PicnicResearch

SA50

Comparison of the Performance of Patient-Mediated Medical Record Retrieval and Tokenization-Based Linkage to Generate Complete and Longitudinal Real-World Data

Ashley Cogell, PhD, Reema Patel, MPH, Kristen Hahn, PhD, MPH

Background

Real-world evidence (RWE) studies increasingly rely on longitudinal patient-level data to assess treatment patterns, disease progression, and long-term outcomes. Two fundamentally different data generation mechanisms are commonly employed: patient-mediated medical record retrieval and tokenization-based linkage (TBL) to various data sources, such as administrative claims. Each approach has distinct data generation processes that may influence longitudinal completeness, retention, and suitability for specific research objectives.

Patient-mediated retrieval (PMR, as performed by PicnicHealth) involves direct acquisition of electronic health records (EHR) through patient authorization, yielding dense clinical documentation including physician notes, laboratory results, and specialty care encounters. Tokenization-based linkage uses probabilistic matching algorithms to connect patients across data sources, enabling access to claims data (including both open and closed claims) that capture billing events and healthcare utilization.

This study compared data generation between TBL to a claims data source (TBL cohort) and patient-mediated EHR retrieval (PMR cohort) among multiple sclerosis (MS) and hemophilia A (HA) populations.

Key Objectives:

- Characterize baseline linkage performance by quantifying the proportion of PicnicResearch patients successfully linked to claims via tokenization.
- Compare longitudinal data completeness by describing observable periods, retention patterns, and condition-specific specialty care (i.e., neurology, hematology) capture between patient-mediated EHR retrieval and TBL to claims.
- Identify token quality issues by quantifying token degradation patterns including token splits, broken linkages, and weak longitudinal coverage.

Methods

The analysis employs a within-patient comparison design, leveraging the subset of patients who have both PicnicHealth medical records and tokenization-based linkage to claims data comprised of open and closed claims. This design enabled direct assessment of data source differences while controlling for patient-level confounding.

COHORTS

- The MS and HA cohorts were defined using physician-documented diagnoses extracted from existing PicnicResearch studies (PMR cohort).
- Patients present in the PMR cohort and successfully linked to closed claims with a token assigned using soundex name, date of birth, and sex, represented the TBL cohort.

DEFINITIONS

- The **observation window** was a 5-year period (September 1, 2019 to September 1, 2024) and served as the temporal anchor for all longitudinal metrics (retention, age stratification, token degradation timing).
- **Retention** quantifies the presence of any encounter/claim at a given time point.
- **Token split** indicates that a patient has multiple distinct token IDs over the observation window.
- **Token collision** indicates multiple patients map to the same token ID whereby distinct individuals incorrectly share a token.
- **Broken token** reflects where medical records continue >6 months after claims cease, suggesting claims data ended while patient remained active in a PicnicResearch study.
- **Weak coverage** indicates that the claims observable period is substantially shorter than the period for which PicnicResearch has medical records.
- **Time to token degradation** is defined from observation window start to last claim in claims data among patients with broken tokens.

ANALYSIS

- Descriptive statistics are reported.
- All comparative analyses of longitudinal outcomes (including data availability and retention), token quality, and data source discordance are restricted to patients with ≥1 medical encounter in the observation window, as determined through patient-mediated retrieval encounter in the observation window, ensuring comparability with the TBL cohort (which by definition requires ≥1 claim).

Table 1: Cohort Linkage at Baseline

	Multiple Sclerosis	Hemophilia
Patient-mediated retrieval cohort (≥1 encounter)	4669	254
Tokenization-based linkage cohort (PMR cohort with ≥1 claim)	4432	243
Linkage rate at baseline	94.9%	95.7%

