Reflections on the Importance of Data

Assignment 05 — Data Engineering

Introduction

The problem I've chosen to explore is: **What factors drive patient no-show rates in Medicare home health scheduling, and how can a healthcare organization use data to reduce missed visits and improve care?**

Patient no-shows are an important part of my current work where we are trying to optimize for intelligent geographic growth while also maximizing provider efficiency to allow for targeted profitable growth in the next 18 months. Every cancellation wastes provider and scheduler time and decreases productivity, loses revenue and reduces quality of care and patient outcomes. I believe the problem sits at a nexus between a plethora of data points that I believe could predict these events with meaningful accuracy and get them rescheduled (or at least dynamically handled) in real time. The goal of this exercise is to map that nexus, assess what data is important and also obtainable, and identify what could actually be done to improve these outcomes before they happen and respond to them more reliably than present state.

Data Inventory and Ratings

Rating Scales

Importance:

- **(A) Absolutely Critical —** Cannot meaningfully analyze the problem without this data
- **(B) Nice to Have —** Adds depth or context but not essential
- **(C) Probably Unnecessary —** Marginally related; unlikely to move the needle

Accessibility:

- **(1) Impossible to Acquire —** Does not exist in structured form or is legally/ethically restricted
- **(2) Time Consuming, but Possible —** Requires significant effort, negotiation, or manual collection
- **(3) Readily Accessible with Right Permissions or Purchase —** Available via internal systems, paid vendors, or data agreements
- **(4) Readily Accessible to Anyone —** Public datasets etc.

2×2 Grid

	Difficult to Access (1–2)	Readily Accessible (3–4)
Critical (A)	Drive time/distance matrices, Patient barriers to compliance, Competing appointments/jobs/responsibilities	Historical visit records, Patient demographics, Provider schedules/caseloads, EHR diagnosis/care plan
Not Critical (B–C)	Scheduler/outreach comms, Provider knowledge of patients, Caregiver engagement, Patient mental health, Household income, Provider satisfaction/turnover, Census demographics, Neighborhood crime, Patient social media, Patient calendars/schedules	Patient geo/zip, Local weather, RAF scores, Patient satisfaction surveys, National benchmarks

---

Detailed Data Table

#	Data Item	Importance	Accessibility	Why It's Important	Why It's Accessible (or Not)
1	Historical visit records (scheduled vs. completed)	A — Absolutely Critical	3 — Readily accessible with right permissions	This is the core metric that would predict future behaviors absent a major change in the patient's status across the board. We also would have to have this for analysis going forward. This would be the starting point for any predictive model.	This data is stored in our Databricks environment for prior appointments with HC, for other health organizations it would be much more difficult or impossible to acquire.
2	Patient demographics (age, gender, language, race/ethnicity)	A — Absolutely Critical	3 — Readily accessible with right permissions	Age, language and ethnicity are reliable predictors of appointment adherence in Medicare populations.	This data is in our EHR system/feed. It is HIPAA protected and there would be ethical concerns with how it was used but it is available internally

#	Data Item	Importance	Accessibility	Why It's Important	Why It's Accessible (or Not)
					for operational analytics.
3	Provider schedules and caseloads	A — Absolutely Critical	3 — Readily accessible with right permissions	Overbooking, caseload imbalance, and provider-patient matching patterns directly influence whether visits happen as planned and are key to route optimization.	We have internal scheduling system data as well as call log metrics and provider panel data available within our Databricks environment.
4	Patient geographic coordinates/zip code	B — Nice to Have	3 — Readily accessible with right permissions	Geography provides context for rural vs. urban, demographic extrapolation from zips, and focuses many of the other metrics mentioned below.	Stored in patient records and easily geocodable with the right tools and queries.
5	Local weather data	B — Nice to Have	4 — Readily accessible to anyone	Extreme weather can drive cancellations on both the provider and patient side. Also, even mild weather can discourage people from following their routines or going outside.	If geography is known, then weather data becomes relatively easy to surface by patient/by day.
6	Drive time and distance matrices (provider-to-patient)	A — Absolutely Critical	2 — Time consuming, but possible	Long drive times compress schedules, cause provider lateness, and increase the likelihood of missed or rescheduled visits. Also essential for route optimization and geographic growth planning.	We could do it, but with patient churn and other changes this would require a significant lift for the Data engineers.
7	EHR diagnosis and care plan data	A — Absolutely Critical	3 — Readily accessible with right permissions	This would directly affect visit cadence and patient readiness for appointments. Different patient conditions drastically affect their visit compliance.	Lives in our system and is accessible. Certain metrics may create ethical concerns for direct targeting

