The Key to Clinical Trial Success: Patient Tokenization

Real-world data and patient tokenization can streamline clinical trials and ensure successful results by employing connected, robust, and reliable patient population insights.

Kwame Marfo
Market Strategy Lead,
Clinical Development
Komodo Health

Ivy Weng
Head of Clinical Development
and RWE/HEOR
Komodo Health

Ryan Moog
Clinical Trial Solutions Lead
Datavant

INTRODUCTION

Patient healthcare data collected during clinical trials provide a limited view of a participant's medical life. This data collected before, during, and after clinical trials are often compiled by different parties and are challenging to bring together from disparate, siloed sources while maintaining patient privacy. This ultimately renders the data difficult to use for meaningful post-trial insights. In this paper, an overview is provided of how clinical trial data can be tokenized and connected to valuable, fit-for-purpose, real-world data while preserving privacy. Cases are presented in which recent advancements in data connectivity and analytics allow real-world data to be collected before, during, and after clinical trials. As this process streamlines the drug development lifecycle, successful results in clinical practice are ensured.

DATA SILOS VERSUS DATA CONNECTIONS

There is inherent value in the ability to merge clinical trial data with mortality, claims, electronic health records [EH], genomics, and other data. Historically, clinical trials were data silos which collected specific, limited information (**FIGURE 1**). Merging trial data with data from other sources required a patient identifier to share individual patient data and connect it across data silos. In the last few decades, probabilistic matching has increased the likelihood that, for example, a 33-year-old male from one data set and a 33-year-old male from another data set can be identified as the same patient based on available information. However, achieving regulatory-grade data that generates good evidence requires high-precision matching across data sources while maintaining patient privacy.

TOKEN GENERATION AND LINKAGE

The first step in this process is to generate a patient token or privacy-preserving record locator, which is an alphanumeric code for an individual patient that does not, in any way, indicate the identity of that patient from each data source (**FIGURE 2**). These tokens are encrypted and hashed as one-way, randomized codes that enable linkage with other data sets that have also been tokenized. This can be done directly with clinical providers,

contract research organizations [CROs], or even clinical study sites, depending on where the patient's personally identifying information [PII] is located. The token itself then becomes the patient-linkable component inside the data set. Because the patient-specific tokens are created using the underlying patient identifiers before they are removed, the tokenization process enables the creation of a consistent token from any dataset. Every code is unique, and at no point is the same token generated across data sets for any patient.

Using these tokens, data are then linked across datasets using proprietary software which recognizes various tokens as the same patient using high-accuracy, probabilistic matching [**FIGURE 3**].

FIGURE 1: Historically, clinical trials were data silos collecting specific information.

