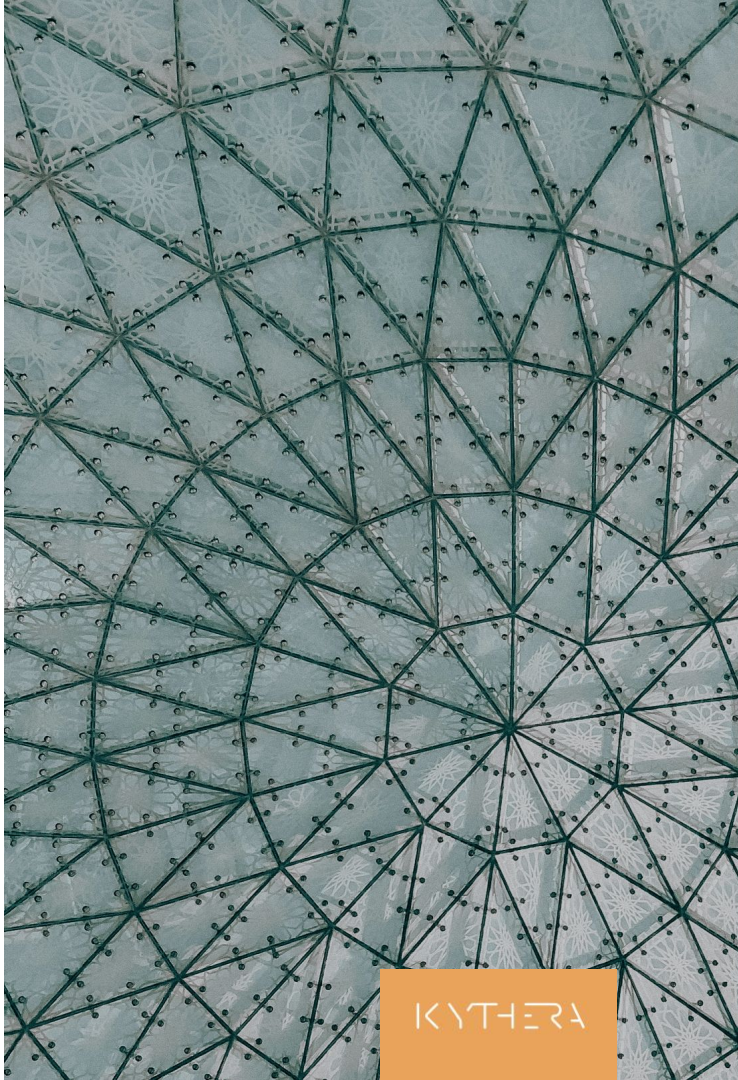
Data Workshop:
Tackling RWD Integration Challenges

KYTHERA
DECIPHERING HEALTHCARE | NOVARTIS

Kythera Labs' Mission

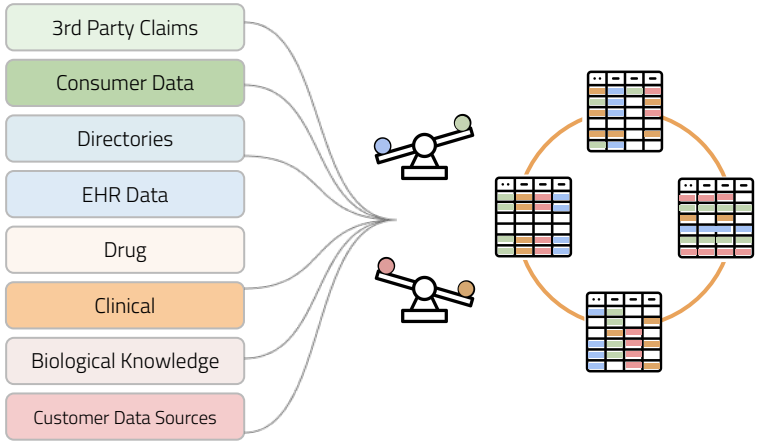
Reduce the uncertainty in the use of real-world data.



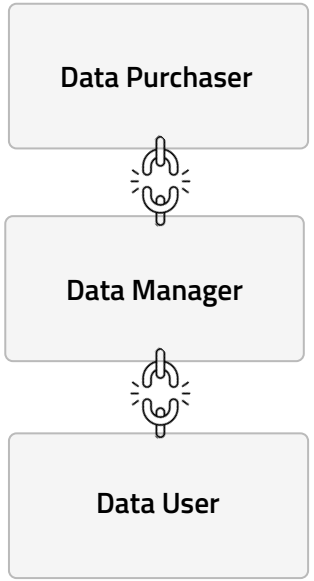
©2025 Kythera Labs, Inc. All rights reserved.

Understanding the Real-World Data Integration Problem

1. Data source **bias**, **fragmentation**, and **overlap**



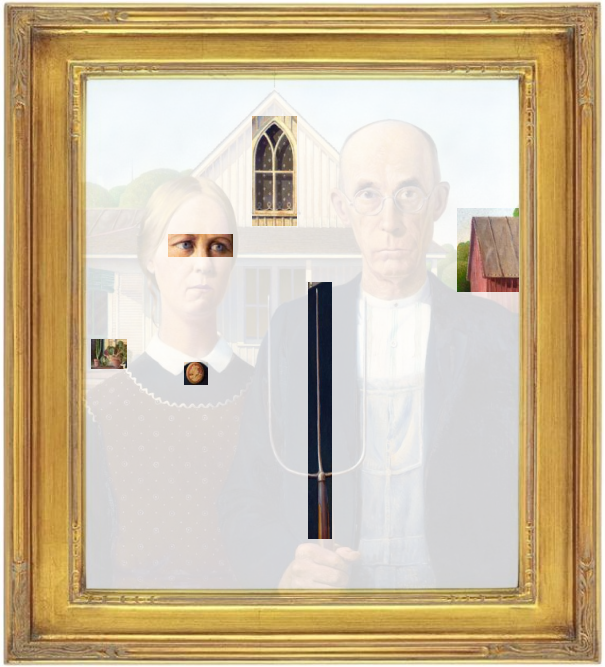
2. Common **disconnect** between data purchasers, managers, and users



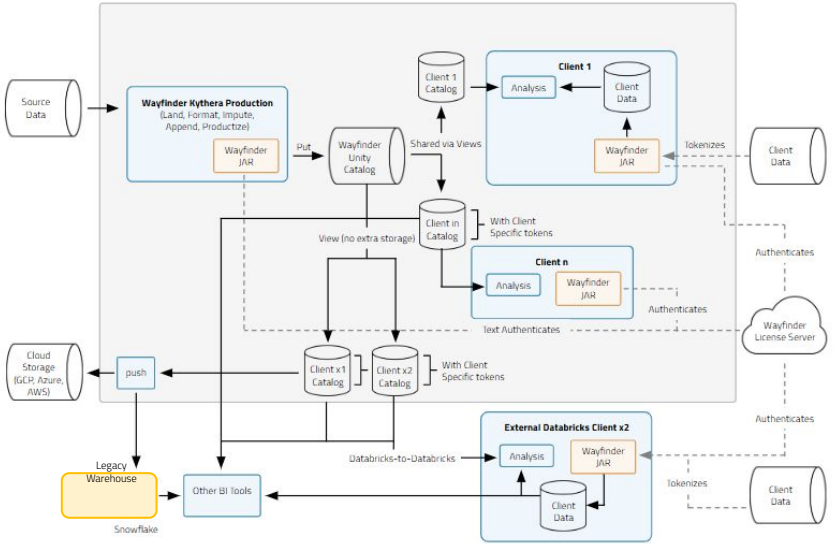
©2025 Kythera Labs, Inc. All rights reserved.

