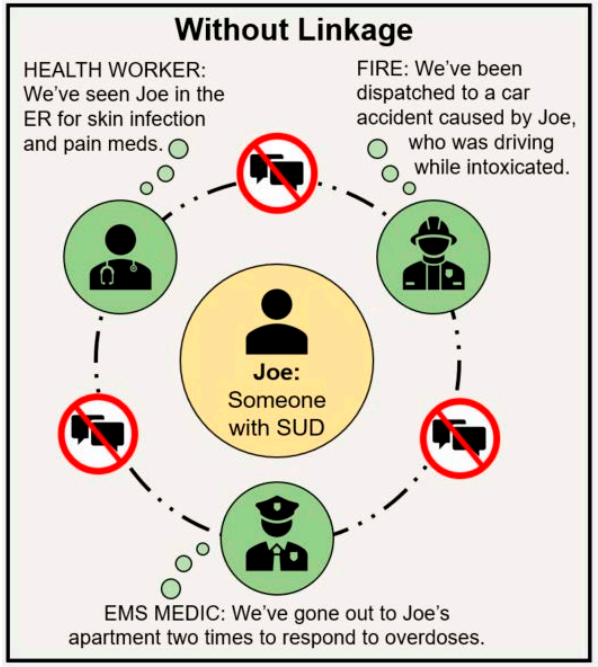
A Statewide Substance Misuse Data Commons: An Artificial Intelligence-enabled Service with Multi-Stakeholder Input and Team Science Design

Majid Afshar, MD, MS

NIH/OD: Building a Substance Use Data Commons for Public Health Informatics Administrative Supplements to Support Collaborations to Improve the AI/ML-Readiness of NIH-Supported Data

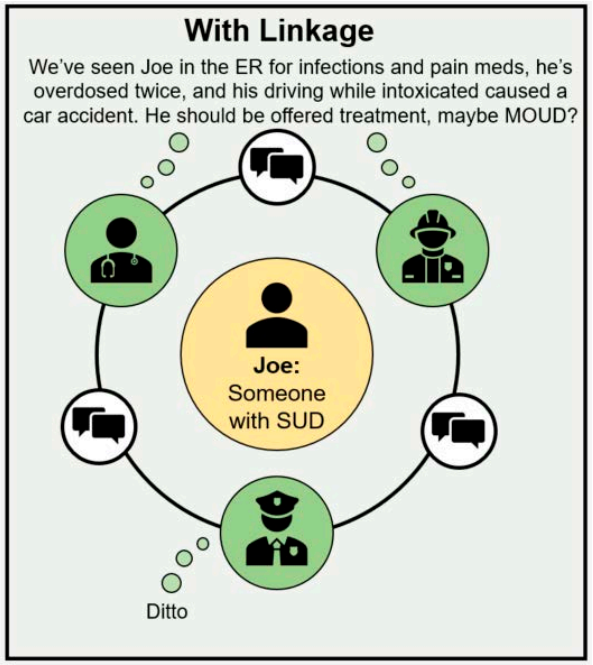
The Problem

- Unhealthy substance use can preordain poor outcomes.
- Repeated encounters with ED or first-responders.
- Fragmented data systems make it difficult to see full picture

