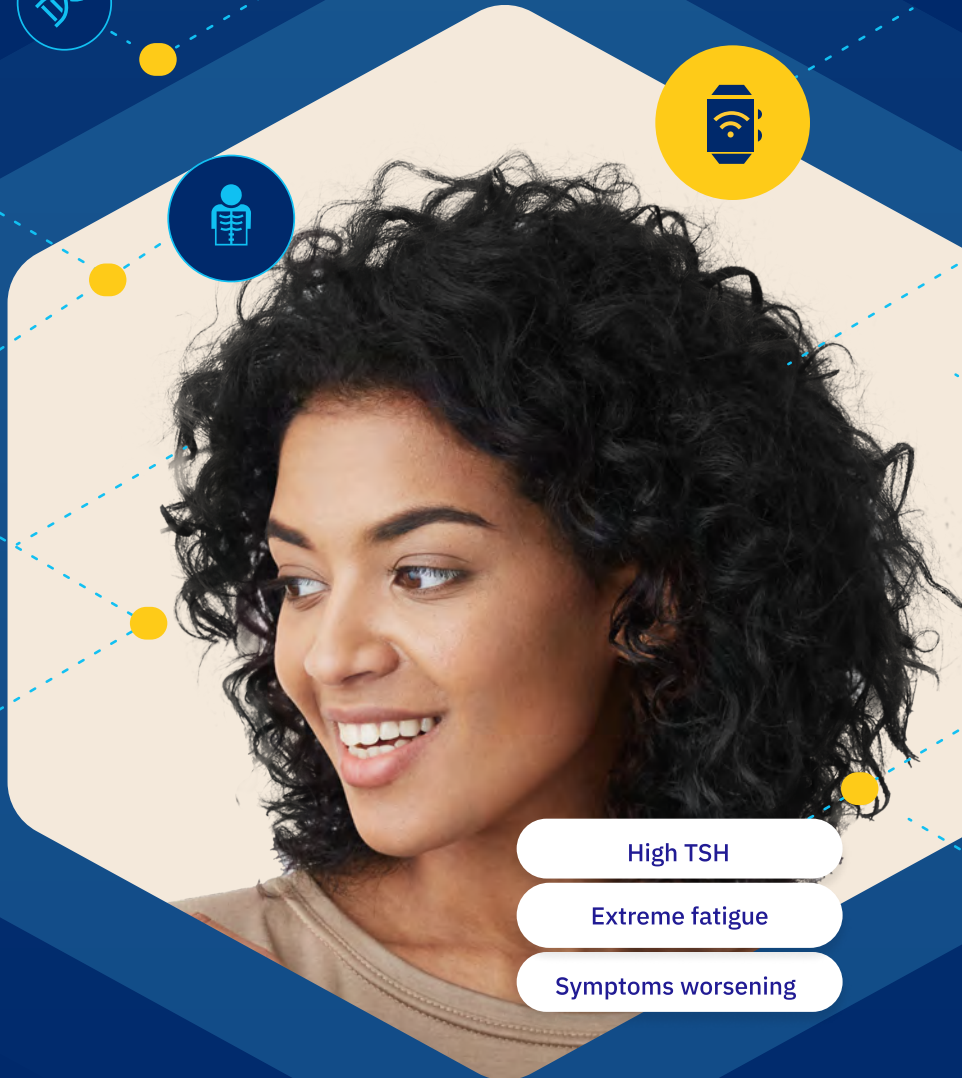


# Matching Patients Across Healthcare Datasets

A New Research-Grade Standard for  
Privacy-Preserving Record Linkage



High TSH

Extreme fatigue

Symptoms worsening

## Executive Summary

Modern-day privacy-preserving record linkage (PPRL) presents an opportunity to set a new standard for research-grade patient matching. High-precision matching is the key to unlocking research, insights and innovation for regulatory applications.