Understanding the Real-World Data Integration Problem

3. Seeing only **parts** of the picture



4. A **simpler** solution is highly **complex**



©2025 Kythera Labs, Inc. All rights reserved.

KYTHERA

The True Costs of Real-World Data

Visible Costs

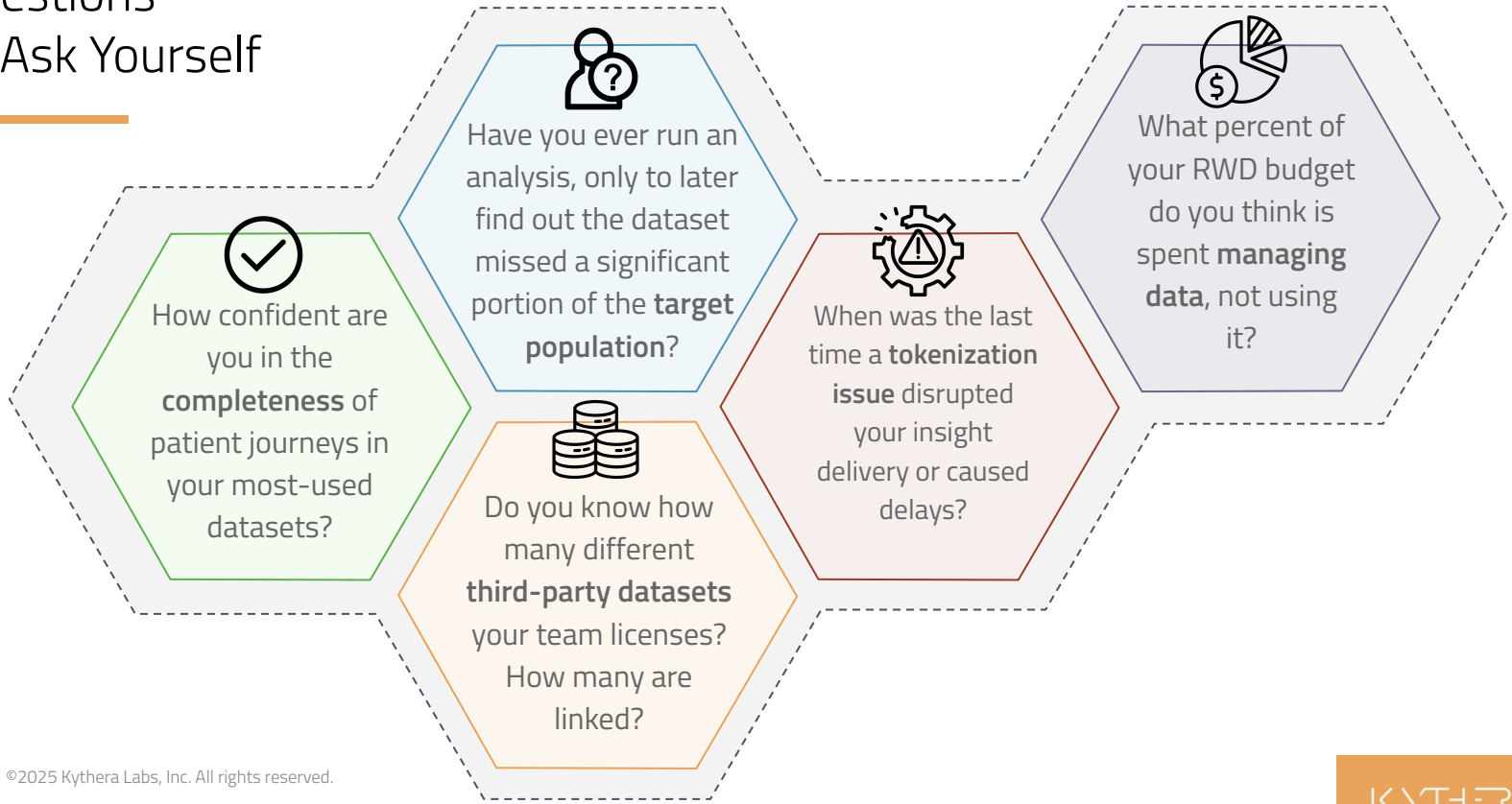
- Data Licensing Fees
- Integration Tools
- Internal Resources
- External Resources

Not-So-Visible Costs

- Time delays
- Insufficient clean-up results
- Redundant purchases
- Opportunity costs
- Failed use cases
- Missed regulatory insights
- Misaligned commercialization

©2025 Kythera Labs, Inc. All rights reserved.

Relevant Questions To Ask Yourself



©2025 Kythera Labs, Inc. All rights reserved.

High Level Outline

- Use case can benefit from combination of 3rd party data sets to improve accuracy and completeness of analyses
- 3rd party data often de-identified for privacy and compliance
- De-identification introduces new challenges, patient mastering addresses these
    - Patient duplication within a single dataset
        - Gaps in patient history
        - **Method: Identify “suspect tokens”**
        - **Method: Match tokens and dedupe**
    - Token scheme mismatches between datasets
        - Token “Drop-off” when joining
        - **Method: Token bridging**
- Assessing Completeness of data to support analyses
    - **Method: Effective overlap analyses**
    - **Method: Identify unique uplift of a data set**

KYTHERA

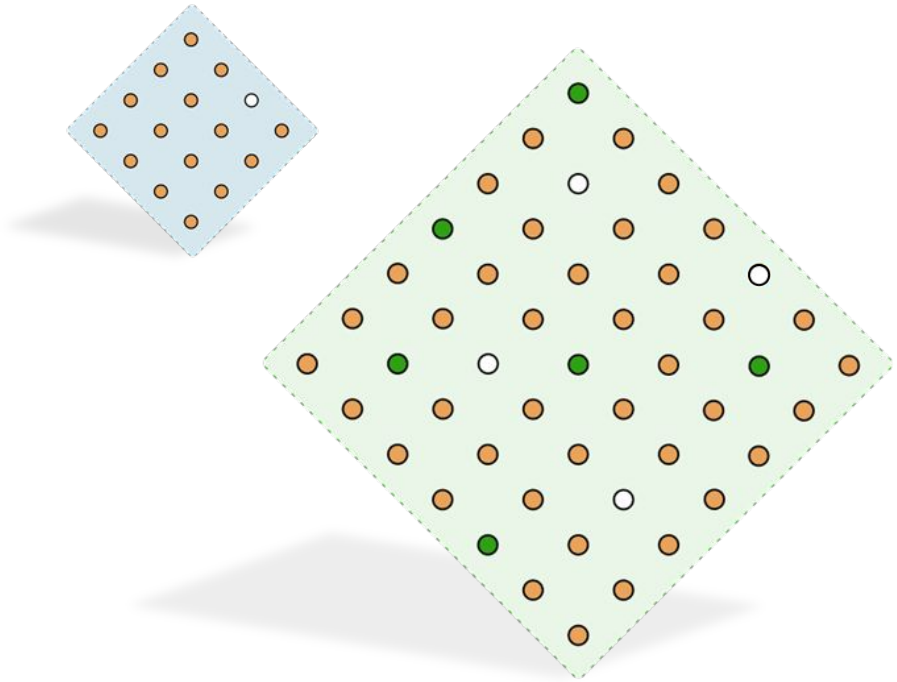
Closed & Open Claims

Patient Encounter Data:

