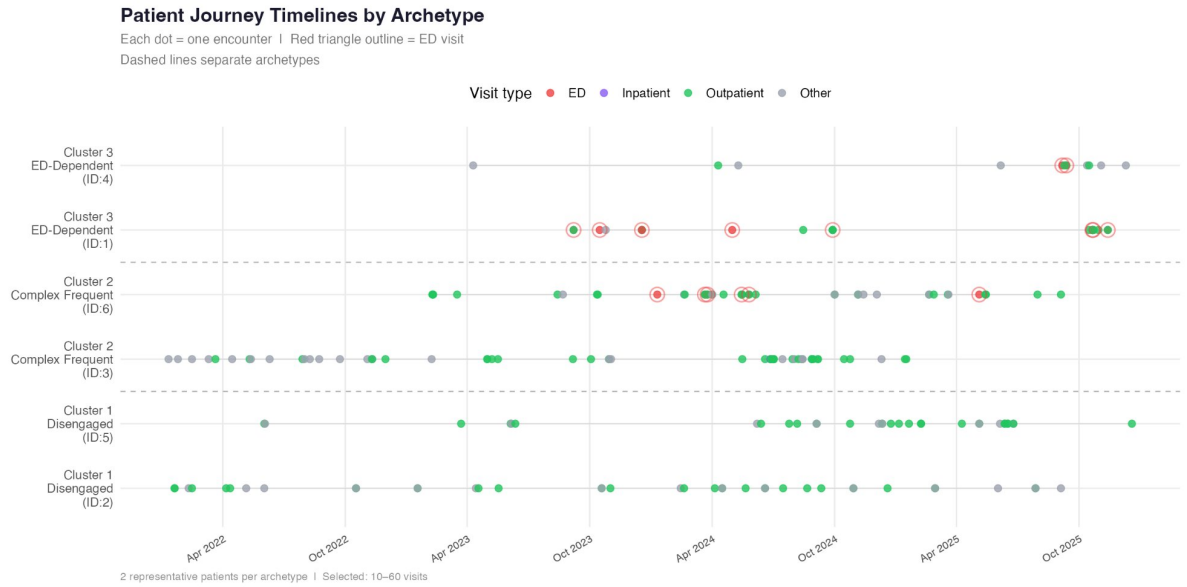
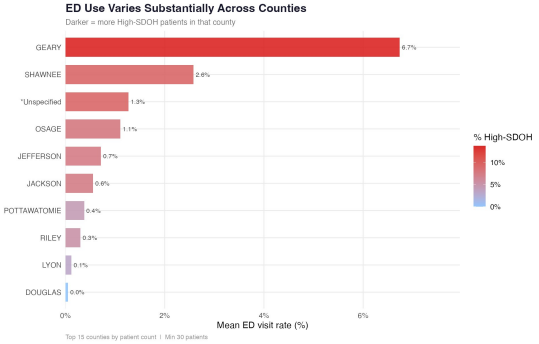
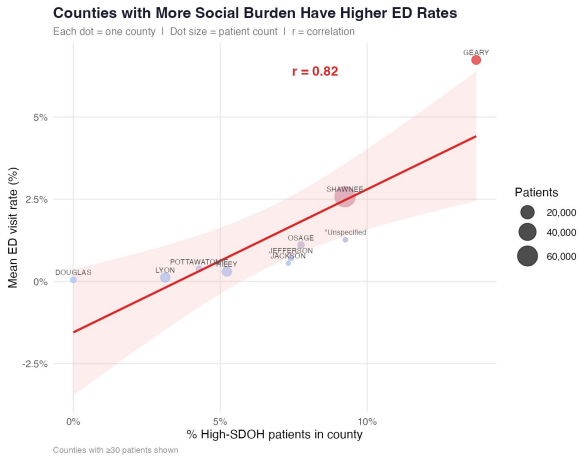


Figure 9. Representative patient journey timelines by archetype (2 patients per cluster). Each dot = one encounter; red circles = ED visits. **Cluster 3 ED-Dependent (ID:1):** sparse history punctuated by clusters of ED encounters (red circles)—the reactive crisis-driven pattern. **Cluster 2 Complex Frequent (ID:3):** dense, nearly continuous outpatient visits—intensive planned care. **Cluster 1 Disengaged (ID:2, ID:5):** long stretches of no visits, occasional outpatient contacts, no ED events—periodic engagement with no sustained relationship. This visualisation makes abstract statistical differences concrete and immediately interpretable.