Strong research-grade matching solutions must:

- De-identify and match patients with high accuracy
- Minimize risk of re-identification
- Provide transparency and audit trails for regulators

Many healthcare organizations have struggled with implementing consistently accurate and reliable patient matching. Some have seen a small measure of success with simple schemes, while others continue to grapple with high rates of error and lack of scalability.

The consequences of inadequate matching are significant. For example, physicians cannot easily gather information on patients with chronic conditions, resulting in duplicate tests, missed diagnoses and delayed care. Older adults, who often see more than 10 different care providers a year at multiple locations, fall through the cracks. An allergic reaction listed on one record but not another raises the potential for patient harm when those records are not correctly matched and linked.

This white paper addresses the need for highly accurate PPRL and patient matching solutions using de-identified data – and how this opens up a new era of research opportunities.

### **Datavant Match is the industry leader for highly accurate, research-grade matching**

- Higher accuracy than simpler approaches
- Industry-leading performance: 99.4% precision, 95.1% recall
- Secure, transparent, auditable
- Machine learning model trained on billions of records with matches that are validated against rich referential data of patient records with known true matches
- The largest healthcare ecosystem for secure, research-grade matching and data connectivity and exchange
- Simple, secure and consistent patient identifier (DVID) for easier, more accurate matching

## The challenges of patient matching with de-identified data

Imagine patient Abigail Lane. Abigail has suffered from chronic hypothyroidism her entire adult life, experiences extreme fatigue and was prescribed a hypothyroid medication five years ago.

As a result, Abigail's health data resides across a variety of databases:






1. Patient-reported fatigue levels
2. Clinical electronic health record (EHR) data from visits to endocrinologist
3. Lab tests measuring thyroid stimulating hormone (TSH) levels over time
4. Pharmacy data tracking medication adherence
5. Insurance claims data on number, location, cost of visits

### Meet Abigail Lane

Health information gets fragmented as she goes through the healthcare system. This can compromise the quality care.



Abigail also got married, changed her last name, and moved out of state. She goes by “Abby” on occasion and sometimes uses a shared email address when registering for a wearable device. Abigail’s social security number is filled out in adjudicated insurance claims forms but is missing or incomplete in various other instances (e.g., filled in as 000-00-0000 or only last four digits). Therefore, each database has a different level of identifying information, also known as protected health information (PHI) and patient identifying information (PII).

	First Name	Last Name	Sex	Age	Email	Zip code	Social Security Number (SSN)
 Clinical EHR Endocrinologist	Abigail	Lane	F	33	abigail@aol.com	11122	111-22-3333
 Patient-reported Fatigue Levels	Abby	Lane	F	33	----	11122	NULL
 Lab Tests TSH levels	Abby	Lane	F	33	----	33444	000-00-0000
 Claims Care costs	Abigail	Jones	F	33	paul@aol.com	33444	111-22-3333
 Pharmacy Adherence	Abby	Jones	F	34	----	33444	111-22-3333

Today, given these common discrepancies in Abigail’s identifying information, it is extremely difficult to match her profile across diverse sets of care locations with a high degree of accuracy and utility.

Many current approaches to matching struggle with managing these data discrepancies, resulting in low match rates that disqualify the data for use in research. Another challenge is ensuring patient privacy. For example, while a strong match result can be generated by passing through higher levels of identifying information, this can also increase the likelihood of re-identification.

For many commercial matching applications, there is often little to no visibility into the logic behind how matches are determined. This lack of transparency severely hinders customers’ ability to assess the accuracy of matching methods and use these solutions for real-world evidence (RWE) programs designed for regulatory submission.

## Matching is difficult, even when using fully identified information

A 2012 survey of CIOs at health systems found that even when using identified master patient indexes based on available patient identifiers, nearly one in five hospital systems reported that patient mismatches (false positives) led to at least one adverse event.