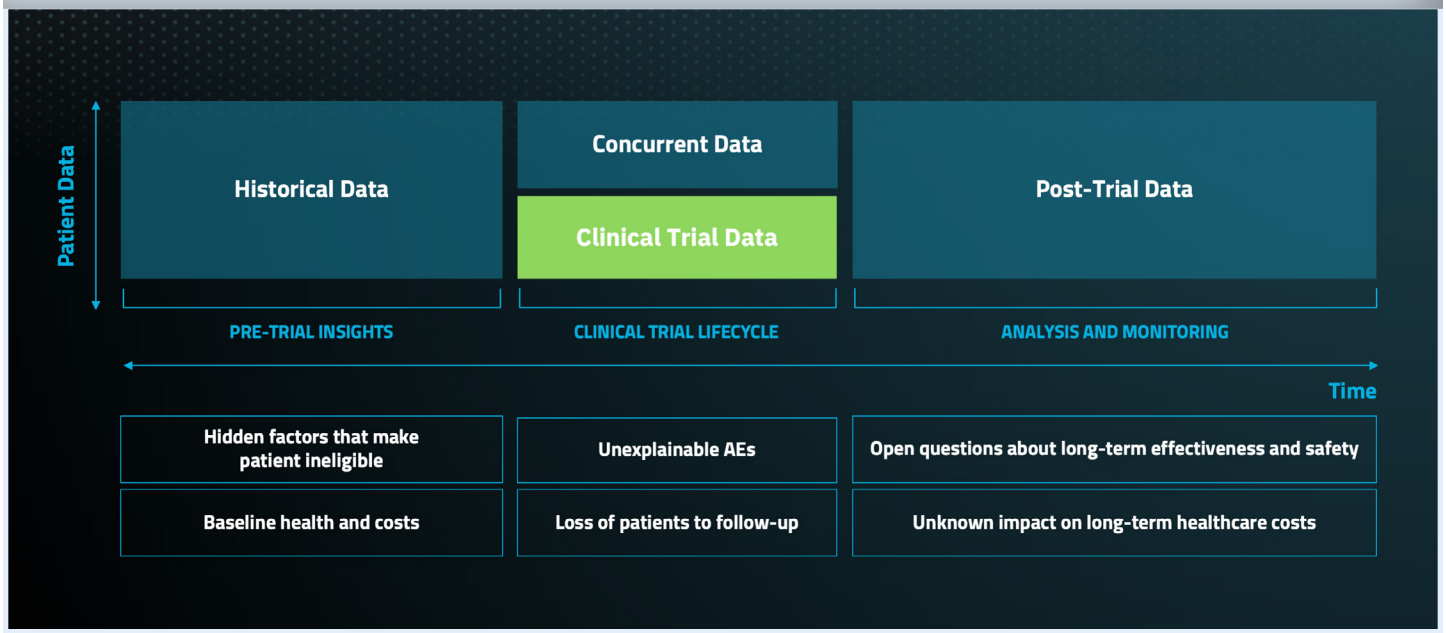


FIGURE 2: Patient tokenization.

A process that supports de-identification by removing patient identifiers and generating a patient-specific encrypted token.

Patient-specific tokens are created using the underlying patient identifiers before they are removed, the tokenization process enables the creation of a consistent token from any dataset.