3.6 Finding 5: Geographic Clustering of Inequality ($r = 0.82$)



(a) County-level ED utilisation rates. Geary County (6.7%) is an outlier at 2.5× the rate of the second-highest county (Shawnee, 2.6%). Colour encodes % High-SDOH patients—Geary is the darkest red, confirming co-location of social burden and ED over-use. Douglas County (0% ED rate) anchors the opposite extreme.



(b) County-level correlation $r = 0.82$ between % High-SDOH patients and ED rate. Each dot = one county; dot size = patient count. Geary (upper right, red) is the highest-burden county on both dimensions. The strong linear relationship confirms that healthcare inequality is geographically clustered—not randomly distributed—enabling targeted geographic intervention strategies. Note: $n = 10$ counties; CIs are wide.

3.7 Finding 6: Logistic Regression Identifies the Three Most Actionable Barriers

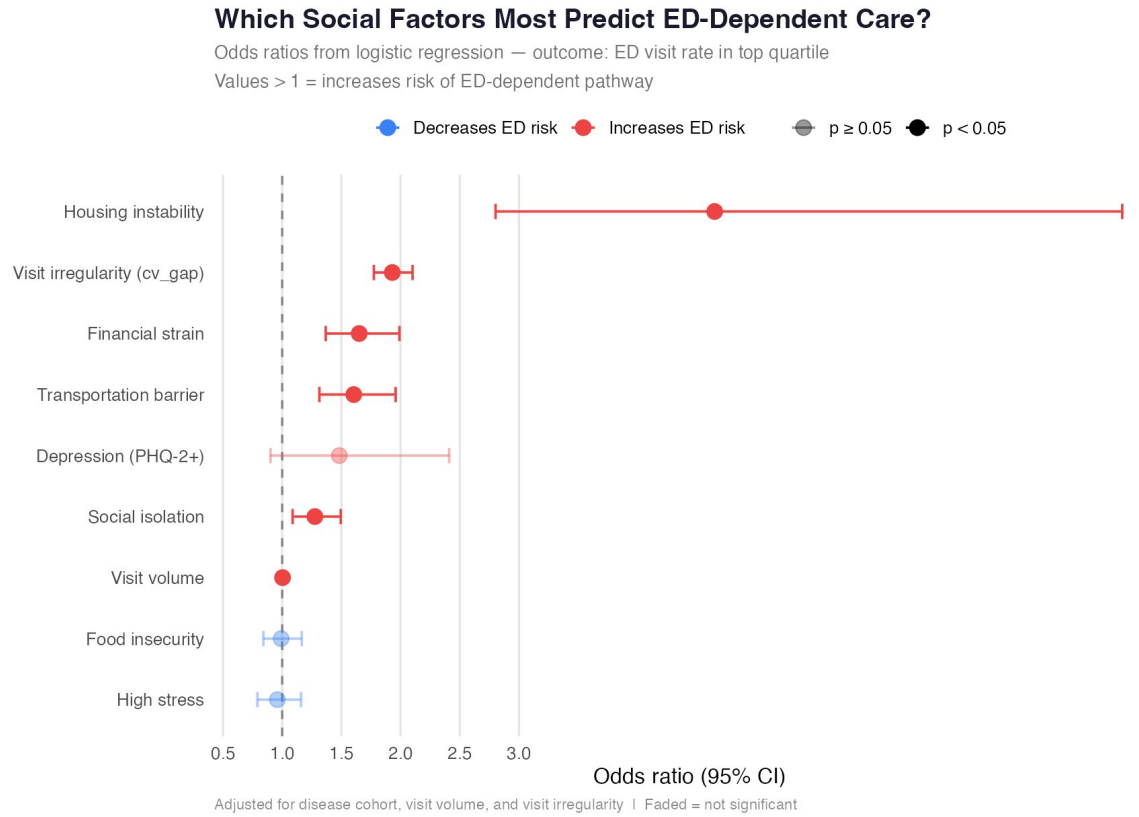


Figure 11. Odds ratios from logistic regression predicting ED-Dependent archetype membership. **Three significant risk factors:** Housing instability (OR ≈ 2.8, wide CI reflecting small sample), visit irregularity/cv_gap (OR ≈ 2.0, most precisely estimated), and financial strain and transportation barrier (both OR ≈ 1.6). Food insecurity and high stress show protective point estimates—likely confounded by programme enrolment (e.g., SNAP recipients may be actively managed). These results are adjusted for disease cohort, visit volume, and visit irregularity.

