Follow Me, If You Can: How Tokenized Linkage Extends Patient Observations in Real-World Databases — the Komodo Research Dataset Example

Ting-Ying Huang, Yuqin Wei — Komodo Health, New York, NY, and San Francisco, CA

ISPOR 2025
RWD134

Introduction

- Federated data networks[1,2] have arisen as a solution to sample size and privacy challenges in observational studies. Yet their distributed structure suffers from an inability to track individuals across multiple databases or insurance plans (in the case of claims datasets) over time.
- Tokenization[3] is an advanced method for addressing all three issues listed above – sample size, privacy preservation, and longitudinality.

Objective

- Using a next-generation tokenized database as example, this study assessed the impact of tokenized linkage on data capabilities of sample size and longitudinality in health insurance claims sources.

Methods

Study Design

- This retrospective cohort study examined the Komodo Research Dataset, a claims database with pseudonymized, person-level linkage across sources enabled by encrypted, tokenized identifiers.

Komodo Research Dataset (KRD): *Composed of administrative data and claims, KRD captures routinely collected health services utilization records and expenditures for over 330 million de-identified unique individuals in the U.S. Native to HIPAA-compliant, privacy-preserving tokens, KRD offers extended patient-level observations of medical encounters and outpatient pharmacy dispensings via linkage across health and pharmacy insurance plans. Data availability is as early as 2016. Specialty datasets such as genomics, laboratory test results, and electronic medical records are readily accessible via additional linkage. KRD is the optimized schema of the underlying Healthcare Map® from Komodo Health for real-world evidence (RWE) generation and health economics and outcomes research (HEOR).*

Komodo Race & Ethnicity (KRE): *Self-reported or health care provider/system-assigned race and ethnicity information for over 200 million unique individuals obtained from assorted categories of data sources, including EHR, patient intake forms, payer enrollment files, and statistically reliable consumer reporting agencies. Value-standardized and pre-certified for dataset linkage via privacy-preserving tokens.*

Inclusion/Exclusion Criteria:

Eligible members must have

- Contributed claims from payer-complete (closed) sources,
- Had ≥1 payer information recorded over time, and
- Had ≥1 enrollment day in medical and drug plans between January 2016 through October 2024

Enrollment span: defined as the continuous period in which a member had simultaneous medical and drug coverage, with gaps of ≤45 days allowed

Main Analyses

- Distributions of the continuous enrollment spans separately with and without tokenized linkage upon payer change — payer-neutral vs payer-segmented enrollment spans
- Distributions of span extensions at both person and span levels after tokenized linkage

Results

- Of 243,383,067 total individuals (as represented by the unique token keys) who were identified from closed sources and with 1+ day enrollment, 197,556,385 (81.2%) also satisfied the payer information requirement and met eligibility criteria
- Among eligible members, the mean age was 34.7 years, 52.2% were female, and 59.4% were commercially insured as of the first day of the most recent payer-segmented enrollment span (Table 1)
- At the enrollment span level, from payer-segmented to payer-neutral spans:
  - Number of accrued enrollment spans consolidated by 10.6% from 316,099,620 to 282,445,948 spans (Figures 1-2)
  - Mean (SD) length of enrollment spans increased by 34.9% from 789 (804) to 1,064 (906) days (Figure 3)
  - Median (IQR) length of enrollment spans increased by 55.8% from 489 (974) to 762 (1,279) days (Figure 3)
- At the person level, comparing the longest payer-neutral spans versus (Figure 4):
  - Shortest payer-segmented: mean (SD) increase by 554 (789) days, despite 47.2% members had 0-day extension
  - Longest payer-segmented: mean (SD) increase by 271 (527) days, despite 61.3% members had 0-day extension
- Sensitivity analysis suggested that when enrollment gap allowance relaxed from 45 days to infinite (i.e., all gaps bridged):
  - Span: 23.9% consolidation in the number of spans, respectively 68.1% and 136.9% increase in their mean and median lengths
  - Person: average increase by 744 and 505 days, respectively, in the length of the longest payer-neutral span from the shortest and longest payer-segmented spans