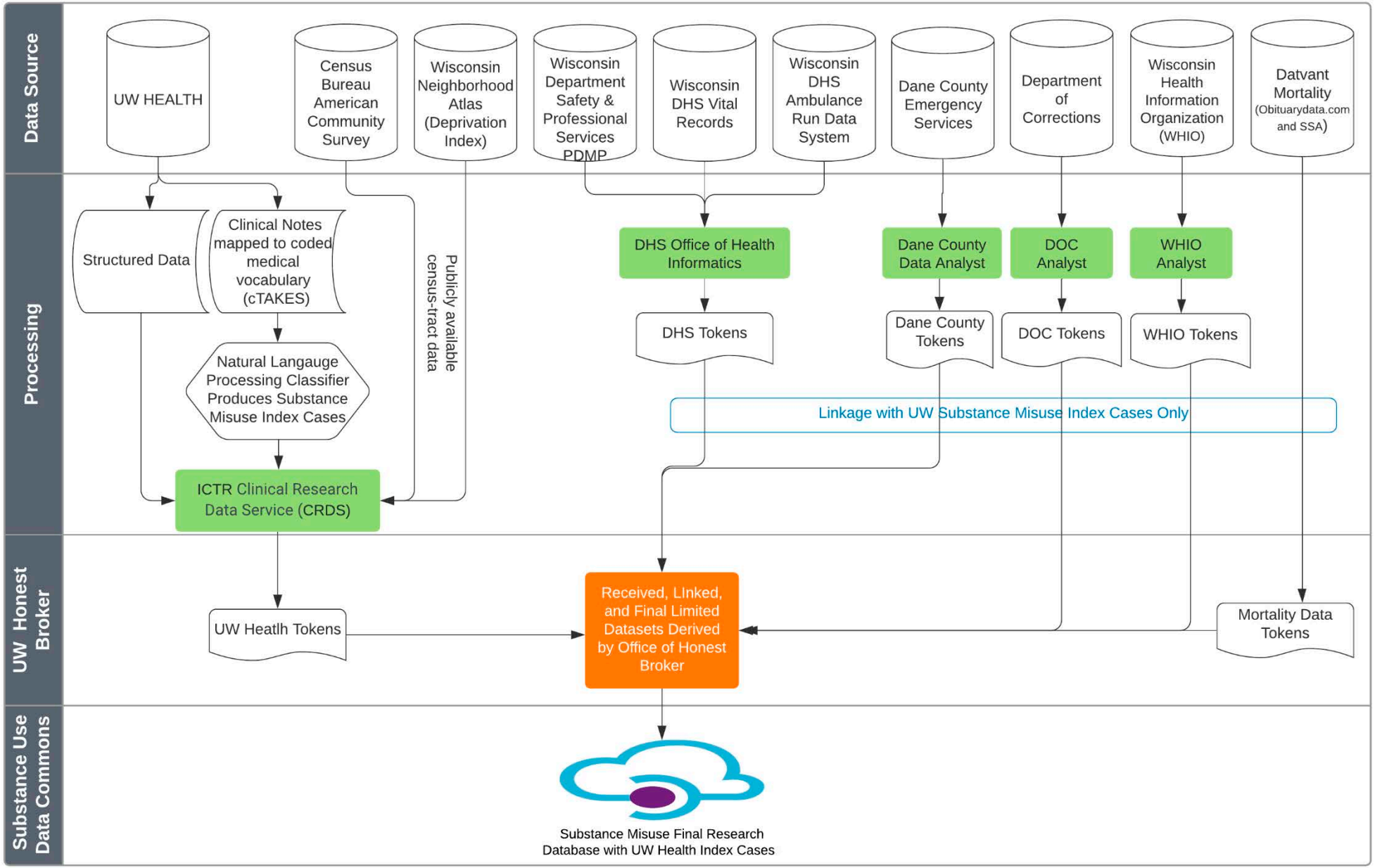


The Solution

- Linked, comprehensive data may allow us to reliably identify, risk stratify, and prioritize care for prevention of substance use conditions

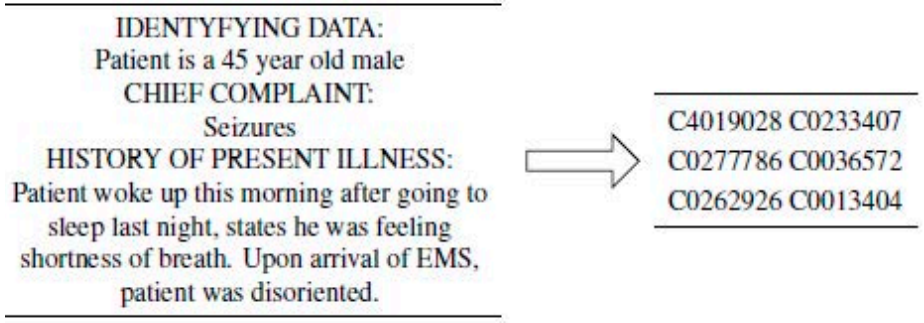


The Solution – Privacy Preserving Data Linkage



Innovation

- Privacy preserving linkage with tokenization to allow line-level data
  - FIPS-140-2 cryptographic modules: SHA-256, AES-128
  - Deterministic and Probabilistic Net Token Approach
- Deidentification of EHR notes with concept mapping to UMLS
- Cloud-computing environment
  - NIST 800-53a security compliance framework
- Addiction treatment regulated by 42 CFR Part 2
  - Meeting legal and regulatory requirements with data owners
- Longitudinal data collection (limited data set) with national obituary feeds of mortality data every 2 weeks
- Linkage Honest broker service
  - Escrow for the cryptographic hash codes (tokens)