Table 2: Longitudinal Data Completeness

Multiple Sclerosis (N=4432)	**PMR Cohort**	**TBL Cohort**	**P-value**
Observable period (years), median (IQR)	4.45 (3.49–4.77)	4.77 (4.24–4.93)	<0.001**
Total encounters or claims, median (IQR)	87.0 (37.0–196.0)	114.0 (54.0–215.0)	<0.001**
Specialty encounters or claims, median (IQR)	14.0 (7.0–26.0)	7.0 (3.0–15.0)	<0.001**
Retained at 3 years, %	93.4	96.4	<0.001**
Retained at 4 years, %	73.4	90.2	<0.001**
Hemophilia (N=234)	**PMR Cohort**	**TBL Cohort**	**P-value**
Observable period (years), median (IQR)	4.52 (4.01–4.83)	4.49 (3.75–4.85)	0.009*
Total encounters or claims, median (IQR)	192.0 (82.5–389.0)	59.0 (23.0–125.5)	<0.001**
Specialty encounters or claims, median (IQR)	35.0 (18.0–58.5)	4.0 (1.0–9.0)	<0.001**
Retained at 3 years, %	97.9	92.2	0.004*
Retained at 4 years, %	90.5	80.2	<0.001**

*/** P-values from Wilcoxon signed-rank tests (continuous outcomes) and McNemar's tests (retention outcomes). Significance: *p<0.01, **p<0.001

Table 3: Condition-Specific Specialty Care Capture Among Retained Patients

	Multiple Sclerosis		Hemophilia	
	PMR	TBL	PMR	TBL
Retained at 3 years	4138	4271	238	224
Specialty record capture from prior 12 months	64.5%	47.8%	87.8%	35.7%
Specialty record capture from prior 24 months	96.2%	67.2%	100%	62.1%
Retained at 4 years	3251	3996	220	194
Specialty record capture from prior 12 months	95.5%	51.1%	96.3%	41.2%
Specialty record capture from prior 24 months	99.5%	69.7%	100%	64.9%

Table 4: Token Quality Metrics

	Multiple Sclerosis (N=4432)	Hemophilia (N=243)
Any token issue, n (%)	421 (9.5%)	57 (23.5%)
Token split, n (%)	32 (0.7%)	2 (0.8%)
Token collision, n (%)	0 (0.0%)	0 (0.0%)
Broken token, n (%)	344 (7.8%)	53 (21.8%)
Weak coverage, n (%)	172 (3.9%)	23 (9.5%)
Mean (SD) time to token degradation, years	3.21 (1.07)	3.26 (1.01)

Results

- Baseline tokenization linkage from available medical records to claims was high (94.9% in MS, 95.7% in HA; Table 1), enabling within-patient comparisons of the PMR and TBL cohorts.
- For MS, retention was greater within the TBL cohort at 3 (93.4% vs. 96.4%; p<0.001) and 4 (73.4% vs. 90.2%; p<0.001) years.
- For HA, retention was greater within the PMR cohort at 3 (97.9% vs. 92.2%; p=0.004) years and 4 (90.5% vs. 79.8%; p<0.001) years.
- The PMR cohort captured approximately 2x more median neurology encounters per MS patient than the TBL cohort (14.0 vs 7.0; p<0.001) and approximately 9x more median hematology encounters per HA patient (35.0 vs 4.0; p<0.001) (Table 2).
- For both MS and HA patients retained over time in the TBL cohort, condition-specific specialty claims were absent for 35%-64% of patients at 3 and 4 years (Table 3).
- Broken tokens occurred in 7.8% (MS) and 21.8% (HA) of patients and the mean time to degradation was approximately 3.2 years in both (Table 4).

Conclusion

- Baseline tokenization-based linkage of the PMR cohort with claims data was high (>94%).
- Both TBL and PMR cohorts demonstrated >70% retention, based on presence of claims or encounters, at 3 and 4 years.
- Overall, the PMR cohort resulted in significantly more clinically relevant specialty care encounters at 3 and 4 years than the TBL cohort, indicating better recovery of disease-relevant clinical signals to meet study objectives.
- Limitations include the implementation of tokenization-based linkage using a specific token type and a defined observation window; findings may not generalize to other tokenization approaches, vendors, or linkage configurations.

Disclosures

Authors are employees of PicnicResearch.