

The Utility of Data Tokenization in Clinical Trials

Paul Petraro¹, Carla Heywood²,
Christian Niyonkuru¹, Ling Zhang¹,
Devin Gilliam³, Gordon Cummins²

1-Boehringer Ingelheim Pharmaceuticals Inc., Ridgefield Connecticut USA

2-Science 37, North Carolina USA

3-Datavant, San Francisco USA



Background

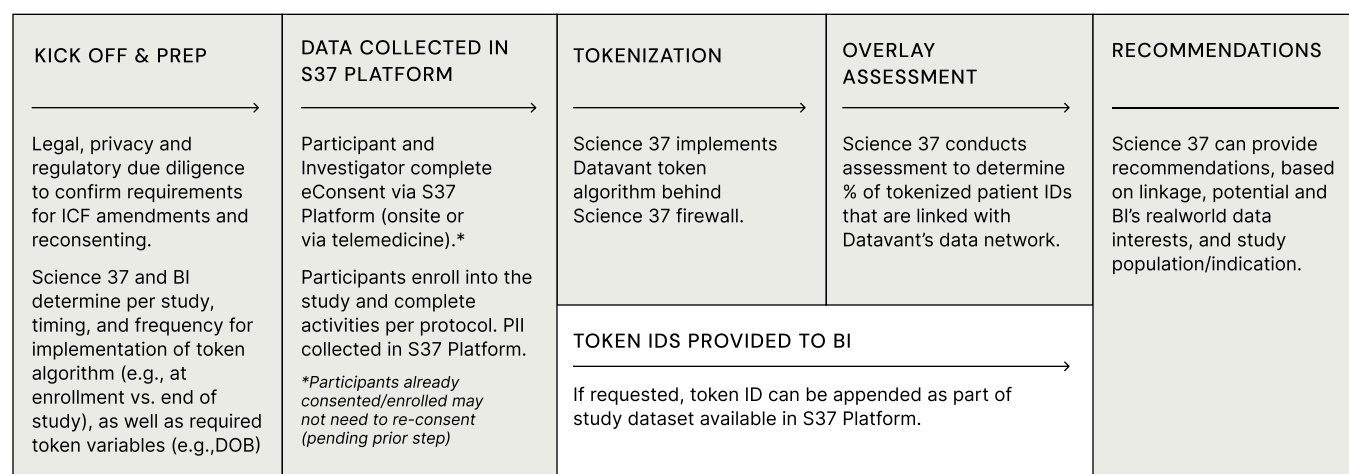
With the growing abundance of real-world data (RWD) sources, and rapidly evolving use of data technologies, it is important to understand the benefits and risks associated with linked RWD to better understand a patient's health status, and for individual and population-level healthcare decision-making.



Objective

Evaluate the utility of anonymized data linking technology using token algorithms to supplement clinical trial data with real world data sources.

Figure 1. Tokenization Process



Methods

Phase II clinical trials have been selected to utilize data tokenization via a token algorithm software to assess the available data sources to be linked to existing patient information collected in the trials.

Most token algorithms utilize complete personal identifiable information (PII) including first/last name, date of birth, gender, and address (including zip code).

Local legal and data privacy considerations must be identified and carefully evaluated.

By implementing a token identification (ID) algorithm, PII is transformed into an unrecognizable token ID that is not linkable to the patient.

As global regulations vary around the collection of PII, only the CNS trial (US only) contained all required PII elements while the CD trial only contained the patient's email address.

Key to the analysis will be determining the number and type of PII required to maximize the utility of a token algorithm.



Results

Of the 136 patients currently enrolled in the CNS trial to date, all required PII data elements have been available and utilized.

To date, the CD clinical trial has enrolled 33 patients with the availability of an email address for all participants.

These results open various potential real-world data types to be evaluated for efficient linkage to the core clinical dataset. This provides additional insights to be considered as part of the overall real-world evidence strategy.