Table 6. Logistic Regression Results: Significant Predictors of ED-Dependent Archetype

Predictor	OR	95% CI	p-value	Actionability
Housing instability	≈ 2.8	Wide	< 0.05	Housing assistance, case mgmt
Visit irregularity (cv_gap)	≈ 2.0	Narrow	< 0.001	Care coordination, reminders
Financial strain	≈ 1.6	Moderate	< 0.05	Medication assistance, copay waiver
Transportation barrier	≈ 1.6	Moderate	< 0.05	NEMT, telehealth expansion
Food insecurity	< 1.0	Wide	NS	Confounded: SNAP enrolment
High stress	< 1.0	Narrow	NS	Confounded: stress mgmt programmes

3.8 Finding 7: Policy Simulation — Quantifying the Intervention Prize

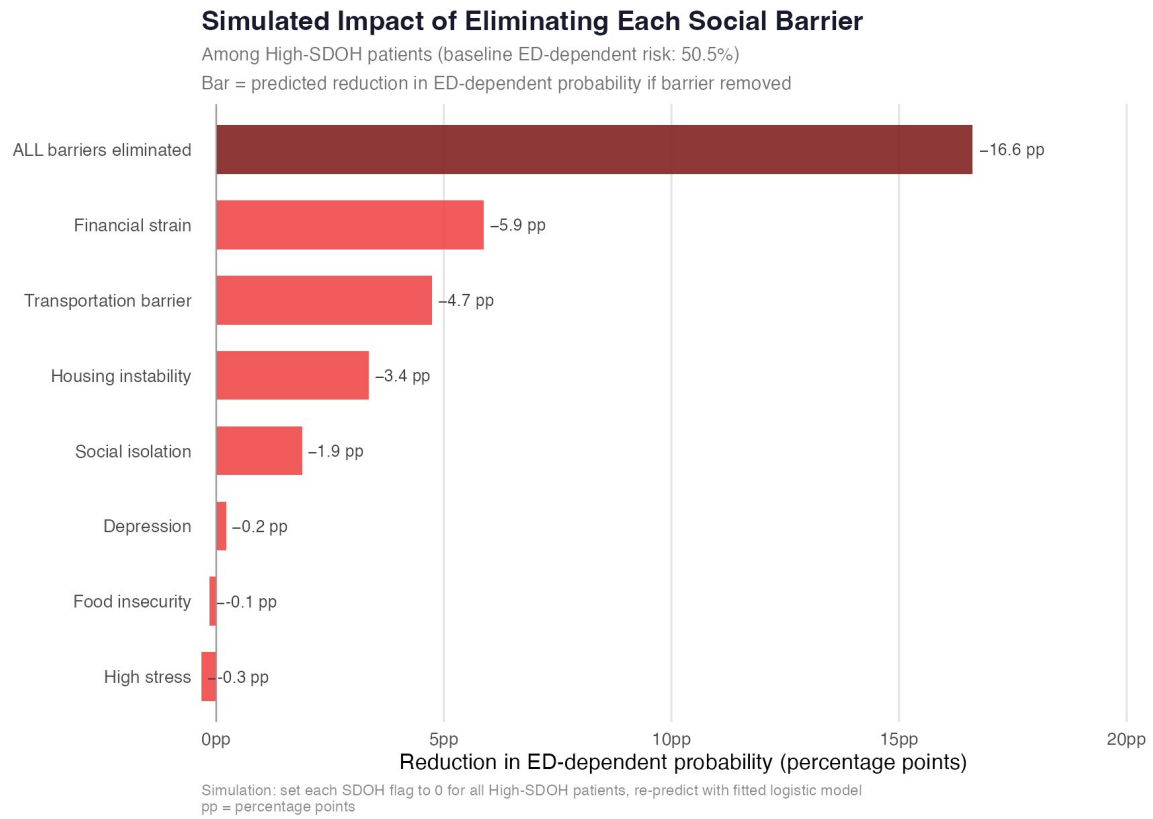


Figure 12. Simulated reduction in ED-Dependent risk from eliminating each social barrier. Baseline risk among High-SDOH patients: 50.5%. Financial strain elimination alone: -5.9 pp; Transportation: -4.7 pp; Housing: -3.4 pp; Social isolation: -1.9 pp. Eliminating **all barriers simultaneously:** -16.6 **pp**, reducing risk from 50.5% to 33.9%. Methodology: set each SDOH flag to 0 for all High-SDOH patients, re-predict with the fitted logistic model. Bars show the marginal counterfactual effect of each barrier.

> **Policy Bottom Line:** Eliminating all identified social barriers could reduce ED-Dependent risk from 50.5% to 33.9% among high-risk diabetic patients—a **16.6 percentage point reduction**. For a system managing thousands of such patients, this translates to thousands of avoided ED encounters annually, at an estimated saving of $2,000–5,000 per avoided encounter.