Closed Claims

PROS
- Adjudicated claims data
- Full records of care
- High rate of data completion

CONS
- Expensive
- Payer-biased
- Region-biased
- Latency
- Short window of visibility
- Small sample size



- Encounter Data
- Kythera Imputed data
- Missing data

Patient Encounter Data:

Open Claims

PROS
- Less costly
- Frequently updated
- Longer window of visibility
- Much larger sample size
- Nationally representative

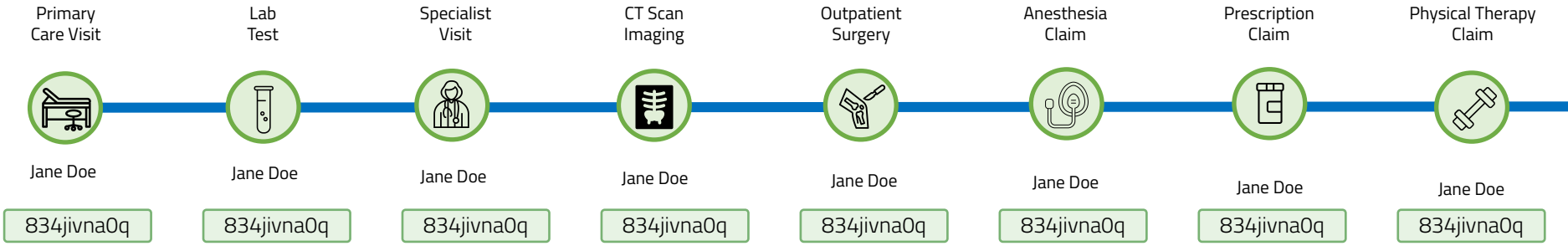
CONS
- ~~Provider biased~~
- ~~Input errors~~
- ~~Missingness~~
- ~~Patient duplication~~

©2025 Kythera Labs, Inc. All rights reserved.

KYTHERA

Patient Journey Ideal State

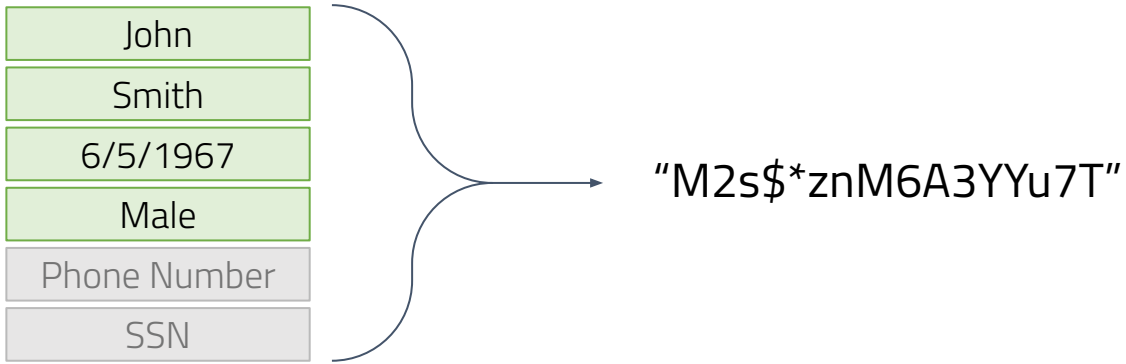
One Patient : One Journey



©2025 Kythera Labs, Inc. All rights reserved.

Tokenization Intro

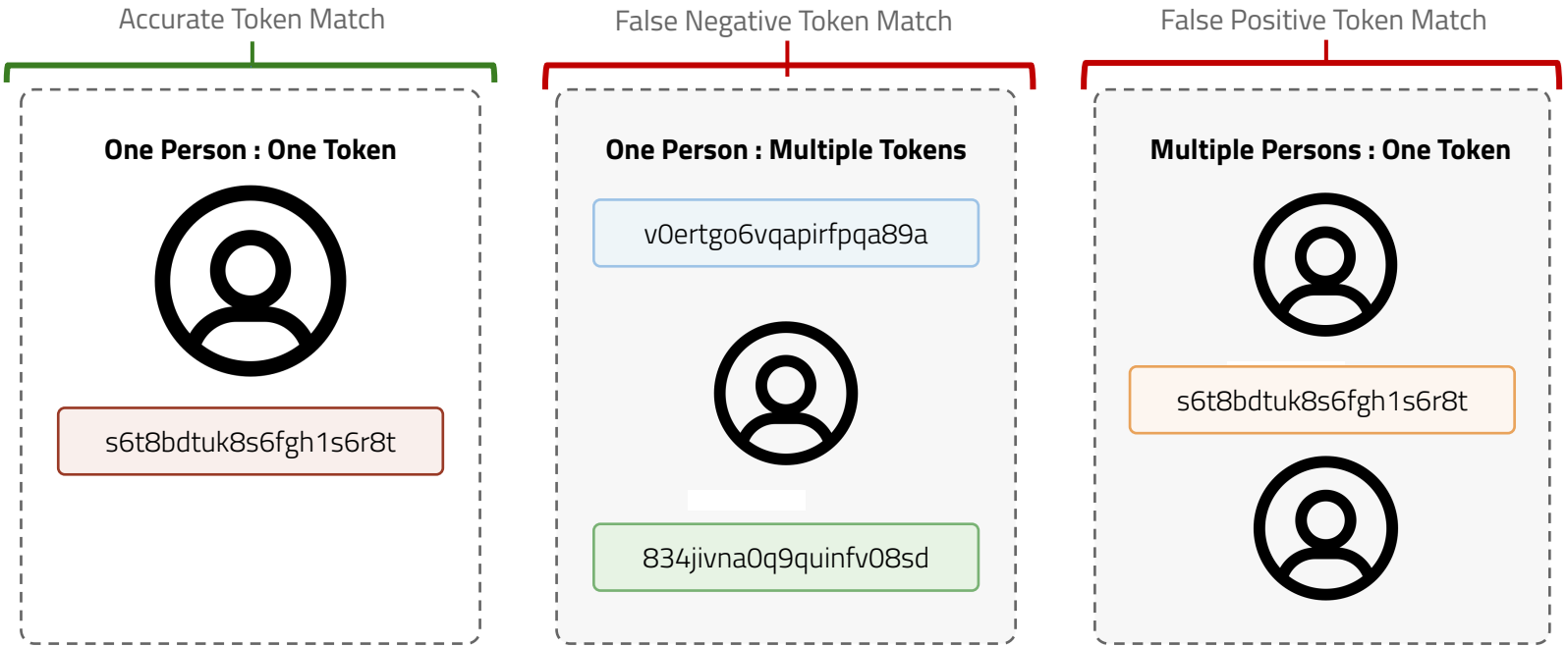
- De-Identification
  - John Q. Smith → “M2s$*znM6A3YYu7T”
  - Increase privacy
  - Increase acceptable use cases for data
- Security
  - Prevent leaks of PII
  - Enable collaboration between organizations that require patient identifiers



©2025 Kythera Labs, Inc. All rights reserved.

