

GPC LEC 2026

Learning Engagement Conference

Medical College of Wisconsin

Kerrigan Auditorium / Alumni Center

April 23-24, 2026

Speaker Introductions



Alyna Dada
Strategic Engagement Lead,
Research Institutions
Datavant
alyna.dada@datavant.com



Claire Manneh, MPH
Sr. Practice Lead, Research Institutions
Datavant
claire@datavant.com

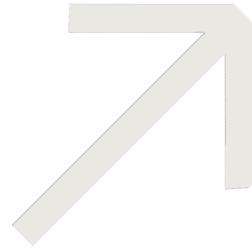


Albert M. Lai, PhD, FACMI, FAMIA
Chief Research Information Officer
Assistant Dean for Biomedical Informatics
Deputy Director, Institute for Informatics, Data
Science and Biostatistics
WashU Medicine
amlai@wustl.edu



Dave Bell
Programmer Analyst
Medical College of Wisconsin,
Clinical & Translation Science Institute
of Southeast Wisconsin
dabell@mcw.edu

Agenda

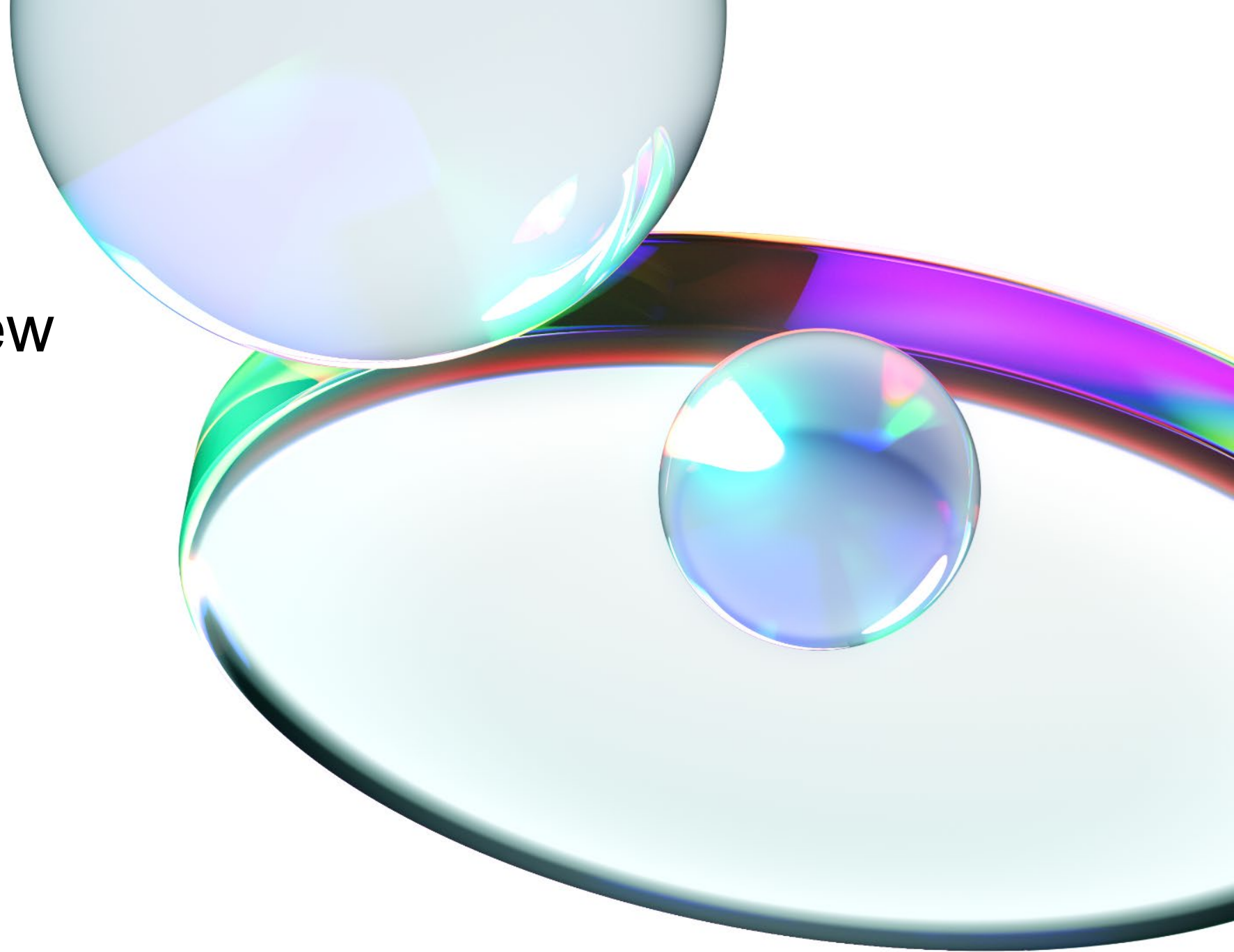


- 01 1:30pm | Welcome and Introductions
- 02 1:30- 1:50pm | Setting the Stage
Datavant Overview
Privacy & Unstructured Data
- 03 1:50- 2:10pm | Putting into Practice
WashU Medicine Use Case
- 04 2:10- 2:30pm | Putting into Practice
MCW Use Case
- 05 2:30 - 2:45pm | Wrap Up & Questions

Datavant Overview

Alyna Dada

Claire Manneh



Every breakthrough begins with data

Observed in the lab.