3.9 Summary Figure: The Complete Evidence Chain

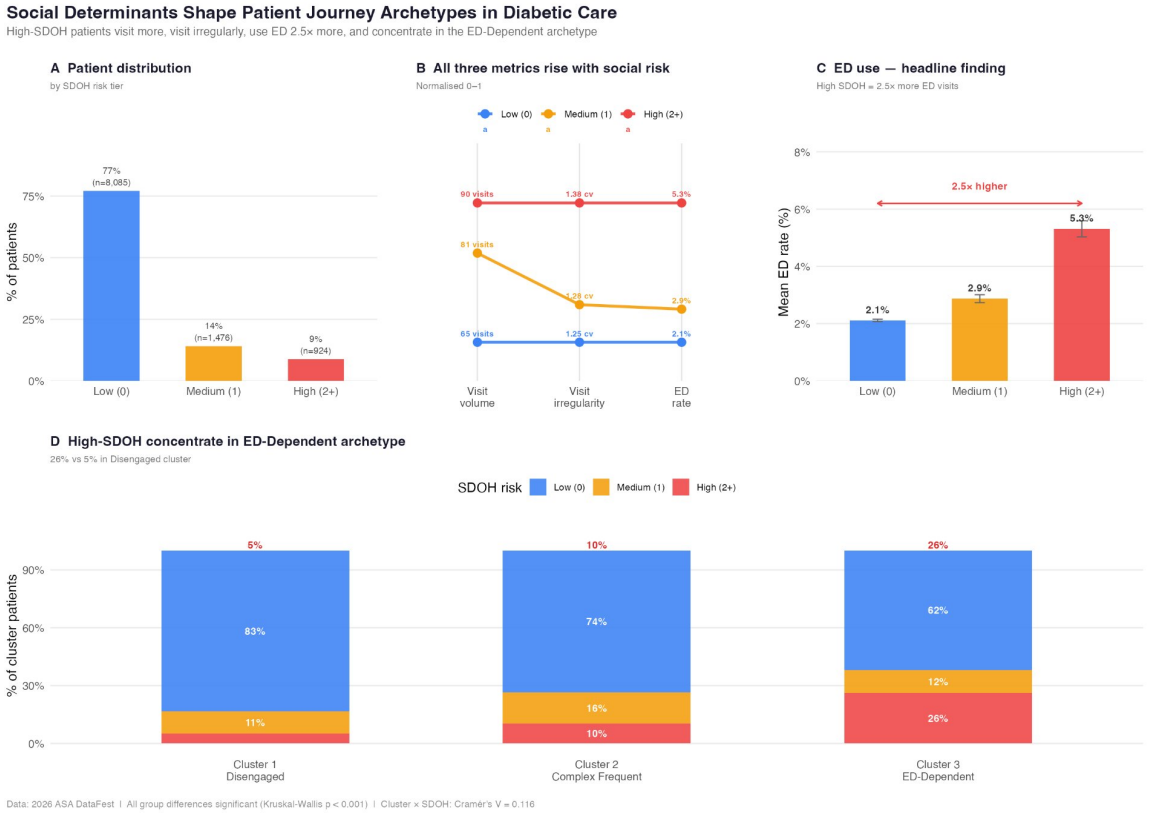


Figure 13. Four-panel summary of the core evidence chain. Panel A: Patient distribution across SDOH tiers (77% Low, 14% Medium, 9% High). **Panel B:** All three metrics—visit volume, visit irregularity (cv_gap), and ED rate—increase monotonically with SDOH risk, demonstrating a system-wide gradient. **Panel C:** Headline ED rate finding ($2.5\times$). **Panel D:** SDOH concentration in each archetype (5%→10%→26%), confirming that social risk deterministically shapes journey type. All group differences significant (Kruskal-Wallis $p < 0.001$); Cramér's $V = 0.116$.

4. Statistical Validation

All key findings were subjected to formal statistical testing before incorporation into the analytical narrative. Triangulation across four independent analytical approaches (descriptive, unsupervised, spatial, causal) substantially strengthens the evidentiary basis.

Table 7. Statistical Tests Supporting Key Findings

Test	Purpose	Result	Interpretation
Two-sample proportion z-test	High vs. Low SDOH ED rates	$p < 0.001$	ED rate difference not due to sampling chance
Kruskal-Wallis + Dunn post-hoc	Gap distributions across SDOH tiers	$p < 0.001$	Distributions differ significantly
Chi-square test	Cluster × SDOH tier	$p < 0.001$, $V = 0.116$	Non-random association
Spearman rank correlation	County % High-SDOH vs. ED rate	$r = 0.82$, $p < 0.05$	Geographic concentration predicts ED over-use
Logistic regression Wald tests	Individual SDOH factor ORs	Housing, Finance, Transport $p < 0.05$	Three barriers have independent significant effects
Silhouette analysis	Optimal cluster number	$k = 3$ maximises score	Three-cluster solution most internally coherent

5. Policy Implications and Practical Applications

5.1 The Three Priority Intervention Targets

[Finance] Intervention 1 — Financial Barrier Reduction (-5.9 pp)
Financial strain prevents medication adherence and delays care-seeking until symptoms are severe. Interventions: medication assistance programmes, waiving copays for chronic disease management visits, integrating financial counsellors into care teams.

[Transport] Intervention 2 — Transportation Access (-4.7 pp)