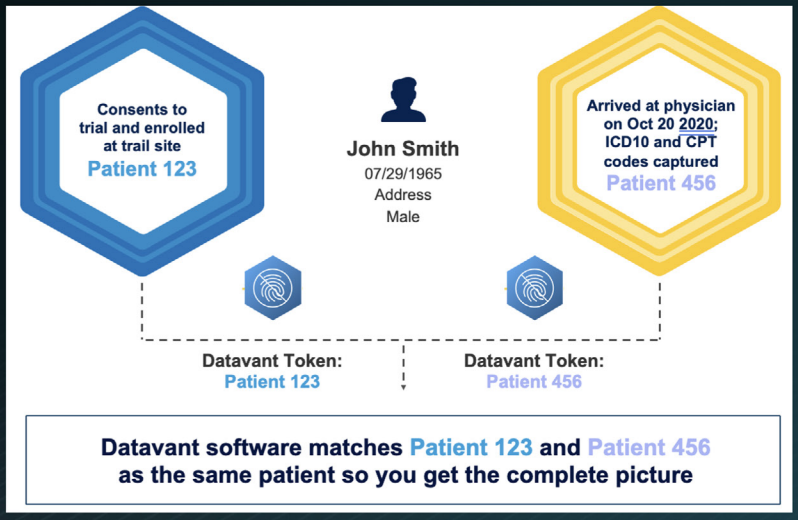
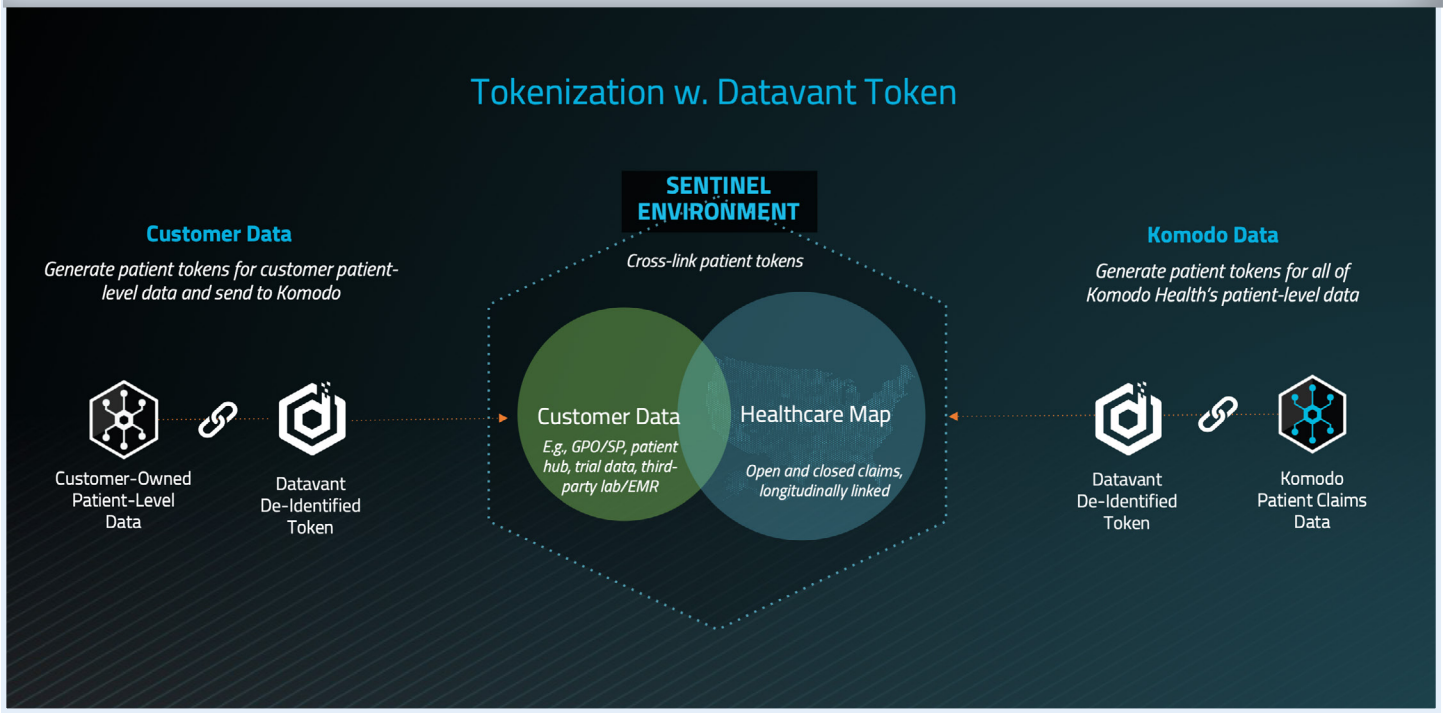


FIGURE 3: Patient-level linkage for real-world insights.



Real-world data from a trial participant's previous medical history can be linked with electronic case report forms [eCRFs], electronic patient reported outcomes [ePROs], insurance claims, ongoing clinical encounters, pharmacy orders, etc. Active engagement across the healthcare system outside of the trial can then be tracked without compromising the participant's privacy.

Ultimately, Datavant acts as a neutral, ubiquitous tokenization partner to multiple data providers and generators, while Komodo Health brings best-in-class data sources together, including administrative claims, EHRs, laboratory results, specialty pharmacy, and genetics. Together, these technologies connect de-identified patient data, bringing together multiple sources of real-world and clinical trial data and generating a longitudinal dataset which details the patient journey. This has resulted in Komodo's comprehensive Healthcare Map™ of the United States [US], which tracks the longitudinal journeys of clinical encounters for 330 million US residents with self-identified race and ethnicity demographics.