Recorded in the clinic.

Lived in the real world.

EMRs →

Pharmacy →

Labs →

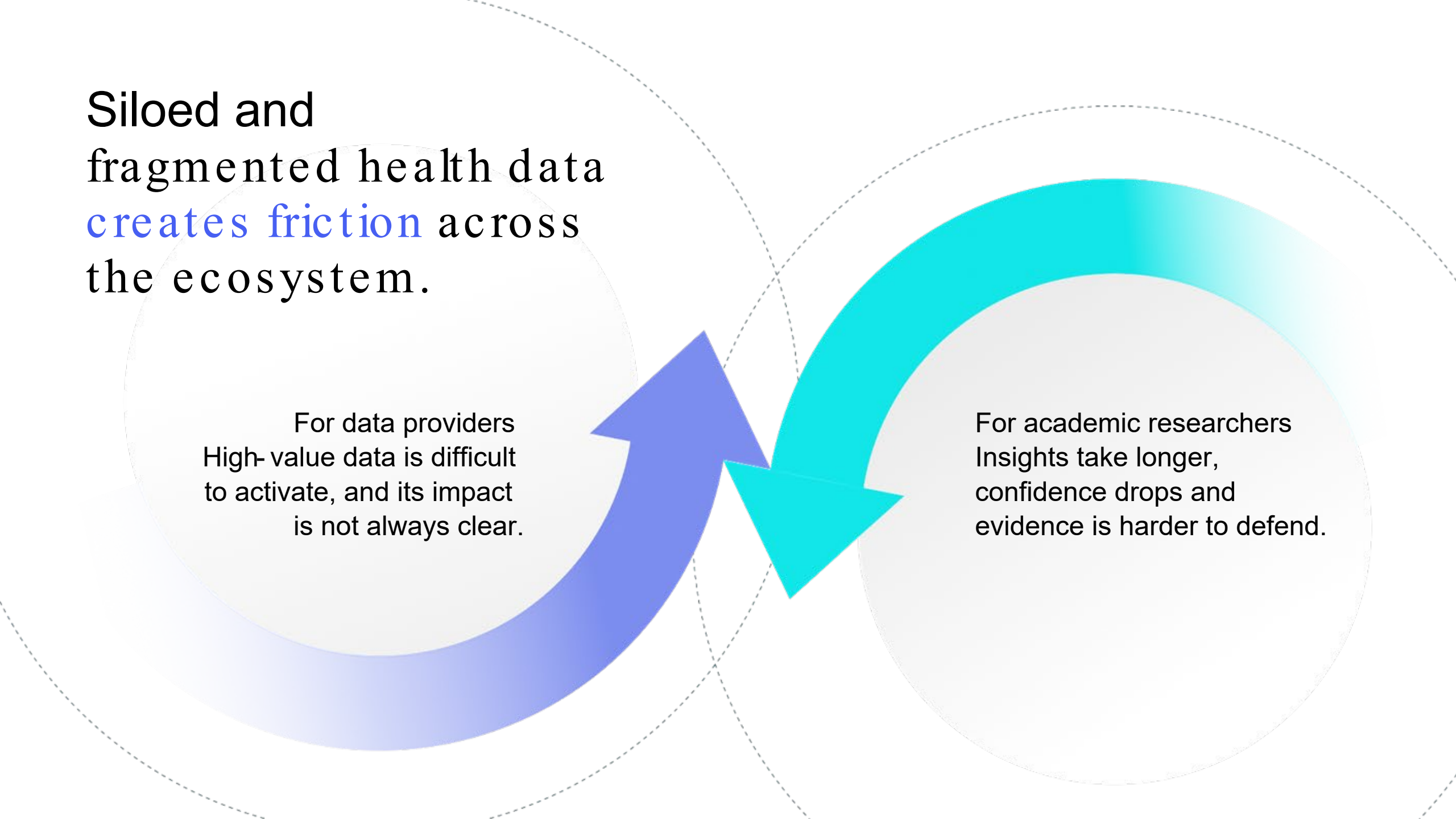
Claims →

Patients →

Siloed and
fragmented health data
creates friction across
the ecosystem.

For data providers
High-value data is difficult
to activate, and its impact
is not always clear.

For academic researchers
Insights take longer,
confidence drops and
evidence is harder to defend.