**Source:** *Summary of CHIME Survey on Patient Data-Matching, College of Healthcare Information Management Executives, May 16, 2012:*

[https://chimecentral.org/wp-content/uploads/2014/11/Summary of CHIME Survey on Patient Data.pdf](https://chimecentral.org/wp-content/uploads/2014/11/Summary_of_CHIME_Survey_on_Patient_Data.pdf)

## Introducing Datavant Match

Datavant Match is the industry's leading validated patient matching solution to consistently achieve a high rate of precision and recall for research-grade enterprise matching and data connectivity at scale.

Datavant Match has demonstrated a 99.4% rate of precision and 95.1% recall in repeated performance tests. This is based on internal studies and tests being run on billions of records in consumer and EHR datasets with high data quality.

Patient privacy, data quality, security and transparency are at the core of all Datavant products. Privacy experts and technology run ongoing checks to minimize risk of re-identification throughout the process. When underlying pass through identifying information is needed for matching, privacy experts confirm the use of these fields. This lets researchers and healthcare providers connect data with confidence for the study of longitudinal patient journeys, outcomes and cost-effectiveness, improvement of patient care and treatment, and regulatory-grade, real-world evidence generation programs.

### Precision → 99.4%

Probability that two records predicted to be a match in fact correspond to the same person. Higher precision minimizes false positives.

		ACTUAL	
		Patient records match	Patient records don't match
PREDICTED	Patient records match	994	6
	Patient records don't match	False Positive	True Negative

### Recall → 95.1%

Likelihood that two records belonging to the same person will be captured as a true match. High recall minimizes false negatives.

		ACTUAL	
		Patient records match	Patient records don't match
PREDICTED	Patient records match	994	False Positive
	Patient records don't match	51	True Negative

Using Datavant Match, the discrepancies in Abigail's underlying identifying information, including differing rates of capture and changes in name, geography and email, are accounted for and she is resolved as a single patient across these databases with high accuracy. Datavant's independent privacy experts and automated remediation technologies minimize the risk of re-identification, in accordance with HIPAA regulations and expert determination standards, throughout the process of patient matching and identity resolution, record linkage and data distribution.

## Getting to high accuracy matching

Datavant Match uses a transparent, privacy-first approach built on robust data quality procedures and an advanced machine learning model trained on billions of patient records within Datavant’s healthcare ecosystem, the largest in the industry.

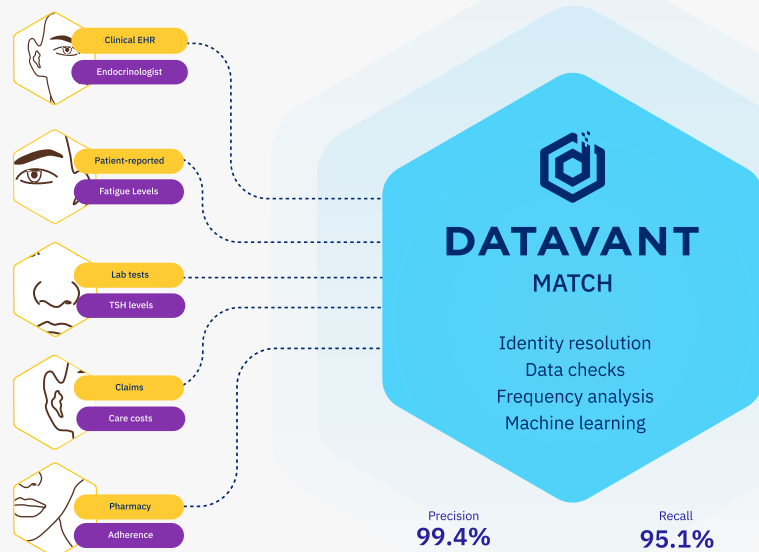
Let’s go back to hypothyroid patient Abigail Lane, who used a nickname and a different zip code at different points of care.