DV-ID	pseudoID	DOB	Gender	Token1	Token2	Token3…
0	ABBCC	7/4/1990	F	Sadfl;j234dsaf08u	3r908nmdli9d	Dsafkjl;…
1	ABBCD	7/5/2000	M	Asdfiienwd907898	324nadsfvuion	DOHEWN..
2	ABBCF	7/6/2001	F	@#sdfklj32jfasdh312	2hadsf9lkewrytnd	Jcoopdusf…
2	55435552	7/6/2001	F	@#sdfklj32jfasdh312	hhdshsdhfaskdkkl	Jcoopdusf…
3	78687768	2/2/1992	F	Fdsalk;jadfslk;efw	Djsefwohiew	Dfsjwefoi…
4	78888889	3/3/177	M	0sf3r2ojeljksdfa	][]dsaf9^4sdfhj	7nsadfh23…
4	65431239	3/3/177	M	0sf3r2ojeljksdfa	][]dsaf9^4sdfhj	7nsadfh23…

Datavant Match Basic for Linux tool was run. The Net-Tokens match model was used.

Results: Linkage and Mortality

- Data linked across 37,162 UW Health patients and their 65,275 encounters between 1/1/2008 and 12/31/21

- State ambulance run data system:
  - Missing data was less than 1% on identifiers with first/last name, gender, date of birth, and Zip3

- 20,318 (54.6%) of the UW index patients were linked to the statewide database and 6.6% had duplicates

- 8,355 (22.5%) deaths were identified from the national mortality files linked to our patient cohort
  - Inpatient deaths (n=1,416): 267 could not be found in the national mortality files
    - 81.1% sensitivity/recall.
- > 93% of the deaths between the EHR death timestamps and national data timestamps were within 30 days
  - 3.1% (n=1,092) were deaths outside of Wisconsin.

Prediction of 60-day readmission or mortality

- Largest gain in Absolute Reclassification Index was between EHR and EHR/ADI/ACS models at 4.1%.
- Most important variables:
  - leaving against medical advice, sex, heart rate, alcohol testing, and admission labs with chloride level, red cell distribution width, calcium level, sodium level, carbon dioxide level, and hemoglobin level.
- No ACS census-tract variables were in the top 20.

Model performance for predicting 60-day mortality or readmission in patients with substance misuse

Models	AUC (95% CI)	P-value
EHR	0.691 (0.678-0.703)	-
EHR + ADI	0.692 (0.68-0.704)	0.09
EHR + ACS	0.694 (0.68-0.7064)	0.27
EHR + ADI + ACS	0.695 (0.683-0.707)	0.18

EHR = electronic health record
ADI = Neighborhood Atlas Area Deprivation Index
ACS = Census Bureau American Community Survey

Highlights of Work

- Two 2023 AMIA Informatics Summit Submissions
- NIH/NIDA Clinical Trials Network Data Science Workshop Invited Lecture
- SUD Data Commons Road Map Manuscript Under Review
- NLP SUD Cohort Discovery Tool published in Lancet Digital Health
- Data Commons now accessible with credentialing and authorization
- Plan for Public GitLab Repository
  - Currently private to finalize open-source license and Data Dictionary

Challenges and Future Work

- Legal and Regulatory Hurdles for data sharing by data owners
- Dissemination of privacy-preserving linkage technology
- Data governance and auditing
- Scaling NLP pipeline and building common data model

- Plan for NIH R01 Submission for new aims with multicenter data
  - Infrastructure design for scaling in Data comons

THANK YOU

https://www.medicine.wisc.edu/apcc/icu-data-science-lab

Twitter: @UW_ICU_DataSci

Majid.afshar@wisc.edu