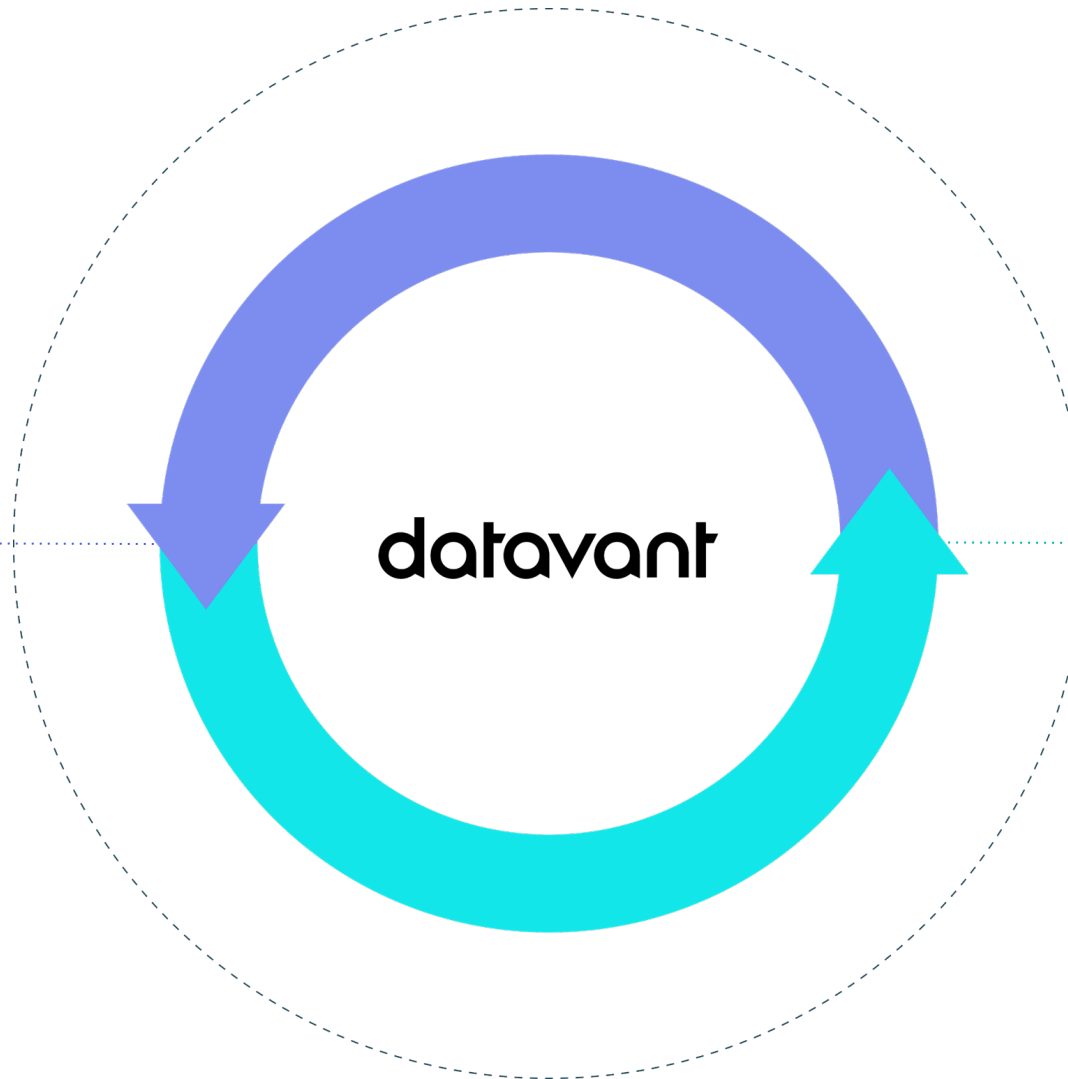


Datavant streamlines the path to critical evidence and insights

Data providers

Data providers transform, enrich and activate high-value data assets—maximizing their utility, accessibility and impact across life sciences use cases.

- Public Sector & Specialty Programs
- Health Systems & Provider Organizations
- Data Sources & Aggregators
- Enterprises



Academic Research Institutions

Academic Research Institutions find, connect and analyze real world and clinical data with confidence —accelerating research, evidence generation and decision-making.

- Principal Investigators (PIs) ○
- Data Analysts & Biostatisticians ○
- Project Managers & Study coordinators ○
- Patient & Stakeholder Engagement ○