Table 1. CNS Overlay Assessment (Prognos Claims Datasets)






DATASET	DATA TYPES	OVERLAPPING INDIVIDUALS
<p>Prognos All Data Types 2016–2021</p> <ul style="list-style-type: none">• >200M patients• Deliveries: real time, daily, weekly, monthly, quarterly, yearly, one time)• <1 week latency• >5 years of history <p>Prognos® Health is accelerating the discovery and application of real-world data to improve health through the Prognos Marketplace – the largest collection of integrated medical records on 325 million de-identified U.S. patients. The Marketplace is built on Prognos Factor®, a specialized healthcare analytics platform that leverages a patent-pending database management system enabling no-code exploration of hundreds of billions of medical records at interactive speeds. The Marketplace allows healthcare clients to assess the value of data before purchasing and buying only the data needed. Embedded standardization and linkability make the data analytics-ready, accelerating speed to value. Use cases include targeting specific patient/provider populations, commercial and HEOR process optimization, clinical research studies, and medical underwriting risk assessment.</p>	<ul style="list-style-type: none"> Medical Claims Retail Rx Claims Lab Specialty Rx Claims	 103 (83.06%)

Table 2. CNS Overlay Assessment (KYTHERA Labs Claims Datasets)









DATASET	DATA TYPES	OVERLAPPING INDIVIDUALS
<p>KYTHERA Labs</p> <ul style="list-style-type: none">• >200M patients• Deliveries: real time, daily, weekly, monthly, quarterly, yearly, one time)• <1 week latency• >5 years of history <p>Kythera Labs provides a health data and analytics platform that applies machine learning to remaster and improve the quality of information by looking for signals which can predict behavioral patterns among patients, practitioners, provider systems, and payers. We supplement our fast, cost-efficient platform with data representing over 300 million US patients plus include directories of practitioners, provider systems, and payers. Our team has been solving problems for decades using proven data science concepts and unique technologies to enable decision-making and growth for healthcare organizations. We focus on being humble and curious in all our work and understand that trust is at the heart of innovation. Get your gears turning with Kythera Labs.</p>	<ul style="list-style-type: none"> Consumer Medical Claims Medical Claims Remittance Other Lifestyle Retail Rx Claims Specialty Rx Claims Electronic Medical Records	<p>Electronic Health Records</p> <p>52 (49.14%)</p> <p>Prescription Claims</p> <p>87 (70.16%)</p> <p>Medical Claims Submitted</p> <p>103 (83.06%)</p> <p>Electronic Health Records</p> <p>91 (73.39%)</p>



Table 3. Combined Dataset Analysis – Illustrative Overview


BI has the ability to identify patients across separate datasets & combine them for a complete view of BI's trial participants

	INDIVIDUALS					TOTAL
Phase II Data	A		C		E	3
Dataset 1	A	B	C	D		
Overlaps (Phase II x Dataset 1)	X		X			2
Dataset 2		B		D	E	
Overlaps (Phase II x Dataset 2)					X	1
Phase II x Combined Datasets	X		X		X	3

Illustrative Overview

- BI's Phase II dataset consists of 3 individuals (A, C, E)
- Dataset 1 has 2 overlapping individuals in BI's Phase II dataset (A, C); Dataset 2 has 1 overlapping individual (E)
- Combining the datasets gives full coverage of BI's Phase II participants

Table 4. Combined Claims Datasets (Prognos & KYTHERA Labs)

DATASET	DATA TYPES	OVERLAPPING INDIVIDUALS	OVERLAPPING INDIVIDUALS Combined datasets
Prognos All Data Types 2016–2021 <ul style="list-style-type: none"> • >200M patients • Deliveries: real time, daily, weekly, monthly, quarterly, yearly, one time) • <1 week latency • >5 years of history 	<div>Medical Claims</div> <div>Retail Rx Claims</div> <div>Lab</div> <div>Specialty Rx Claims</div>	<div>53</div> <div>(42.74%)</div>	<div></div> <div>106</div> <div>(85.48%)</div>
KYTHERA Labs <ul style="list-style-type: none"> • >200M patients • Deliveries: real time, daily, weekly, monthly, quarterly, yearly, one time) • <1 week latency • >5 years of history 	<div>Electronic Health Records</div> <div>Prescription Claims</div> <div>Medical Claims Submitted</div> <div>Medical Claims Remittance</div>	<div>52 (49.14%)</div> <div>87 (70.16%)</div> <div>103 (83.06%)</div> <div>91 (73.39%)</div>	

Conclusion

The ability to tokenize trial data in the development lifecycle allows for earlier access to real world data on specific patient populations, enabling the potential to validate real world populations prior to marketing authorization. This approach also provides the ability to link additional data directly into the clinical trial programs.

Disclosures/disclaimer

This study was supported by the Boehringer Ingelheim. The authors were fully responsible for all content and editorial decisions, were involved at all stages of poster development, and have approved the final version.

Petraro P, Heywood C, Niyonkuru C, Zhang L, Gilliam D, Cummins G; The Utility of Data Tokenization in Clinical Trials. Poster presented at ICPE 2022, Aug. 24-28, Copenhagen, Denmark.