Tokenization Can Introduce New Errors

Challenges with tokenization for the patient journey

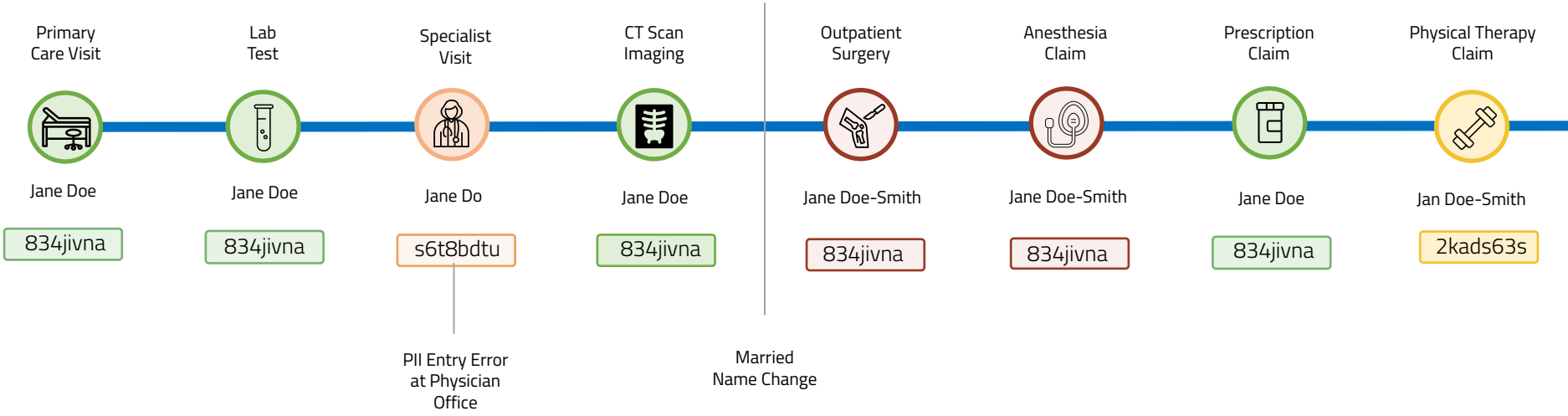


©2025 Kythera Labs, Inc. All rights reserved.

KYTHERA

Patient Journey Fragmented State

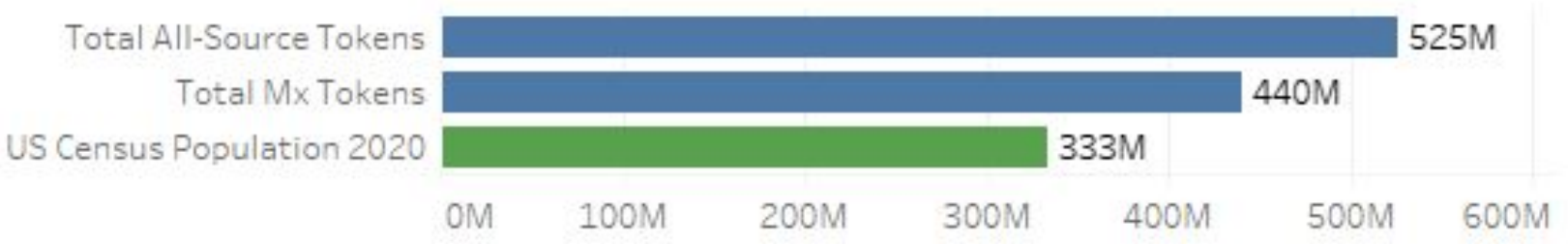
One Patient : Multiple Tokens



©2025 Kythera Labs, Inc. All rights reserved.

Extent of the Problem: Single Data Source

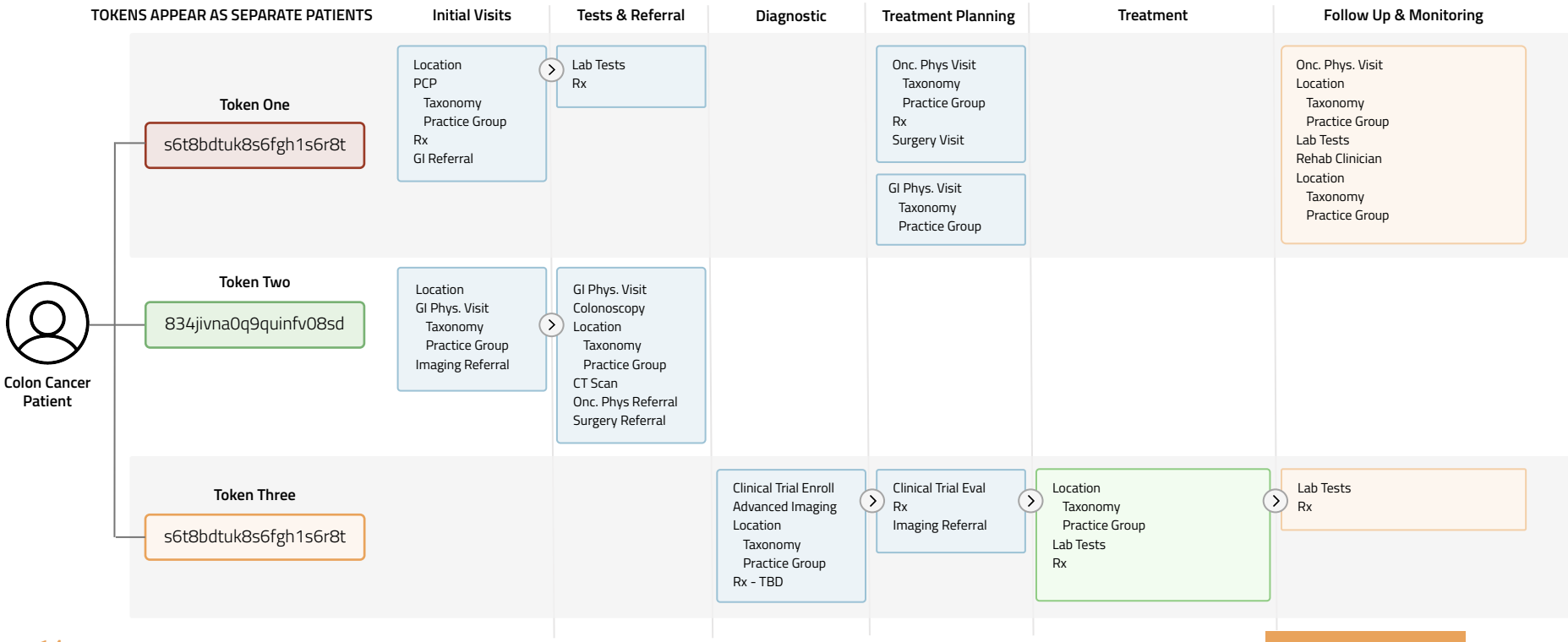
- From a single vendor of clearinghouse claims (Mx and Rx) Kythera sees over **520M** distinct token combinations (token1 and token2 together), which is clearly impossible to for a dataset containing only information on US individuals if those are supposed to be distinct patients
- Considering the medical submits only there are **440M** distinct token combinations – better, but still far too high



©2025 Kythera Labs, Inc. All rights reserved.

Token Misalignment

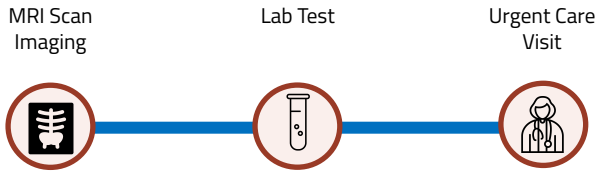
Multiple tokens result in disparate, unreliable, inaccurate event capture



©2025 Kythera Labs, Inc. All rights reserved.