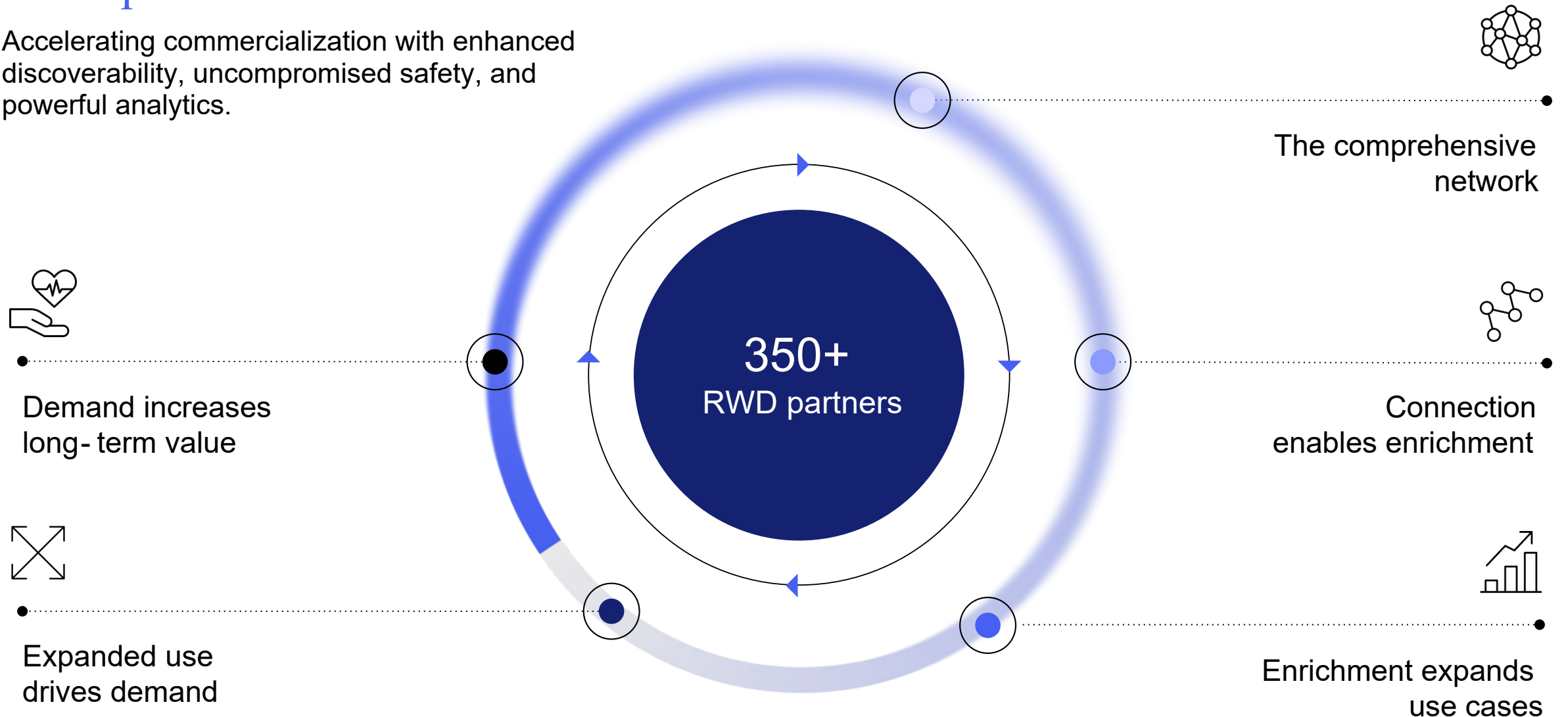
Agentic platform, purpose-built for healthcare

An integrated suite of products and services to source, link and analyze healthcare data securely and confidently.



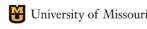
Network scale becomes commercial advantage for data partners

Accelerating commercialization with enhanced discoverability, uncompromised safety, and powerful analytics.



Connecting de-identified data securely across 350+ real world data partners

datavant

Lab & Genomics		EHR		Claims	SDoH, Consumer, & Mortality		Registry	Health Systems and Clinical Research Networks	
									
									
									
									
									
									
									
									

Industry impact at an unmatched scale

• 20
of the top
20 Pharma

• 350+
real-world
data partners

• 75%
of the top 100 U.S.
health systems
connected

• 41B
patient-level
records processed

• 100+
Epidemiologists,
biostatisticians, and
privacy experts

• 240+
peer-reviewed
publications
enabled

PCORnet built the nation's largest clinical research network, but lacked a centralized patient view across institutions and health plans to conduct medical research. With Datavant, they were able to power people - centered research to ultimately produce better, faster, and more relevant evidence in a way that scales to serve the nation.

Use Case

The PCORnet mission is to improve patient centered health research for the American people:

- 47 Million Americans
- 8 research networks
- 78+ health systems
- 1,000+ Community Clinics
- 3,500+ primary care

Their research was slowed by incomplete views of patients, coupled with constant project - by-project hurdles.

PCORnet decided to conduct a formal study to validate a new, secure approach for improving patient data quality and governance across its network.

How Datavant Powered Patient - Centered Research for PCORnet

PCORnet identified 4 sites and tokenized ~170 million patient records across 78 hospitals using Datavant Connect.

They were able to better examine patients journeys, including demographic and clinical characteristics of the population.



> BMC Res Notes. 2022 Oct 31;15(1):337. doi: 10.1186/s13104-022-06243-5.

Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet®, the National Patient-Centered Clinical Research Network

Daniel Kiernan ¹, Thomas Carlton ², Sengwee Toh ³, Jasmin Phua ⁴, Maryan Zirkle ⁴, Darcy Louzao ⁵, Kevin Haynes ⁶, Mark Weiner ⁷, Francisco Angulo ⁸, Charles Bailey ⁹, Jiang Bian ¹⁰, Daniel Fort ¹¹, Shaun Grannis ¹², Ashok Kumar Krishnamurthy ¹³, Vinit Nair ¹⁴, Pedro Rivera ¹⁵, Jonathan Silverstein ¹⁶, Keith Marsolo ¹⁷, 18

Affiliations + expand
PMID: 36316778 PMCID: PMC9620597 DOI: 10.1186/s13104-022-06243-5

Abstract

Objective: The aim of this study was to determine whether a secure, privacy-preserving record linkage (PPRL) methodology can be implemented in a scalable manner for use in a large national clinical research network.