REAL-WORLD DATA LINKAGE

Patient-level data on the Datavant token can be easily linked to yield real-world insights. A patient overlap feasibility analysis is first run by leveraging tokens representing data from multiple data sources, such as Veradigm, Quest, TriNetX, and Prognos, to identify common and unique patients between datasets. Necessary data sets are identified as required fit-for-purpose for a specific trial. The Datavant software generates tokens for clinical trial participants from PII such as first and last name, gender, date of birth, or social security number, and connects the clinical trial participants who are tokenized to the various data sources. Komodo brings in claims data and other clinical records relevant to the given clinical trial to create a customized "data universe," which is a unique and powerful linked dataset specific to that trial. Retrospective historical data can be tracked for each patient, and data is refreshed daily to capture patient data even after the clinical trial is complete.

This universe also allows the customer to look both backward and forward within the broader target patient population to understand the full health journeys and outcomes using first-party data as well as other sources. Compliance with the US Health Insurance Portability

and Accountability Act of 1996 [HIPAA] in this regard is certified by a third-party statistical expert, with data either aggregated to a sufficient level or with certain features removed to eliminate the possibility of re-identification. Informed consent and Institutional Review Board [IRB] approval are not required when real-world data are used in this manner since the sources have been de-identified by tokenization.

REAL-WORLD APPLICATIONS

This HIPAA-compliant collection of patient-level data powers six key use cases (**FIGURE 4**). Real-world data linked to clinical trial data can be used to optimize protocols, study design and site selection, understand social determinants of health (including where patients are underserved in terms of access to clinical trials), build external control arms for rare diseases or personalized medicine, understand which patients respond better to certain trials, perform long-term follow-up for safety and efficacy, and ensure unmet needs are served.

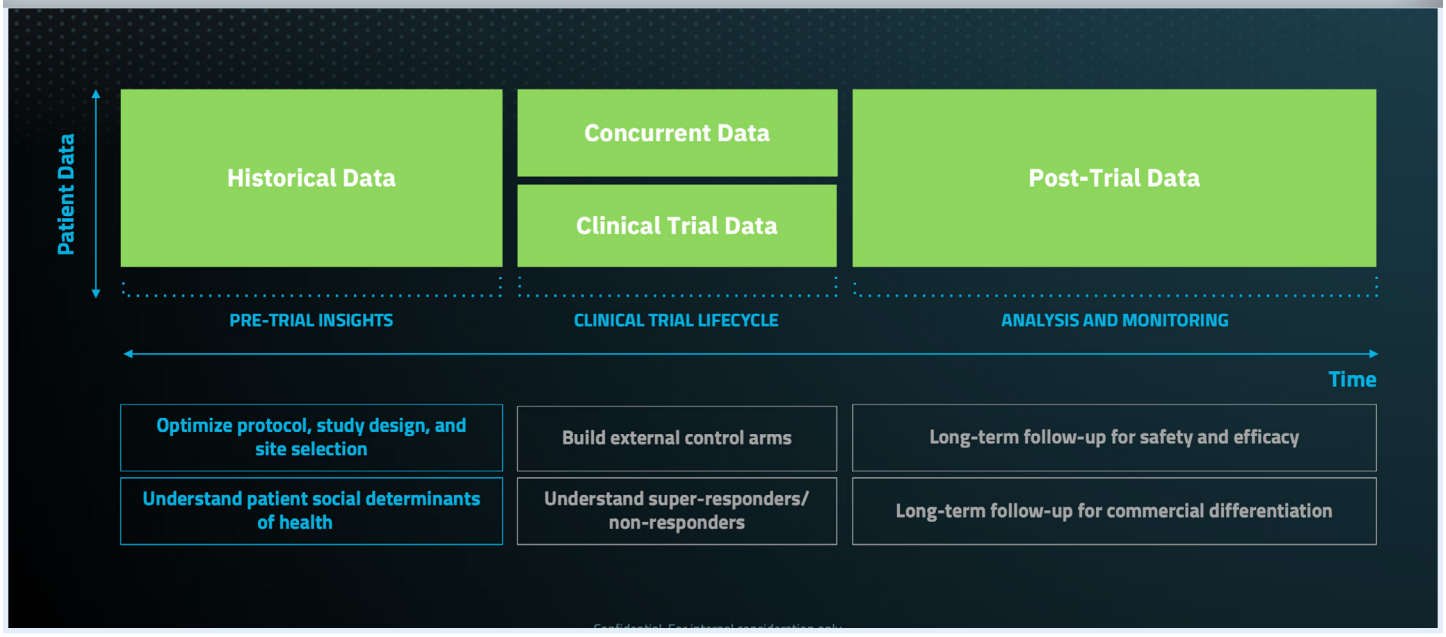
Case Study: Optimize Study Design and Cohort Selection