Suspect Token Identification

- Suspect tokens are tokens that are statistically likely to represent only a fragment of a complete patient journey.
- Suspect tokens may appear for many reasons, such as by a patient changing their name or an input error at the provider level.

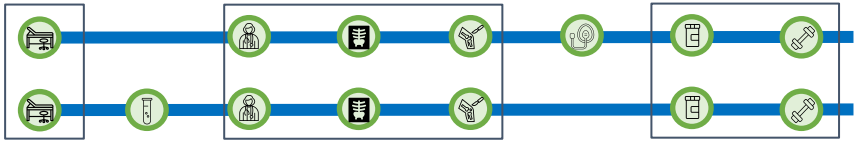


Age:	63 years of age
Gender:	Females
Diagnosis Code:	M4726 (Degenerative Spine Condition)
Procedure Code:	72149 (MRI)
Number of patients in the Kythera claims set who present with both this primary diagnosis and procedure	1,119 patients
Mean chronological count	166 (for the 1119 patients, the average number of claims generated by the patients before the MRI for spondylolysis of the lumbar region is 165).
Standard Deviation of the chronological count	220
Chronological count for the lowest 1% of the population	149

©2025 Kythera Labs, Inc. All rights reserved.

Patient Dedupe Methodology #1

- Train a statistical ML model using closed claims
  - Likelihood of patients appearing at same provider, same day, same procedure
  - Prior distribution on likelihood of patients sharing same chronic diagnoses and comorbidities
- Identify candidate token matches using similarity measures
- Evaluate the quality of the candidate match

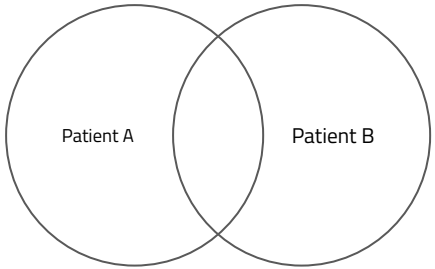


Patient A:

- # of claims
- # of claims overlapped with patient B
- # of chronic conditions similar to patient B
- # of chronic conditions exclusive to A

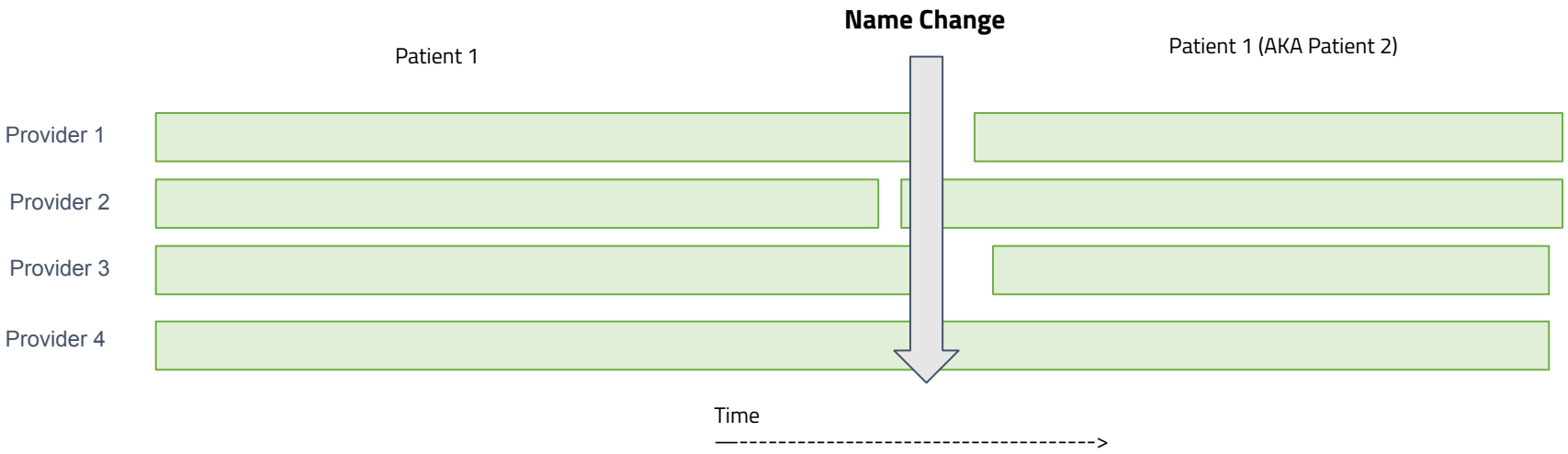
Patient B:

- # of claims
- # of claims overlapped with patient A
- # of chronic conditions similar to patient A
- # of chronic conditions exclusive to B



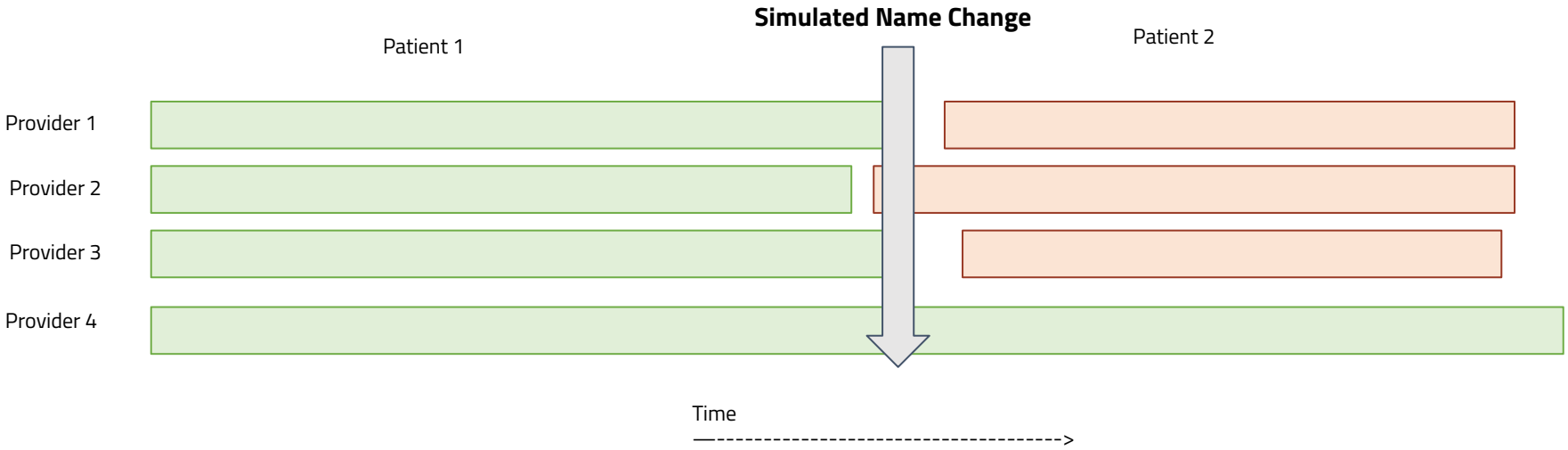
©2025 Kythera Labs, Inc. All rights reserved.

Patient Dedupe Method #2: Synthetic Data



©2025 Kythera Labs, Inc. All rights reserved.