Results: We established the governance and technical capacity to support the use of PPRL across the National Patient-Centered Clinical Research Network (PCORnet®). As a pilot, four sites used the Datavant software to transform patient personally identifiable information (PII) into de-identified tokens. We queried the sites for patients with a clinical encounter in 2018 or 2019 and matched their tokens to determine whether overlap existed. We described patient overlap among the sites and generated a "deduplicated" table of patient demographic characteristics. Overlapping patients were found in 3 of the 6 site-pairs. Following deduplication, the total patient count was 3,108,515 (0.11% reduction), with the largest reduction in count for patients with an "Other/Missing" value for Sex; from 198 to 163 (17.6% reduction). The PPRL solution successfully links patients across data sources using distributed queries without directly accessing patient PII. The overlap queries and analysis performed in this pilot is being replicated across the full network to provide additional insight into patient linkages among a distributed research network.

Keywords: Medical record linkage; Multicenter studies; Patient data privacy.

© 2022. The Author(s).

Peer reviewed Journal Article Publication

Example findings

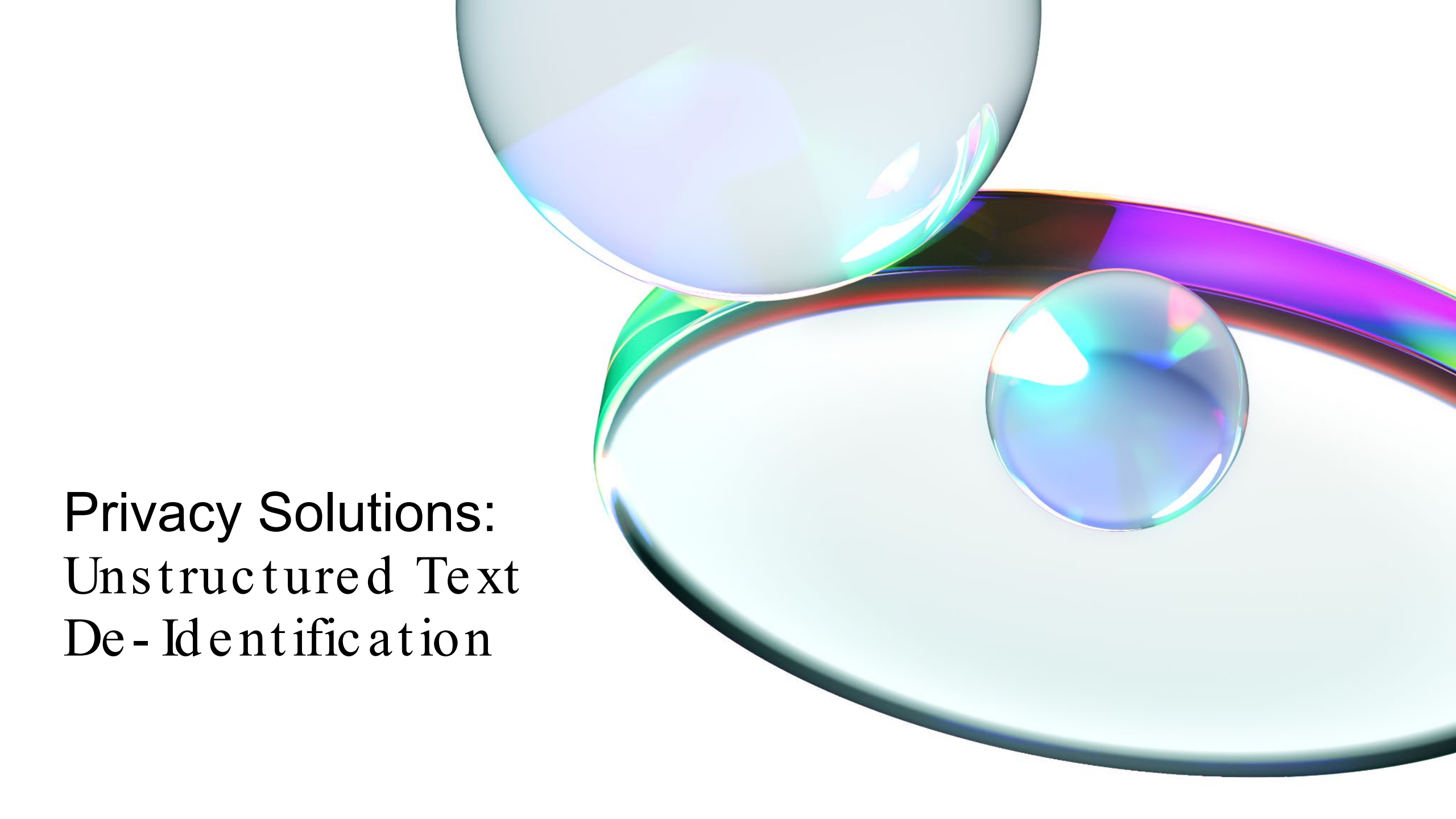
~138M records (81%) identified as unique patients

17.6% reduction in other/missing values in sex

There was increase in disease prevalence in every single clinical characteristics

Results and Value Delivered

“The solution successfully links patients across data sources using distributed queries without directly accessing patient PII. The overlap queries and analysis performed in this pilot is being replicated across the full network to provide additional insight into patient linkages among a distributed research network”.