#	Data Item	Importance	Accessibility	Why It's Important	Why It's Accessible (or Not)
					and campaigns however.
8	Scheduler/Outreach communication success/failure	B — Nice to Have	2 — Time consuming, but possible	Whether a patient was successfully reached on their reminder call or read or confirmed their appointment through text (and when they did it) would be predictive with enough data.	It exists in our Databricks environment, but there hasn't been a meaningful consolidation of this data at the patient level to my knowledge.
9	Patient-reported barriers to visit compliance	A — Absolutely Critical	2 — Time consuming, but possible	I have this as critical, but maybe it's more of a nice to have. If we know WHY they have cancelled in the past we could easily predict similar occurrences in the future. (Patient misses an appointment on June 12th because it's their son's birthday for example.)	Must be collected on follow up outreach calls and/or requested at the time of cancellation. Would be impossible for no-shows in which follow up outreach fails.
10	Provider knowledge of patients (tribal/institutional knowledge)	B — Nice to Have	2 — Time consuming, but possible	Experienced providers often know which patients are likely to cancel and why, but this knowledge lives in their heads and informal notes rather than structured data.	Would require the provider to annotate additional information on their reports, time they likely don't have.
11	Real-time patient RAF scores	B — Nice to Have	3 — Readily accessible with right permissions	RAF scores indicate patient acuity and complexity, which may correlate with no-show patterns, but they're more of a reimbursement metric than a direct predictor of visit compliance.	Available through our data feeds, currently only exists in our bronze layer, so doable and

#	Data Item	Importance	Accessibility	Why It's Important	Why It's Accessible (or Not)
					accessible but need to surface to a silver or gold layer to maintain consistent data retrieval and matching.
12	Caregiver engagement level	B — Nice to Have	1 — Impossible to acquire	For homebound Medicare patients, caregiver presence and engagement is a strong predictor of visit completion. Patients with active caregivers are far more likely to be home and prepared.	Not systematically documented anywhere. Partially noted in care plans, but this is often inconsistent.
13	Patient mental health status	B — Nice to Have	1 — Impossible to acquire	Depression, dementia, and cognitive decline affect a patient's ability to remember appointments, answer the door, or communicate cancellation. Relevant but hard to operationalize.	Diagnosis information is available but only for diagnosed conditions. Severity not so much to my knowledge. Extracting reliable and actionable data to get to a meaningful predictive model would be difficult.
14	Competing appointment schedules, jobs and other responsibilities	A — Absolutely Critical	1 — Impossible to acquire	Patients juggling multiple providers, work schedules, or family obligations frequently cancel home health visits due to conflicts. This is likely one of the most common real-world reasons for no-shows.	Patients rarely share this proactively. Constant change, too much data to re-compute every day.

#	Data Item	Importance	Accessibility	Why It's Important	Why It's Accessible (or Not)
15	Patient household income details	B — Nice to Have	1 — Impossible to acquire	Income correlates with transportation access, phone availability and family support/caregiver availability but this is sensitive information.	Not collected usually, would require self report and doesn't come through any feeds.
16	Patient satisfaction surveys	B — Nice to Have	3 — Readily accessible with right permissions	Dissatisfied patients may be more likely to cancel or disengage. Could reveal provider-specific or service-specific issues driving no-shows.	Anonymous data collected internally through Press Gainey. Response rates vary, and we could only use them generally, but still could be helpful. Perhaps some regional or location based scoring. (I think they go down to the market level so we could differentiate by Rural/Urban and/or metro area.)
17	Provider job satisfaction and turnover	B — Nice to Have	2 — Time consuming, but possible	Disengaged or overburdened providers may contribute to no-shows through lateness or poor bedside manner and focus.	Would require bringing a new datasource from HR records into Databricks which (surprisingly) isn't currently in our Databricks environment. I believe it's a 2026 priority though.

#	Data Item	Importance	Accessibility	Why It's Important	Why It's Accessible (or Not)
18	Neighborhood crime data	C — Probably Unnecessary	2 — Time consuming, but possible	Provider safety concerns in certain areas could theoretically cause visit avoidance or rescheduling, but this is rarely a primary driver of patient no-shows.	Would be fairly easy to get a free resource, but reporting accuracy varies by jurisdiction and normalizing it would be time consuming.
19	Patient social media activity	C — Probably Unnecessary	1 — Impossible to acquire	Unlikely but could potentially contribute to no-show calculations. Would be more valuable in answering other metrics and dimensions noted elsewhere.	Not accessible, ethical concerns with using in this way.
20	Patient calendars and schedules	B — Nice to Have	1 — Impossible to acquire	Knowing a patient's personal schedule would help avoid conflicts, but this overlaps significantly with the competing appointments/jobs/responsibilities item.	Patients manage their own calendars and there's no mechanism to access or integrate them. Constantly changing and entirely in the patient's control.

Closing Reflection

This exercise reinforced what I see in my current role at Harmony Cares, the most explanatory data is often the least accessible. We need to see into these patients lives to understand cancellations and no-shows and most of that information is either not accessible or ethically constrained. The remaining data can get us close though, at least close enough to make a valuable prediction and/or signal outreach personnel on who to call and when. This reminded me to stop chasing easy to get data that may not move the needle and focus on what actually matters to the problem at hand.