Patient Dedupe Method #2: Synthetic Data



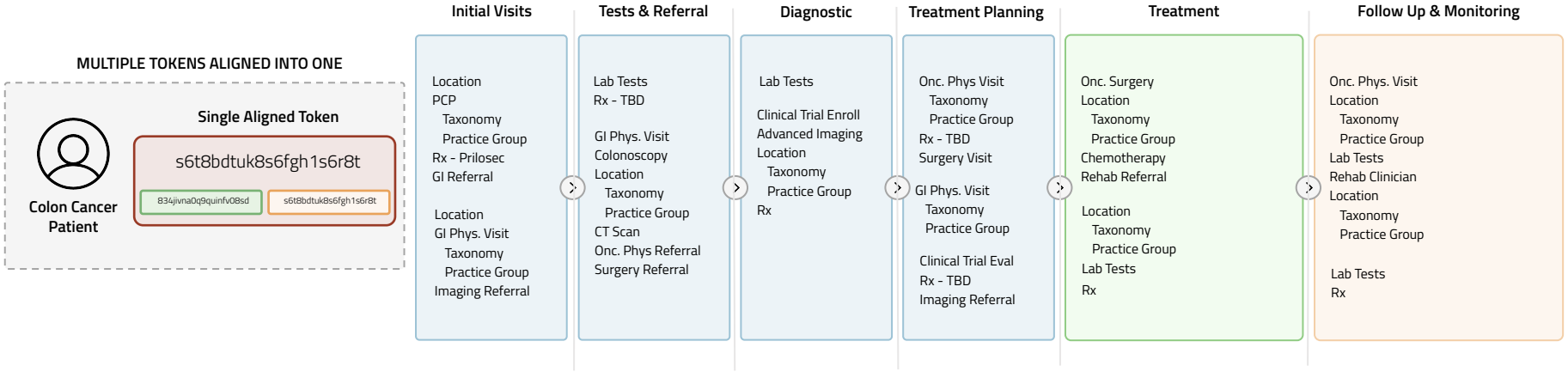
Simulate a “bad” match by modifying closed claims to product synthetic training data

- Use claims from two separate patients
- Truncate them before and after an arbitrary “name change date”

©2025 Kythera Labs, Inc. All rights reserved.

Token Alignment

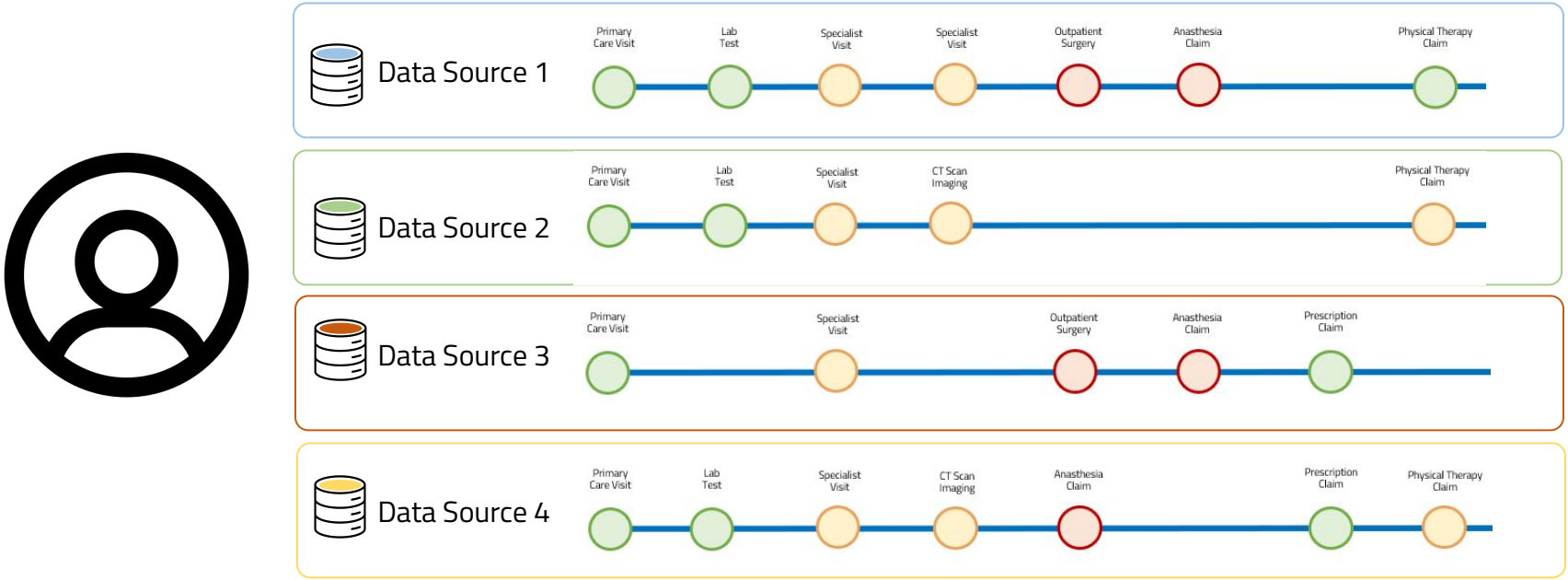
Token alignment leads to far more accurate event capture



©2025 Kythera Labs, Inc. All rights reserved.

Patient Journey Fragmented State: Multiple Sources

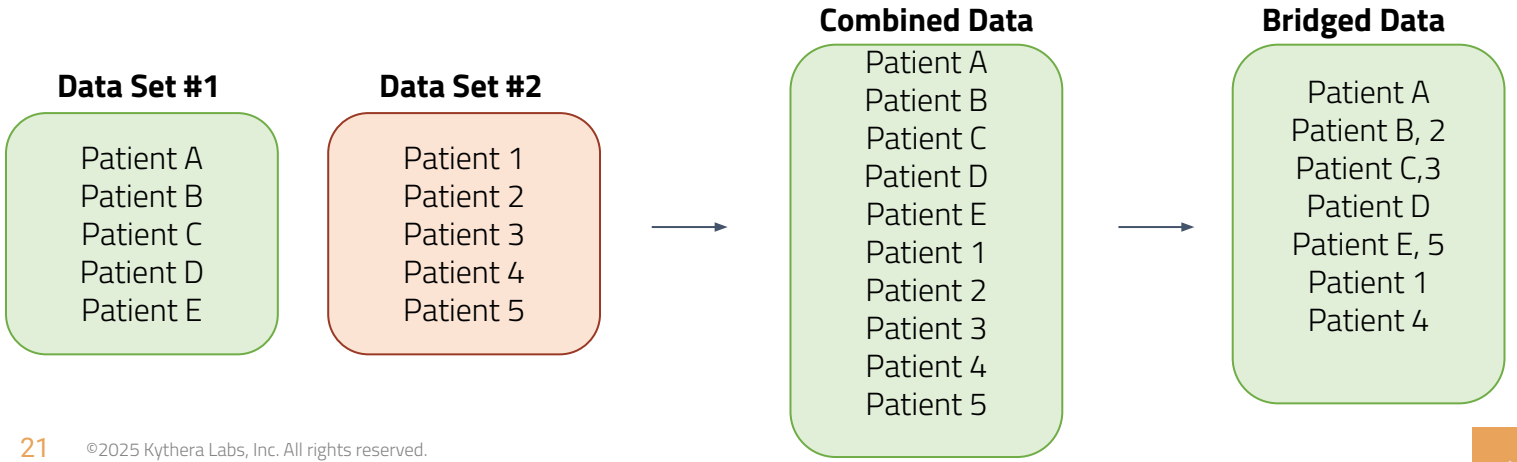
One Patient : Multiple Tokens : Multiple Data Sources



©2025 Kythera Labs, Inc. All rights reserved.