Privacy Solutions: Unstructured Text De - Identification

A significant portion of the world's health data is unstructured

Free Text Data

For Unstructured Data, unstructured *free text* in particular is critical to understanding the complete patient experience...



Radiology narrative reports



Discharge summaries



Clinical trial narratives



Clinician notes



Pathology reports



Emergency room records



Operative reports



Genetic counseling notes



EHR notes

..but it is **challenging to access and analyze** while preserving patient privacy.

In order to derive value from free text datasets, achieving legal de-identification is essential. This is done through **Redaction & Obfuscation** and then **certified compliant through an Expert Determination*** for unstructured data.

Expert Determination

Now available for structured and unstructured datasets.

What is an Expert Determination?

- One of two methods by which you can achieve de-identification in accordance with the **HIPAA Privacy Rule** (the alternative method is Safe Harbor)
- Performed by **data science experts** with appropriate knowledge of and experience utilizing generally accepted statistical and scientific principles and methods for rendering information not individually identifiable
- Performed on **individual** or **linked** data sets (or a representative sample)

Benefits of Expert Determination:

- Scale your business by maximizing data utility while addressing HIPAA Privacy Rule De- ID Standard
- Ensure you obtain a high - quality statistical analysis performed by industry - leading team of privacy experts, who have performed in combination more than 2,000 Expert Determinations
- Receive details on dataset risk and necessary remediation steps
- Address compliance for data types including unstructured text data

Datavant Unstructured Text De-Identification & Certification

Datavant Solutions to Enable Your Use Case

Your unstructured text data



Notes (e.g., clinician, discharge, EHR, trial)

Reports (e.g., radiology, pathology, operative)

Transcripts (e.g., physician patient visits)

Remediation options

Redaction: remove all detected direct + indirect identifiers

Michael → [REDACTED]

Obfuscation: replace detected direct + indirect identifiers with synthetic alternatives

Michael → **Ethan**

De-identification

Datavant will provide an **Expert Determination** + the remediated de-identified dataset for compliant use

Note: *If you have your own proprietary methodology, Datavant can review and certify the output data with an Expert Determination.

Use-cases unlocked



- Commercialization
- Research
- Internal Analytics
- Insights
- Internal Compliance
- Model Training

Unstructured Text De-Identification

1

Redaction & Obfuscation: De-Id identifying information

Datavant PHI Redaction provides this solution, using the latest in Large Language Models and Named Entity Recognition technology

Datavant can obfuscate the identifiers replacing with a “like” option to increase data utility (e.g., Mary to Marissa)

2

Expert Determination: Review & Certify Output Dataset

Datavant Privacy Solutions conducts a review of redacted data to confirm necessary PII is removed, & complete the de - identification process.

Partner with Datavant to minimize privacy risk, enable wider and more compliant data handling across your organization, and **unlock new business opportunities for your unstructured text data**

Redaction

- Removes all detected direct + indirect identifiers
- Can fragment flow of text due to missing words/entities
- Very small but non-zero residual risk due to undetected identifiers

Obfuscation

- Replaces detected direct + indirect identifiers with plausible synthetic alternatives
- 100% success rate from internal tests that a team of medical proofreaders could not re-infer original identifier
- Increased privacy - any undetected identifiers receive quantifiable privacy protection from ambiguity of their real vs obfuscated status
- Enhanced utility
 - Seamless, plausible text without redactions
- Not appropriate for binary flags - e.g. if a patient has veteran status

Example Original Text: Diabetes was diagnosed in Michael

Diabetes was diagnosed in [REDACTED]

Diabetes was diagnosed in Ethan

Redaction offering removes direct & indirect patient identifiers

Original unstructured free text with patient identifiers

PatientID, Notes, Notes
Surgery_a4, "
name: Ben Stinson
PT ID: 6843115 65
DOB: 10/05/1940

Home Phone:(385) 3528050
Address: 979 Atrium street, Greendale, Bargersville, NE 40065
Insurance No: 85 0348

Date of visit: 25-Aug-63

Huntington Hospital
Address: 100 Ws California Blvd, Pasadena, CA 91105 Drs D Pateniak and vs Naipaul