Matching solutions that fail to take nicknames, variations or missing data into account will generate significantly lower matching results – missing true positive patient matches and resulting in a higher rate of false positives, or incorrectly matching two different patients together.

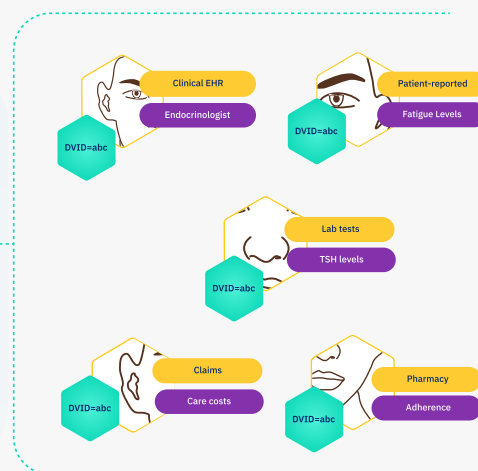
To avoid this, patient matching solutions require high impact data quality and improvement processes during de-identification. This is a critical step to preserving patient privacy and optimizing for high precision rates. These standards allow Datavant to obtain the highest fidelity match between records.

### How it works

#### De-identified records



#### Records matched across datasets with same DVID



### What is the DVID

As part of Datavant’s matching process, a consistent and locally encrypted patient-level Datavant ID (DVID) is assigned to each matched patient record. These DVIDs are automatically appended to datasets for further linkage and distribution – and are the strongest privacy preserving IDs available today for patient matching and record linkage.


For a patient who moves to a different zip code, goes by a nickname, or utilizes multiple email addresses, Datavant Match resolves these into the same patient profile, accurately consolidating what would have otherwise looked like multiple patients into one.



## The advantages of Datavant Match

Traditional methods to address complexities and data discrepancies are labor-intensive, static and result in less than optimal match rates. Limitations include the inability to detect changes in identifying information that negatively impact match rates, and the inability to accommodate the rapid growth in digital patient data.

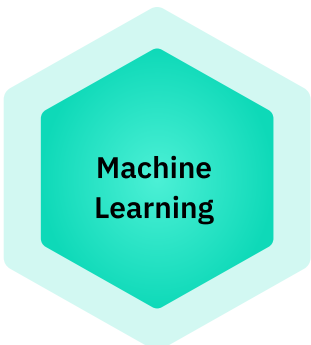
Datavant Match is the industry-leading solution for PPRL, with a 99.4% rate of precision in repeated performance tests, on high quality data. Datavant Match uses a multi-faceted approach to determine a high-accuracy match.



### Underlying data cleaning

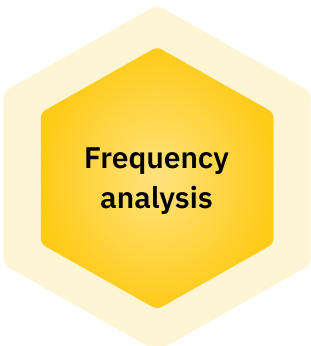
#### Datavant's data quality steps include:

- Removing filler values
- Removing invalid social security numbers (SSN), date of birth and email values
- Standardizing and normalizing of misspellings, nicknames, different addresses and phone numbers, etc.



### Machine Learning

- Simultaneously compares numerous de-identified records across datasets, versus just looking at two records in isolation.
- Goes beyond de-identified data contained in the record to also look at additional information allowed to pass through by privacy experts.
- Model trained on billions of patient records with known true matches and representative of the U.S. population.
- Assigns a highly secure, locally encrypted and consistent patient-level identifier (DVID) to the same record across all selected datasets.
- Continuously learns and improves over time.



