Accurate Patient Record Matching

Patient record matching on de-identified data is a critical component of healthcare research and analytics, preserving patient privacy while allowing for multiple datasets to be connected for a holistic view of patients.

Datavant Match combines a proprietary best-in-class algorithm for matching records with a turnkey experience, to ensure you can accurately connect data and uncover deeper patient-level insights.

Our Approach

- **Sophisticated**
  
  Enhanced matching capabilities are powered by a machine learning model trained on six billion record pairs, representing 300 million patients across consumer and EHR datasets.

- **Turnkey**
  
  Append a single, easy-to-use patient ID for matching records without having to build custom token-based match logic.

- **Secure**
  
  There's no need to share PHI, and records are processed in our secure cloud environment. Your Datavant IDs are encrypted with an encryption key specific to your organization.

In short, Match makes it simple for you to optimize your matching strategy, without sacrificing accuracy.

Match does not require you to send any PHI, and instead uses de-identified records with Datavant tokens. Given de-identified data is far more accessible and commercially available for healthcare analytics than identified data, with Match you can connect your data to a wide variety of real-world datasets in the Datavant ecosystem, including claims, lab data, EHR, and consumer data -- all without having to expose PHI.

DVIDs are encrypted to be data-pool specific, meaning they are specific to your dataset and any additional reference data you specify. This prevents your match results from being affected by external data that may not have been cleaned and standardized, and ensures consistent ID assignment.
Validated with Real-World Data

Datavant’s matching solutions can be used to achieve over 99.5% precision with full patient information. With basic patient information, studies conducted by third-party institutions have consistently shown over 99% precision and over 85% recall.**

- **Global Analytics Company:** This study was based on an identified consumer dataset used to evaluate de-identified data matching via PPRL. Analysis of 150 million records derived from SSN, date of birth, and gender yielded 99.8% precision with slightly lower recall.

- **Academic Research Center:** This study was based on a clinical data warehouse in one of the largest, most diverse metropolitan areas in the United States. Analysis of only Token 4, derived from first name, last name, date of birth, and gender yielded 99.9% precision.

- **Centers for Disease Control & Prevention:** These studies, based on the National Hospital Care Survey (NHCS) & CDC National Death Index, evaluated the performance of PPRL technology in linking two datasets as compared to “gold standard” linkages, performed using standard linkage algorithms used by the CDC. Preliminary results yielded up to 98.7% recall.

How Datavant Match Works

Datavant Match determines whether a pair of records should be considered a match based on (1) comparing tokens across datasets and (2) a machine learning model trained on consumer and EHR datasets representative of the U.S. population. These datasets contain known patient matches that the model can use to inform whether two records are a match.

What are tokens?

A token is a secure, encrypted hash generated from the underlying PII elements in a dataset. Datavant’s technology can generate many possible tokens based on the different combinations of PII present.

Note that match rates precision and recall are ultimately constrained by the underlying dataset which is why results vary depending on the available data.
Even with just 6 different tokens on a record, there are more than 700 different ways that two records can match. Each of the 6 tokens could match, mismatch, or be incomplete across the two records. Datavant has a menu of up to 40 tokens based on different combinations of identifying elements, which in theory offers more than a quintillion possibilities for how two records could match.

How are matches declared?

Match will compare the record with all records previously processed within a data pool, declaring each record pair to be a match or not a match. If no match is declared, a new Datavant ID is created for this individual and assigned to the record. If a match is declared, the existing Datavant ID is assigned.

The model makes a determination by having been trained on billions of tokenized records for which a reliable person ID was present. This allows the model to learn about the patterns of token concordance (and discordance) across matched records and to calculate statistical likelihoods for matching.

For example, the model may have learned that if Tokens 1, 2, and 4 match but Token 17 does not, it is 99.9% likely to be a match. Tokens 1, 2, and 4 rely on first name, last name, date of birth, and gender while Token 17 relies on first name, last name, street address and zip 5. Therefore, there is high empirical evidence that if the patient’s street address or zip 5 have changed but name, date of birth and gender have not, those records are, in fact, the same person.

What type of statistical model is deployed?

To calculate statistical likelihood of matching, the algorithm employs an ensemble of gradient boosted trees, a machine learning model that employs successive decision trees, each of which learns from the previous set.

When evaluating a pair of records, the model initially outputs a score between 0 and 1 indicating the probability that the pair of records correspond to the same patient. Two records are declared a match if the score exceeds a threshold. The threshold is determined by comparing precision and recall for record pairs in the model’s training data, and picking the threshold that best balances precision and recall.

While results vary depending on the dataset and input data, precision and recall can exceed 99% with complete patient information (i.e., first name, last name, gender, date of birth, zip, SSN).

The model is updated every quarter with refreshed training data, and Datavant is constantly adding more.

Getting Started with Match

Match appends Datavant IDs (DVIDs) as patient-level identifiers to datasets distributed through Datavant.

1. Tokenize and de-identify your data.
2. Upload to the Datavant portal.
3. Notify your Product Success Lead about intended recipients.
4. Your data will be matched and delivered to you through Datavant’s SFTP with corresponding Datavant IDs appended.
If matching records across multiple parties:
1. You upload your data to the Datavant portal.
2. Your data partner uploads their data to the Datavant portal.
3. Distribute both datasets to one or both of the two parties who will receive the data with Datavant IDs appended.

1. Data is de-identified & tokenized
2. Upload to Datavant
   Datavant Match assigns IDs to records belonging to same patient
3. Datasets are distributed back to you with Datavant IDs appended

1. Data is de-identified & tokenized
2. Both parties upload to Datavant
   Datavant Match assigns IDs to records belonging to same patient
3. Both datasets are distributed back to you and/or your partner with Datavant IDs appended