PATIENT SURGICAL HISTORY

assessment requested by ins company (JAYA INT Ltd).
On 11-Aug 7:00 a.m. Ms. Stinson. was admitted to Hungtington Hospital and under the care and treatment of Surgeon JJ Waters (General Surgeon). At 10:22 a.m. she was taken to the Operating Room holding area and beginning at 12:22 p.m. she underwent a Total Thyroidectomy for a malignant tumor. General Anesthesia was administered and monitored by Dr. Andrews, Anesthesiologist. Fentanyl, Atropine and Droperidol was given.
During the procedure Dr. Waters noted the mass to be "large" and "infiltrating". Frozen Section biopsy revealed Papillary Adenocarcinoma and a 3 ½ hour, Thyroidectomy was completed. Ms. Stinson was taken to the PACU for recovery at 2:45 p.m. and noted in stable condition however, her blood pressure was elevated reading 220/92. (Nurse George Fredrick, RN:4985)

signed: ETK Delafield MD
22-Dec-87",

>

Redaction of patient identifiers through our ML model

PatientID, Notes, Notes
Surgery_a4, "
name: [REDACTED]
PT ID: [REDACTED]
DOB: [REDACTED]

Home Phone: [REDACTED]
Address: [REDACTED]
Insurance No: [REDACTED]

Date of visit: [REDACTED]

Huntington Hospital
Address: [REDACTED]

PATIENT SURGICAL HISTORY

assessment requested by ins company (JAYA INT Ltd).
On [REDACTED] [REDACTED] was admitted to Hungtington Hospital and under the care and treatment of Surgeon [REDACTED] (General Surgeon). At [REDACTED] she was taken to the Operating Room holding area and beginning at [REDACTED] she underwent a Total Thyroidectomy for a malignant tumor. General Anesthesia was administered and monitored by [REDACTED], Anesthesiologist. Fentanyl, Atropine and Droperidol was given.
During the procedure [REDACTED] noted the mass to be "large" and "infiltrating". Frozen Section biopsy revealed Papillary Adenocarcinoma and a [REDACTED], Thyroidectomy was completed. [REDACTED] was taken to the PACU for recovery at [REDACTED] and noted in stable condition however, her blood pressure was elevated reading 220/92. [REDACTED], [REDACTED])

signed: [REDACTED]
[REDACTED],

WashU Medicine's experience with Use of LLMs for converting unstructured data into structured data



Albert M. Lai, PhD, FACMI, FAMIA
Chief Research Information Officer
Assistant Dean for Biomedical Informatics
Deputy Director, Institute for Informatics, Data
Science and Biostatistics
WashU Medicine
amlai@wustl.edu

Extracting cancer phenotypes from EHRs

- Focus on non-small cell lung cancer
- Cancer stage
- Treatment (chemotherapy, radiation, surgery)
- Recurrence (presence of relapse)
- Metastasis / Affected organs during recurrence



Approach

- Zero-shot, no fine-tuning
- GPT-4
- GPT-3.5
- Flan-T5
- Llama-3
- Rule-based NLP (medspaCy, scispaCy)

Model Performance by Entity Category

Model	Staging (F1)	Treatment (F1)	Recurrence (F1)	Organs (F1)	Overall Trend
GPT-4	~0.92	~0.92	~ 0.96 (best)	~0.68	Best overall across all categories
GPT-3.5-turbo	~0.90–0.92	~0.90	~ 0.96 (≈ GPT-4)	~0.62–0.65	Very close to GPT-4
Llama-3-8B	~0.88–0.91	~0.90	~0.83–0.85	~0.68	Strong non-GPT model
Flan-T5-xxl	~0.82–0.85	~0.85–0.88	~0.75–0.77	~0.55–0.58	Mid-tier
Flan-T5-xl	~0.78–0.82	~0.78–0.82	~0.75–0.77	~0.50–0.55	Slightly below XXL
scispaCy	~0.65–0.70	~0.60–0.65	~0.60–0.62	~0.55–0.58	Best rule-based
medspaCy	~0.60–0.65	~0.60–0.63	~0.58–0.60	~0.50–0.55	Weakest overall

Results

- **GPT-4 achieved highest performance (F1 up to ~0.96)**
- **GPT-3.5 performed similarly**
- **LLMs outperformed rule-based methods**, especially for:
 - Context understanding
 - Handling variability in clinical language

Key Takeaways

- LLMs enable **accurate, scalable extraction of complex clinical information**
- Provide **better contextual understanding than traditional NLP**
- Less sophisticated foundational models may provide similar performance to those of larger models
- It may be more cost effective to run smaller models, but it may need to be balanced with the cost of engineering time associated with making the lesser models work

Model compute time and success rate

Model	Successfully analyzed echocardiogram reports, <i>n</i> (%)	Processing time per report in seconds, median (IQR)
Gemma2:9b-instruct	504 (99.4%)	5.737 (5.484-6.06)
Gemma2:9b	503 (99.2%)	3.583 (3.34-3.774)
Llama3:8b	499 (98.4%)	2.509 (2.349-2.594)
Gemma:7b	499 (98.4%)	2.169 (2.064-2.335)
Llama3:8b-instruct	498 (98.2%)	4.89 (4.684-5.109)
Llama3:70b	497 (98.0%)	63.242 (59.91-66.568)

Overall model performance

- Gemma2:9b-instruct
 - Precision: 0.973
 - Recall: 0.959
 - F1: 0.965
-
- Llama3:70b – also very accurate, but computationally expensive
 - Gemma2:9b – slightly less accurate, but faster
-
- **Llama3:70b, Gemma2:9b- instruct, and Gemma2:9b perform high-quality extractions nearing manual gold-standard clinician annotation performance.**