### Frequency analysis

#### Analysis of relative frequencies of identifying data and de-identified records:

- Datavant Match computes a high frequency score for a commonly used first name and last name, even if the record looks to be the same person based on gender and date of birth, to **correctly issue a non-match**. This ensures high precision in the event that the first and last names are common, that is, two John Smiths were born on the same day.
- Social security number is often considered a reliable field to match on. However, if an SSN-based de-identified record is associated with thousands of other de-identified records, it is highly unlikely to be a true match with those thousands of other patients. Datavant Match **correctly issues a non-match**. In this case, Datavant Match does not give more weight to SSN – which traditional methods often use as a more accurate field to match on.

## Datavant's privacy-preserving record linkage and HIPAA compliance

All datasets eligible for matching and linkage must undergo disclosure risk assessments to minimize risk of patient re-identification and comply with HIPAA requirements.

Getting datasets to HIPAA compliance is a process filled with friction and bottlenecks, and becomes more complicated as patient records in two or more datasets need to be connected. Datavant's Privacy Hub (part of the Datavant Switchboard) streamlines the ability to obtain expert determinations for multiple datasets simultaneously.

Each data partner simply uploads datasets into a neutral, secure environment. Only authorized experts conducting privacy assessments have access to this environment. Once disclosure risk assessments on all datasets are complete, Switchboard automatically moves datasets from disclosure risk assessment, to de-identification and matching. The same patient-level DVID is assigned to each matched record for seamless and accelerated connection and distribution of datasets across the Datavant ecosystem.

## Datavant Match for research-grade use cases

Expanding and accelerating the use of real-world data (RWD) for research is one of the biggest challenges in healthcare today. It is also the biggest opportunity, especially as more sources of RWD come online and pharmaceutical and analytics companies explore how to best integrate, connect and exchange deep clinical data from EHR, genomic testing and more patient-reported information.

We first met Abigail Lane as she started her patient journey seeking treatment for hypothyroidism. Eventually, Abigail received a breast cancer diagnosis, underwent genomic testing to find any possible mutations, and enrolled in a clinical trial. She received care at an academic research center and a community hospital and self-administered treatments at home to gather measurements that she logged through her personal device.

In a world where Abigail's data is de-identified, matched and connected across all her care touch points (including clinical trial data and, ultimately, mortality data), researchers can more quickly and deeply understand the real-world impact of their new drug on patients.

### Researchers can answer questions like:

- Did the new drug slow disease progression long term?
- Were there specific clinical characteristics for patients who experienced slower progression?
- What was the cost effectiveness of the new drug compared to standard of care?
- What was the overall survival rate of patients taking the new drug?



Datavant Match, with its advanced, validated algorithms and high accuracy rates, allows researchers and analytics experts to securely combine a diverse array of datasets for real-world evidence studies.

### Examples of research-grade use cases include:

- External control arms
- Clinical trial recruitment
- Public health studies
- Long-term safety studies
- RWE for new drug indications

These emerging use cases will demand a high-accuracy, research-grade, privacy-preserving record linkage solution that can withstand regulatory scrutiny.

Datavant Match brings the industry its leading research-grade, privacy-preserving matching solution, with a 99.4% precision rate. This is enabled through Datavant's Switchboard, which streamlines and automates privacy-preserving tokenization, disclosure risk assessment, and high-accuracy matching across the greatest number of datasets within the largest health data ecosystem.

## Conclusion

In Abigail Lane's patient journey, there were a minimum of eight care touchpoints. These spanned the ecosystem of healthcare providers, from primary care to specialist visits, lab tests to prescription medications, insurance claims, and then to breast cancer diagnostic and genomic tests and clinical trials.

Now imagine all of Abigail's records accurately matched across these datasets despite differing levels of underlying identifying information. Such a scenario would empower leaders in Abigail's healthcare ecosystem to improve her care and conduct innovative research.

Matching solutions that can address the complexity of this endeavor – with the highest rates of accuracy while preserving patient privacy – are among the most needed in the healthcare industry today.

If Abigail Lane's records are matched across the healthcare ecosystem, imagine the scale at which that data could be connected and exchanged – and the possibilities that would be unlocked for insights and research to improve patient lives.