Although linking patient data across multiple sources can be immensely valuable during a clinical trial, the utility of tokenized data is not limited to that specific clinical study. Sponsors are encouraged to tokenize all trial data as those data can provide great insight when looking back at legacy trial data to evaluate patient recruitment, adherence, follow-up management, or even additional patient information that was not considered during trial design. For example, one medical device company combined archived eCRF data with claims and EHR data from two primary sources. Their tokenized trial data allowed them to track previously enrolled patients longitudinally, evaluate the effects of comorbidities on long-term efficacy, and inform cohort design for a follow-on, 750-patient, implantable medical device trial.

Case Study: Reduce Burden of Long-term Follow-up

The clinical collection of long-term, follow-up data can represent a large burden for both study sites and patients, particularly for large trials. In addition, patients may have unplanned clinical encounters between study visits that are important to track but may not always be captured within the site's health system, for example, due to travel. Tokenization of clinical trial data enables automated, long-term follow-up of patients. For example, during a 10,000+ patient, long-term atherosclerotic cardiovascular disease clinical trial, one academic research institute linked their trial data to data from two claims and EHR aggregators. This allowed longitudinal data access to subsequent clinical encounters and mortality events and automated patient follow-up for more than five years, which enabled the institute to correlate ePROs with study outcomes.

FIGURE 4: Enabling richer insights and a better understanding of all patients.



Real-world data can also be used to observe how an endpoint endures and compares over time, from short-term results to long-term follow-up, without requiring a new clinical study. For example, one sponsor had successfully completed a 31-week, Phase IV open-label study examining seizure reduction following 24 weeks of treatment with a new drug. These clinical trial data were tokenized and linked to Komodo's Healthcare Map to extend the study outcome's measurement to six months. This link to the Healthcare Map also enabled monitoring of long-term safety and efficacy for up to 15 years using real-world data from a larger, more varied population of patients than those enrolled in the original Phase IV study.

Case Study: Build External Control Arms

Many clinical trials for rare diseases are perforce conducted with few patients, leading either to failed trials (due to issues like insufficient statistical power) or to single-arm trials that make it difficult to compare results against other therapeutic options. Similarly, in oncology, the move towards personalized medicine means certain therapeutics will be most efficacious for a well-determined but potentially small patient population with frequently poor prognosis. It is both ethically and operationally useful to consider an external control arm for these types of studies, which not only reduces the number of patients who need to be enrolled but can also accelerate the timeline for therapies being brought to market.

Tokenized, real-world data represent an ideal source for these external controls. For example, in a non-small cell lung cancer trial, one sponsor needed patients with a specific subtype of epidermal growth factor receptor [EGFR] gene mutation for their registration trial, which required them to use three different EHR data vendors to obtain enough patients for an external control cohort. The US Food and Drug Administration [FDA] agreed to accept these real-world data, with the condition that the trial sponsor could demonstrate the final control population had no duplicates. Tokenizing the data sets from all three vendors enabled the comparison of patients and removal of duplicates. The use of this tokenized, external control arm cut the trial population in half, providing significant time and cost savings.

CONCLUSIONS

Recent advances in tokenization, data connectivity, and advanced analytics allow patient data collected before, during, and after clinical trials to be accurately linked to fit-for-purpose, real-world data in a way that ensures patient privacy. This enables sponsors to gain a deeper and broader understanding of patients, reducing uncertainty across the drug development lifecycle and increasing focus on what treatments work in specific populations in real clinical practice. This, in turn, can accelerate clinical development timelines and optimize clinical trials to ensure success, ultimately leading to a direct and lasting impact for patients.