Lessons Learned

- LLMs can be very effective at converting unstructured data into structured data
- Performance needs to be evaluated
- Smaller models may work, depending on the scenario
- Use of smaller models are needed in order to scale up to perform wholesale conversion of unstructured EHR data into structured warehouse data

Future plans

- Put EchoLLM into production
 - Structure all historical echo reports and store in warehouse as “computationally derived data”
- New efforts on widely structuring cancer staging data
- Automate deidentification of all clinical notes

Infrastructure needed

- Initially utilizing HIPAA-compliant GPU research compute cluster
- Purchase GPU cluster for local informatics research and research data warehousing needs
 - 8x H200s for large general purpose LLMs
 - 4x RTX PRO 6000 Blackwells for medium sized agentic LLMs
 - 4x RTX PRO 6000 Blackwells for running smaller, faster or other comparator LLMs

MCW Case Study: Fact Discovery in Unstructured Data



Dave Bell IV

Programmer Analyst

Medical College of Wisconsin

Clinical & Translation Science Institute of Southeast Wisconsin (CTSI)

Center for Biomedical Informatics (CBMI)

Geospatial, Epidemiology, and Outcomes Shared Resource (GEOSR)

Integrated Cancer Data Resource (ICDR)

dabell@mcw.edu

Unstructured Data in CRDW

- 1.6TB of unstructured data as clinical notes
- HNO Notes
 - Progress Notes
 - Consults
 - H&P
- Imaging reports
 - Narratives & Impressions
- order result comments
 - Various comments on orders related to procedures and other orders

Electronic Medical Record Search Engine

- EMERSE
- University of Michigan
- Deployed as a self-service tool for CRDW
- Use cases:
 - Cohort discovery
 - chart review
- 18 IRB Protocols
- 89 Users



Patients All Local Patients (1,802,736)

Filters Encounter Type

Terms mold asthma

Results

HIGHLIGHT DOCUMENTS

FIND PATIENTS

Summaries Demographics Trends

4,200 patients matched the search criteria

Electronic Medical Record Search Engine

- No notes extraction
- What you search is literally what you get

LOINC Document Ontology

- Logical Observation Identifiers, Names, and Codes
- Standardization of HNO note types
- Bag of Words approach
 - Subject Matter Domain
 - Dentistry, Allergy, etc...
 - Kind of Document
 - Note, Report, Prescription list, etc...
 - Type of Service
 - Consultation, Procedure, Counseling, etc...
 - Setting
 - Outpatient, Telehealth, Pharmacy, etc...
 - Role
 - Patient, Nurse, Physician, etc...

LOINC Document Ontology

1. Document Metadata

ED Patient Depart Summary, University Hospital, Care
Delivery, UH Emergency

2. BOW

hospital, discharge, care, emergency,
summary, delivery, note, patient, {role},
{setting}

LOINC BOWs

3A. 59258-4

Discharge,
emergency,
summary, note,
{role}

3B. 34112-3

hospital,
note, {role}

3C. 11506-3

progress, note,
{role}, {setting}

Mapping Summary

- Word order is lost in BOW
- “{role}” and “{setting}” wildcards are added.
 - Priority is given to real axis, if available (e.g. Emergency)
- Synonymy list
 - Depart -> Discharge
 - Summary -> Note
- Chosen LOINC Code is **59258-4**

Biomarker Status Extraction From Genomic Reports

- Data entry problem
 - different degrees of completeness
- LLM to replace manual data entry
- PDF reports -> text
- Iterately reached 100% accuracy on sample set
 - Awaiting validation on full dataset via investigators
- GPT 4o with Azure Document Intelligence

CEBPA Result	CEBPA Significance	CEBPA_significance_note	CEBPA_type	CEBPA_type_evidence
Positive	Pathogenic	Reported under 'Pathogenic Variants Detected'.	Monoallelic	CEBPA c.123_456dup, p.Y789* (NM_012345.6). No mention of a second mutation.

Immune Related Adverse Events

- Incident detection and grading post-therapy
 - Limited list of events (140)
 - Steven's johnson syndrome
 - Anemia
 - Nausea
- Severity classification using CTCAE
 - Anemia: HGB 11.9 (Grade 1)
 - Grade 1: Hemoglobin (Hgb) <LLN - 10.0 g/dL; <LLN - 6.2 mmol/L; <LLN - 100 g/L
- Longitudinal summary

Anemia was first noted on March 4, 2024, with a grade of 1, and fluctuated between grades 1 and 2 until the last mention on September 4, 2024, where it was recorded as grade 1 again.

Jet ski Trauma

- Rare event detection
- Case where current tools (EMERSE) fall short without chart review

timing_phrase	incident_date	incident_date_precision	evidence_quality	evidence	injury_description
on 04/04/2004	2004-04-04	Exact	Moderate	Patient was recently hospitalized after a jet ski accident on 04/04/2004, which contributed to his CHF decompensation. He continues to have poorly controlled blood pressure and is being followed for sleep issues.	Not specified

Questions?

Thank you!