Transportation barriers prevent attendance at scheduled outpatient appointments, creating gaps that lead to decompensation and eventual ED presentation. Interventions: non-emergency medical transportation (NEMT) benefit expansion, telehealth for routine diabetes monitoring, ride-share partnerships.

[Housing] Intervention 3 — Housing Stability Support (-3.4 pp)
Housing instability disrupts medication storage, provider relationships, and creates competing stressors that crowd out health management. Interventions: care coordination for patients in unstable housing, warm handoffs to social workers at ED encounters, collaboration with housing assistance programmes.

5.2 Geographic Targeting

The $r = 0.82$ county-level correlation supports a **geographically targeted** strategy. Rather than implementing uniform system-wide programmes, resources concentrated in high-burden counties (exemplified by Geary County) maximise population health return on investment. The Early Warning Dashboard operationalises this: it surfaces a real-time watchlist of patients whose visit patterns (rising cv_gap, recent ED encounters, declining outpatient visits) signal impending

care disruption before emergency presentation.

5.3 System Redesign Imperatives

- **Care coordination reimbursement:** The fee-for-service model does not reimburse proactive outreach (calls to schedule appointments, transport coordination, social work referrals) that prevents ED overuse. Value-based payment models rewarding avoided ED encounters and maintained care continuity create aligned incentives.
- **Integrated SDOH screening:** SDOH screening currently occurs primarily at outpatient encounters—the very encounters high-risk patients least attend. Integrating screening into *every* ED encounter would capture patients invisible to the current social risk identification system.
- **Community health worker infrastructure:** Patients in the Disengaged archetype are unreachable by clinic-based interventions. CHWs embedded in high-burden communities provide a pathway to re-engage this invisible population.
- **Transportation as healthcare infrastructure:** The -4.7 pp ED-risk reduction from eliminating transportation barriers provides the evidence base for treating transportation access as a healthcare capital investment.