Figure 1. Number of Continuous Enrollment Spans by Length

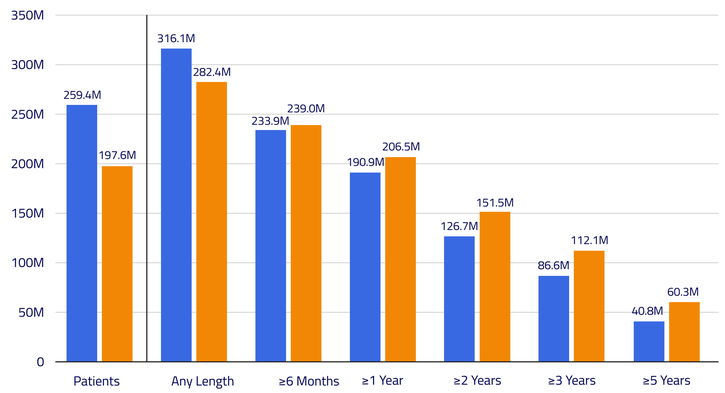


Figure 2. Percentage of Continuous Enrollment Spans by Length

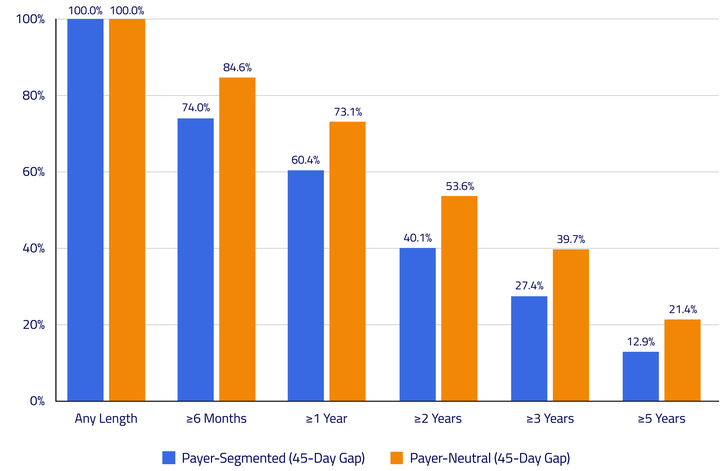


Table 1. Characteristics of Eligible Members

	n	%
Number of patients	**197,556,385**	**100.0**
Age, years (mean, SD)	34.7	22.0
Age group, years		
≤18	55,024,088	27.9
19-44	73,400,674	37.1
45-64	48,326,539	24.5
65-74	13,574,606	6.9
≥75	6,818,622	3.5
Sex		
Female	103,038,086	52.2
Male	92,788,550	47.0
Insurance type		
Commercial	117,389,630	59.4
Medicaid	59,096,781	29.9
Medicare	20,913, 118	10.6
Unknown	156,856	0.1

	n	%
Number of patients	**197,556,385**	**100.0**
Race		
White	78,242,952	39.6
Asian or Pacific Islander	8,617,080	4.4
Black or African American	23,129,260	11.7
Other	11,102,435	5.6
Unknown	76,464,658	38.7
Region		
Boston	9,975,208	5.0
New York	20,896,313	10.6
Philadelphia	19,140,338	9.7
Atlanta	39,690,093	20.1
Chicago	32,230,969	16.3
Dallas	27,063,916	13.7
Kansas City	7,427,074	3.8
Denver	6,135,103	3.1
San Francisco	26,480,629	13.4
Seattle	8,471,983	4.3

Figure 3. Length of Continuous Enrollment Spans

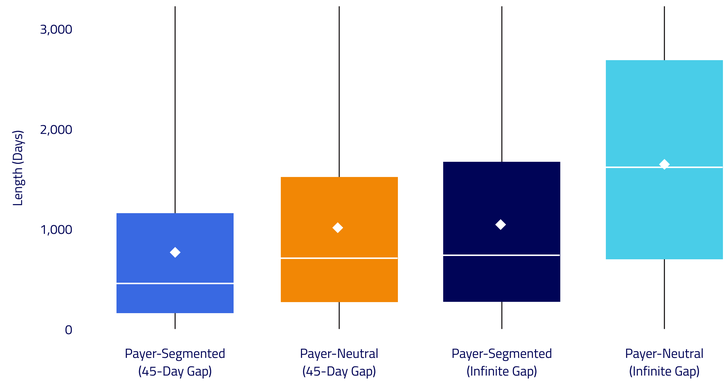
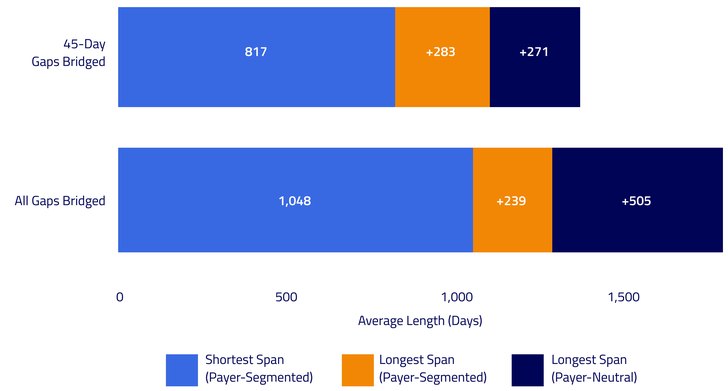


Figure 4. Average Length of Continuous Enrollment Spans Among Eligible Members



Conclusion

- This is one of the first evaluations to quantify the effect of tokenized linkage on a claims database.
- Findings verify how observable time in a claims database can be extended by following patients across multiple insurance plans based on their pseudonymized identifiers in the form of tokens or token combinations.
- Comprising over 240 millions lives and up to 2.4 times longer token-enhanced person-time in the database, the Komodo Research Dataset demonstrates robustness and evidentially offers competitive advantages of scale, continuity, and privacy-preservation among real-world data sources for medical research.

References

[1] Brown, J.S., Mendelsohn, A.B., Nam, Y.H., Maro, J.C., Cocoros, N.M., Rodriguez-Watson, C., Lockhart, C.M., Platt, R., Ball, R., Dal Pan, G.J. and Toh, S., 2022. The US Food and Drug Administration Sentinel System: a national resource for a learning health system. *Journal of the American Medical Informatics Association*, 29(12), pp.2191-2200.

[2] Palchuk, M.B., London, J.W., Perez-Rey, D., Drebert, Z.J., Winer-Jones, J.P., Thompson, C.N., Esposito, J. and Claerhout, B., 2023. A global federated real-world data and analytics platform for research. *JAMIA Open*, 6(2), p.ooad035.

[3] Will Howe et al. Evaluating the Performance of Privacy Preserving Record Linkage Systems (PPRLS). Cancer Moonshot Task Order Phase 2, National Cancer Institute, March 27, 2023. https://surveillance.cancer.gov/reports/TO-P2-PPRLS-Evaluation-Report.pdf

Scan here to download poster or inquire for more info.

komodo®