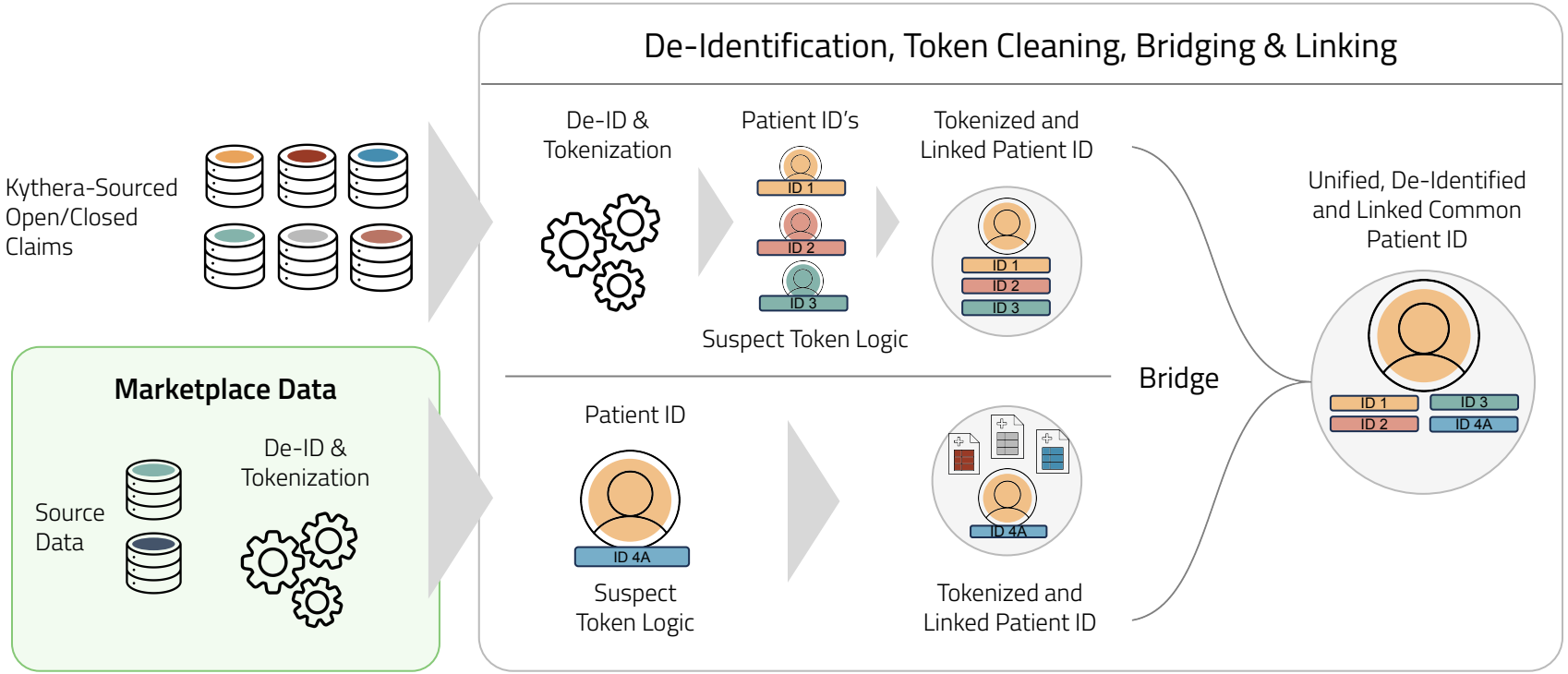
Bridging Method

- First dedupe WITHIN each data set
- Treat separate data sets as one large data set
- Use token matching methods to tie tokens together



©2025 Kythera Labs, Inc. All rights reserved.

Unifying Token Systems



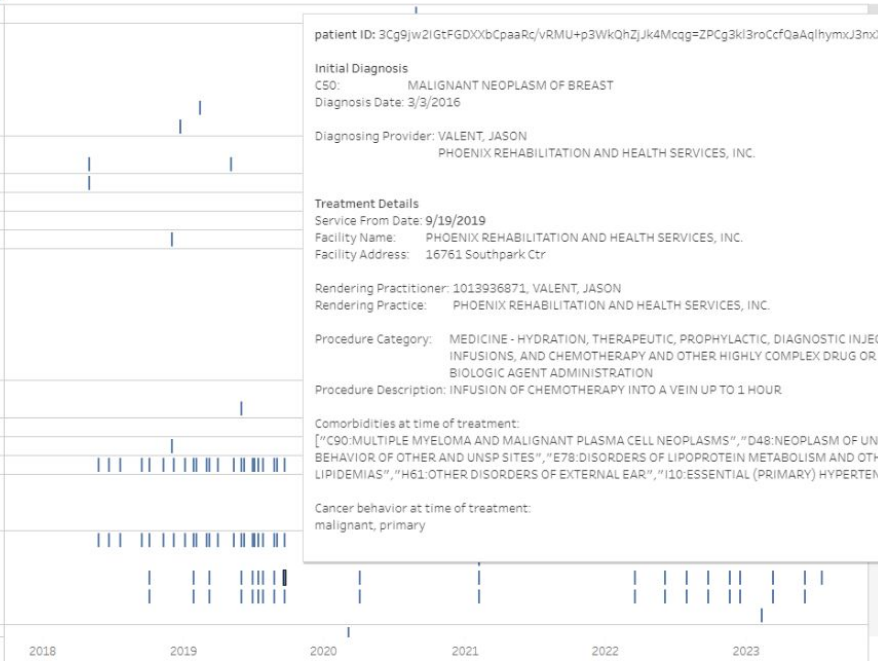
©2025 Kythera Labs, Inc. All rights reserved.

Patient Journey Example: Breast Cancer

Detail and visualize the patient journey

Patient Clinical Pathway

patient_ID	treatment_..	treatment	procedure_proceduregroup	procedure_code
3Cg9jw2lGtFGDXXbCpaa..	DIAGNOSTIC	ADVANCED IMAGING	Null	Null
			RADIOLOGY - DIAGNOSTIC RADIOLOGY	70450
				70486
				71260
				72146
				72148
				72195
		CARDIAC IMAGING	MEDICINE - CARDIOVASCULAR	93010
				93306
			PATHOLOGY AND LABORATORY - ORGAN OR DISEASE-O..	80061
		LOW TECH	Null	G0279
			RADIOLOGY - BONE/JOINT STUDIES	77080
			RADIOLOGY - BREAST MAMMOGRAPHY	77067
			RADIOLOGY - DIAGNOSTIC RADIOLOGY	71045
				71046
				72100
				73110
				73130
				73502
				73564
		ULTRASOUND	MEDICINE - NONINVASIVE VASCULAR DIAGNOSTIC STUDIES	93970
				93971
			RADIOLOGY - DIAGNOSTIC ULTRASOUND	76642
	LOW TECH	OTHER IMAGING SERV..	RADIOLOGY - BREAST MAMMOGRAPHY	77067
	MEDICAL ONCOLOGY	BORTEZOMIB	Null	J9041
		CHEMOTHERAPY	Null	96401
				96413
				96415
			MEDICINE - HYDRATION, THERAPEUTIC, PROPHYLACTIC, DIAGNOSTIC INJECTIONS AND INFUSIONS, AND CHEMOTHERAPY AND OTHER HIGHLY COMPLEX DRUG OR HIGHLY COMPLEX BIOLOGIC AGENT ADMINISTRATION	96401
				96402
				96413
				96415
				96417
		INFUSION	MEDICINE - HYDRATION, THERAPEUTIC, PROPHYLACTIC	96365



patient ID: 3Cg9jw2lGtFGDXXbCpaaRc/vRMU+p3WkQhZJk4Mcqg=ZPCg3kl3roCcfQaAqlhymxJ3nxXWU76FoJGwIizoVnI=

Initial Diagnosis
C50: MALIGNANT NEOPLASM OF BREAST
Diagnosis Date: 3/3/2016

Diagnosing Provider: VALENT, JASON
PHOENIX REHABILITATION AND HEALTH SERVICES, INC.