6. Strengths and Limitations

6.1 Methodological Strengths

1. **Longitudinal design:** Complete patient trajectories enable identification of care patterns invisible in snapshot analyses.
2. **Multi-method triangulation:** Core finding supported by four independent analytical approaches (descriptive, unsupervised, spatial, causal) with different assumptions — convergence substantially strengthens evidence.
3. **Clinically grounded feature engineering:** cv_gap, gap classification, and fragmentation flag designed with clinical interpretability, ensuring policy-actionable findings.
4. **Anchor-based cohort design:** Diabetes as anchor reflects clinical reality (comorbidities co-occur) and maximises SDOH data coverage.
5. **Policy simulation with quantified effect sizes:** Specific, magnitude-quantified intervention effects provide the evidence base for resource allocation decisions.
6. **Performance optimisation:** Vectorised cohort assignment and vroom I/O enable full-dataset analysis without sampling, preserving statistical power.

6.2 Limitations and Caveats

1. **SDOH survey coverage ($\sim$10%):** Patients with SDOH data are likely more engaged than the uncovered 90%. The Disengaged archetype is most underrepresented in SDOH data. True

SDOH impact may be larger than measured.

2. **Synthetic dataset:** Certain artefacts may be present (e.g., food insecurity OR < 1 may reflect a simulation artefact). All effect sizes are illustrative rather than definitive.

3. **Causal inference constraints:** Logistic regression estimates association, not causation. Unobserved confounders (disease severity, health literacy, provider factors) may partially explain the SDOH–ED relationship.

4. **Geographic analysis small n:** $r = 0.82$ is based on ≈ 10 counties; confidence intervals are wide and results should be replicated with a larger geographic sample before informing resource allocation.

5. **Temporal confounding:** No control for time trends within the study period. If high-SDOH patients were more likely in the system during higher-ED-utilisation periods, temporal confounding could inflate the association.

6. **Comorbidity as binary flag:** Severity is not captured. Two patients both flagged for hypertension may have vastly different management complexity.

7. Social Significance and Broader Implications

7.1 Reframing the Health Equity Narrative

This research contributes to a growing body of evidence challenging the framing of healthcare inequality as primarily an *access* problem. Our findings demonstrate that even among patients who ARE in the system—who have diabetes diagnoses, who DO attend visits—social determinants create a secondary layer of inequality in how those patients navigate the system.

This distinction matters enormously for policy design: interventions focusing only on initial access (insurance enrolment, clinic availability) will fail to address the within-system inequality demonstrated here. The solution must address the structural barriers that shape ongoing care behaviour, not merely the front door.

7.2 The Social Cost of ED Over-Reliance

When a diabetic patient with financial barriers presents to the ED in glycaemic crisis instead of attending a scheduled outpatient visit, several inequitable transfers occur:

- The patient experiences a worse health outcome (crisis management vs. prevention).
- The healthcare system absorbs higher costs, distributed across all payers.
- The ED is more crowded for all patients—a negative externality.
- The crisis may precipitate downstream complications (nephropathy, neuropathy, retinopathy), creating future high-cost encounters that compound the initial inequity.

This is a **negative-sum situation**: the healthcare system as a whole performs worse than it would under equitable access to planned care.

7.3 The Early Warning Dashboard as a Model for Proactive Equity

The real-time Early Warning Dashboard developed by the team instantiates a different model of healthcare delivery—one that monitors for warning signs of impending care disruption and intervenes proactively. The dashboard's risk score integrates the exact signals our analysis identifies as predictive—rising cv_gap, recent ED encounters, declining outpatient frequency, SDOH risk tier—enabling care teams to prioritise outreach before crisis presentation.

This model has implications beyond diabetes management. The same analytical approach—identify SDOH risk concentration, monitor longitudinal care patterns, intervene before crisis—generalises to any chronic condition where social barriers create care discontinuity.

8. Conclusions

8.1 Summary of Five Key Findings

1. **High-SDOH patients visit more, but worse:** Greater visit irregularity (cv_gap 1.25 $\rightarrow$ 1.38) and more reactive care patterns.