Treatment Details
Service From Date: 9/19/2019
Facility Name: PHOENIX REHABILITATION AND HEALTH SERVICES, INC.
Facility Address: 16761 Southpark Ctr

Rendering Practitioner: 1013936871, VALENT, JASON
Rendering Practice: PHOENIX REHABILITATION AND HEALTH SERVICES, INC.

Procedure Category: MEDICINE - HYDRATION, THERAPEUTIC, PROPHYLACTIC, DIAGNOSTIC INJECTIONS AND INFUSIONS, AND CHEMOTHERAPY AND OTHER HIGHLY COMPLEX DRUG OR HIGHLY COMPLEX BIOLOGIC AGENT ADMINISTRATION
Procedure Description: INFUSION OF CHEMOTHERAPY INTO A VEIN UP TO 1 HOUR

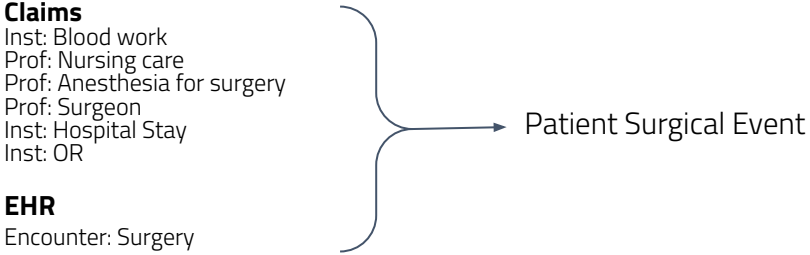
Comorbidities at time of treatment:
["C90:MULTIPLE MYELOMA AND MALIGNANT PLASMA CELL NEOPLASMS","D48:NEOPLASM OF UNCERTAIN BEHAVIOR OF OTHER AND UNSP SITES","E78:DISORDERS OF LIPOPROTEIN METABOLISM AND OTHER LIPIDEMIAS","H61:OTHER DISORDERS OF EXTERNAL EAR","I10:ESSENTIAL (PRIMARY) HYPERTEN

Cancer behavior at time of treatment:
malignant, primary

©2025 Kythera Labs, Inc. All rights reserved.

Understanding Unique Contribution of a Dataset

- "Overlap analysis" typically requested of vendors falls short
  - Prior to patient mastering overlaps are often misleading
  - Patient-only overlaps overestimate the value of a data combination
- Data overlaps best done at the "Patient Event" level
  - Better specificity to use case
  - Record-level data often redundant
  - Records can be of a different type
  - What matters is NEW INFORMATION



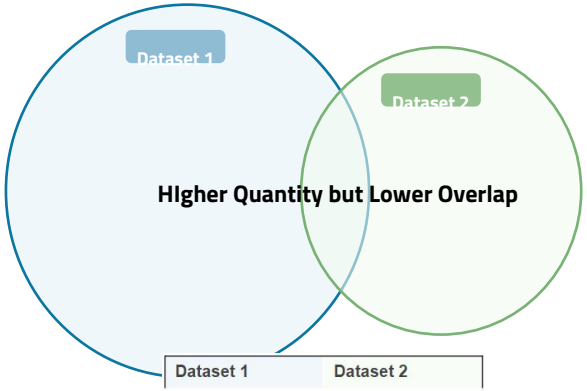
©2025 Kythera Labs, Inc. All rights reserved.

A Note About Overlaps

Before Patient Mastering

Patient A	Token X
	Token Y
	Token Z
Patient B	Token Q
	Token R

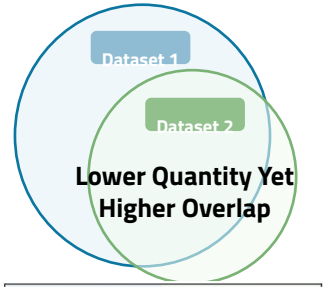
Standard Tokenization



Dataset 1	Dataset 2
Token X	Token Z
Token Y	Token Q
Token R	

(Inflated Patient Count)

After Patient Mastering



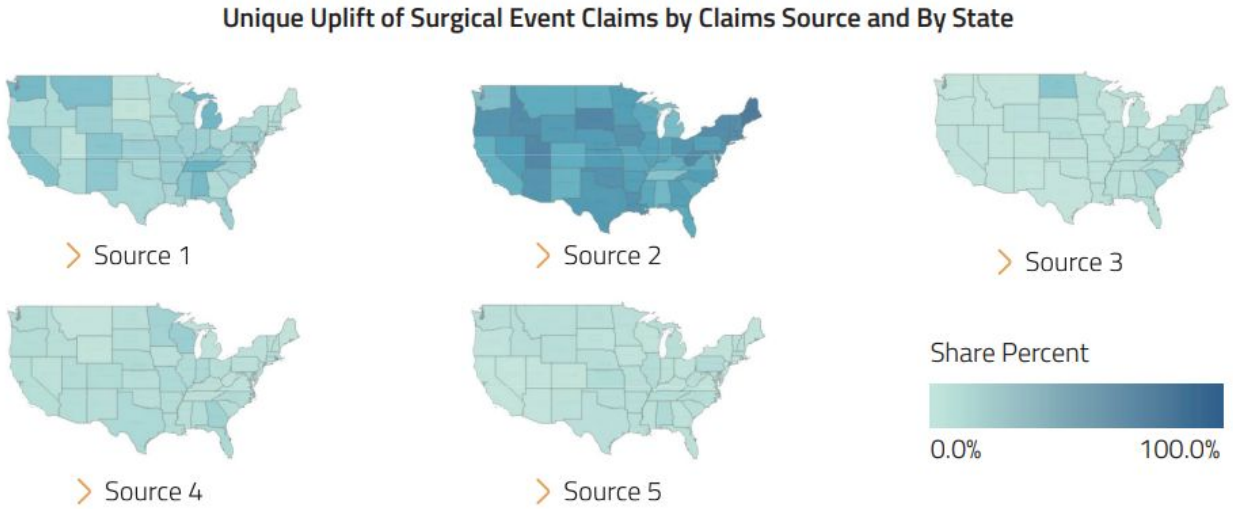
Dataset 1	Dataset 2
Patient A	Patient A
Patient A	Patient B
Patient B	

(More Accurate Patient Count)

©2025 Kythera Labs, Inc. All rights reserved.

Reducing Wasteful Data Spend through Source Assessment Intelligence:

Kythera Customer Case Study



30%

reduction

in data spend was achieved by eliminating redundancy while optimizing their data investments to better support revenue-driving CPT codes.

©2025 Kythera Labs, Inc. All rights reserved.

KYTHERA

Q & A

Contact:
Stuart Head
stu@kytheralabs.com

KYTHERA


Thank You
KYTHERA