2. **ED utilisation is 2.5$\times$ higher** among High-SDOH patients (5.3% vs. 2.1%), a quantifiable and policy-addressable system burden inequity.

3. **Three structurally distinct archetypes exist:** High-SDOH patients are 5$\times$ more concentrated in the ED-Dependent archetype (26% vs. 5%).

4. **Healthcare inequality is geographically clustered** ($r = 0.82$), enabling targeted geographic intervention strategies.

5. **Eliminating social barriers could reduce ED-Dependent risk by 16.6 pp** (50.5% $\rightarrow$ 33.9%), representing both a health equity and system efficiency win.

8.2 The Central Argument

Healthcare inequality in chronic disease management operates not primarily through differential access to care, but through **differential quality of care pathways** once patients are inside the system. Social determinants transform what should be proactive, planned chronic disease management into reactive, crisis-driven emergency care. Addressing this requires interventions targeted at the specific barriers that divert patients from planned to emergency care—financial strain, transportation, housing—not merely expanding the system's front door.

"The data tells us not just who is getting sick,
but how the system is failing them—
and precisely what it would take to do better."

— Mia Zhou, ASA DataFest 2026

A. Technical Reference

A.1 R Package Dependencies

Package	Version	Purpose
tidyverse (dplyr, tidyr, ggplot2, stringr)	2.x	Core data manipulation and visualisation
vroom	1.6.x	High-performance CSV I/O (10–50× faster)
lubridate	1.9.x	Datetime parsing and arithmetic
cluster	2.1.x	Silhouette score computation
factoextra	1.0.x	Cluster visualisation and diagnostics
scales	1.3.x	Axis formatting
ggridges	0.5.x	Ridge plot visualisation
broom	1.0.x	Tidy logistic regression output

A.2 Script Execution Order

#	Script	Input	Output
1	`encounter_cleaning.R`	encounters.csv, diagnosis.csv, departments.csv	encounters_clean.csv
2	`cohort_building_fast`	encounters_clean.csv	enc_cohort5.csv, patient_level5.csv, gap_seq5.csv
3	`sdoh_clean_integration`	diabetes_social_survey.csv, patient_level5.csv	sdoh_patient.csv, patient_final.csv
4	`FINAL_ANALYSIS.R`	patient_final.csv, gap_seq5.csv, enc_cohort5.csv	All figures, patient_clusters.csv
5	`model_policy_sim.R`	patient_final.csv, patient_clusters.csv	model1_odds_ratios.png, model2_policy_simulation.png

A.3 Key Variable Definitions

Variable	Definition	Type
`cv_gap`	`sd(gaps) / mean(gaps)` — visit rhythm irregularity	Continuous, per patient
`ed_rate`	$n_{\text{ED}}/n_{\text{total}}$ encounters	Proportion [0,1]
`sdoh_score`	Sum of 8 binary SDOH indicators (range 0–8)	Integer
`sdoh_risk_tier`	Low (0), Medium (1), High (2+)	Ordered factor
`fragmented`	`max_gap` > 75th percentile of all gaps	Binary flag
`ED-Dependent`	Cluster 3 assignment from K-Means	Binary outcome
`EncounterClass`	ED > Inpatient > Observation > Outpatient > Other	Categorical

A.4 SDOH Binary Coding Rules

Variable	Source Question(s)	Coding Rule
`transport_barrier`	2 Transportation Needs Qs	TRUE if any "yes" response
`food_insecurity`	2 Food Insecurity Qs	TRUE if any positive response
`financial_strain`	2 Financial Resource Qs	TRUE if any "yes" response
`housing_instability`	2 Housing Stability Qs	TRUE if homeless or moved $\geq 2\times$
`physically_inactive`	2 Physical Activity Qs	TRUE if < 3 days/week moderate activity
`high_stress`	1 Stress Q	TRUE if "often" or "always" stressed
`depression_screen_pos`	PHQ-2 Total Score	TRUE if score ≥ 3 (validated clinical cutoff)
`socially_isolated`	5 Social Connections Qs	TRUE if $<$ monthly social contact

— End of Report —

ASA DataFest 2026 · Duke University · March 2026 · Author: Mia Zhou