JAMIA Open, 2023, **6(4)**, ooad092
https://doi.org/10.1093/jamiaopen/ooad092
Research and Applications

Research and Applications

Creation of a data commons for substance misuse related health research through privacy-preserving patient record linkage between hospitals and state agencies

Majid Afshar , MD, MS[1,*], Madeline Oguss, MS[1], Thomas A. Callaci, CISSP, CDA[1], Timothy Gruenloh, MS[1], Preeti Gupta, MD, MPH[2], Claire Sun, BS[1], Askar Safipour Afshar, MS[1], Joseph Cavanaugh, BA[1], Matthew M. Churpek, MD, MPH, PhD[1], Edwin Nyakoe-Nyasani[3], Huong Nguyen-Hilfiger, MPH[3], Ryan Westergaard, MD, MPH, PhD[1,3], Elizabeth Salisbury-Afshar, MD, MPH[1,3], Megan Gussick, MD[1], Brian Patterson , MD, MPH[1], Claire Manneh, MPH[4], Jomol Mathew, PhD[1], Anoop Mayampurath, PhD[1]

[1]School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53706, United States, [2]Division of Pulmonary and Critical Care, University of Illinois-Chicago, Chicago, IL 60607, United States, [3]State of Wisconsin Department of Health Services, Madison, WI 53703, United States, [4]Datavant Incorporated, San Francisco, CA 94104, United States

*Corresponding author: Majid Afshar, MD, MS, Department of Medicine, School of Medicine and Public Health, University of Wisconsin-Madison, 600 Highland Avenue, CSC H4/616, Madison, WI 53792 (majid.afshar@wisc.edu)

Abstract

Objectives: Substance misuse is a complex and heterogeneous set of conditions associated with high mortality and regional/demographic variations. Existing data systems are siloed and have been ineffective in curtailing the substance misuse epidemic. Therefore, we aimed to build a novel informatics platform, the Substance Misuse Data Commons (SMDC), by integrating multiple data modalities to provide a unified record of information crucial to improving outcomes in substance misuse patients.

Materials and Methods: The SMDC was created by linking electronic health record (EHR) data from adult cases of substance (alcohol, opioid, nonopioid drug) misuse at the University of Wisconsin hospitals to socioeconomic and state agency data. To ensure private and secure data exchange, Privacy-Preserving Record Linkage (PPRL) and Honest Broker services were utilized. The overlap in mortality reporting among the EHR, state Vital Statistics, and a commercial national data source was assessed.

Results: The SMDC included data from 36 522 patients experiencing 62 594 healthcare encounters. Over half of patients were linked to the statewide ambulance database and prescription drug monitoring program. Chronic diseases accounted for most underlying causes of death, while drug-related overdoses constituted 8%. Our analysis of mortality revealed a 49.1% overlap across the 3 data sources. Nonoverlapping deaths were associated with poor socioeconomic indicators.

Discussion: Through PPRL, the SMDC enabled the longitudinal integration of multimodal data. Combining death data from local, state, and national sources enhanced mortality tracking and exposed disparities.

Conclusion: The SMDC provides a comprehensive resource for clinical providers and policymakers to inform interventions targeting substance misuse-related hospitalizations, overdoses, and death.

Lay Summary

Substance misuse comprises a heterogeneous and complex set of conditions associated with high mortality and regional and demographic variation. Healthcare providers and public health agencies who design treatment and preventative interventions have focused primarily on fatal events. Recently, the Office of National Drug Control Policy recommended shifting focus to early warning signs—emergency department visits or hospitalizations—that lie on the path to fatality. To aid this transition, we constructed the Substance Misuse Data Commons (SMDC), a first-of-its-kind informatics platform that links hospital data from adult cases of substance (alcohol, opioid, nonopioid drug) misuse from a regional health system to census, national mortality, and state agency data. Our article describes our privacy-ensuring data-linking process and the characteristics of SMDC patients. Over half of the 36 522 SMDC patients had data from statewide ambulance and prescription drug databases. The majority of deaths were attributable to chronic diseases, more so than overdose deaths. There was a 49.1% overlap in death cases across the 3 mortality data sources, highlighting the value in our use of electronic health records, state vital records, and national death databases. With the SMDC, healthcare providers and policymakers may analyze a unified record of information that is useful for informing preventive strategies for both health systems and health departments.

Key words: substance abuse; opioids; alcohol; health information exchange; data commons.

Background and significance

Substance misuse is a complex and multifaceted problem that requires a comprehensive and coordinated response from governments, healthcare providers, and community stakeholders. Death from opioid misuse, nonopioid illicit use (ie, cocaine, methamphetamine), and alcohol misuse continue to increase

Received: August 21, 2023; **Revised:** October 4, 2023; **Editorial Decision:** October 10, 2023**Accepted:** October 16, 2023

annually with significant variation between geographic regions, race, and ethnicity.[1,2] Stimulant-related hospitalizations are also on the rise, with nearly one-third of all substance misuse-related emergency department (ED) visits involving alcohol.[3,4] While poisoning and withdrawal are the most recognized causes of hospitalizations among patients with substance misuse, many indirect causes related to substance misuse include noncommunicable diseases (ie, cardiovascular disease and chronic lung disease), infectious diseases (ie, hepatitis, wound infections, endocarditis, sepsis), and trauma (intentional and unintentional; ie, car accidents).[5–8] Recognizing this, the National Drug Control Strategy (NDCC) recommends that regional public health agencies shift away from focusing solely on substance-related fatalities to using data from nonfatal events for designing treatment, interventions, and policies to combat the substance misuse epidemic.[9] However, a significant impediment to implementing these guidelines is the fragmentation of data across health systems and public health surveillance entities. For example, state agencies collect information on ED visits, prehospital ambulance runs, socioeconomic factors, and housing status but are limited in linking with more comprehensive electronic health record (EHR) data. Therefore, there is a need for a data platform that integrates information from multiple entities, thereby presenting a comprehensive picture crucial to inform harm reduction policies.

Within health informatics, the use of data commons has increased over recent years. A data commons is a cloud-based infrastructure that comprises data storage and computational tools and applications required for managing and analyzing the data.[10] The use of data commons has grown rapidly across many biomedical applications.[11] A data commons, in general, must include key considerations: (1) storage and computational resources must be collocated within a cloud-based infrastructure; (2) permissible use of hosted data must be covered by agreements with data-supplying entities; and (3) data must conform to the FAIR (Findable, Accessible, Interoperable, Reusable) digital compliance principles of the NIH Data Science Strategic Plan.[12] Building a data commons for substance misuse research has additional challenges in ensuring the confidentiality and security of patients with addiction while still providing researchers access to the tools for conducting analyses that will inform prevention efforts and learn how health systems may interact with their communities and local health departments to provide better care.

Objective

In this study, we describe the establishment of the Substance Misuse Data Commons (SMDC), a cloud-based platform that aims to solve the problem of isolation of health systems from the society around them to foster population health research for substance misuse. The goal of the SMDC is to provide a first-of-its-kind informatics platform using public–private partnerships to advance research in substance misuse prevention, treatment mapping, and clinical care.

Methods

Data commons design and study population

Figure 1 describes the data owners, data linkage, and final data curation of the SMDC. We used a population health design approach that identifies substance misuse patients from a health system and connects them to out-of-hospital data modalities, thereby developing a regional perspective of health systems and their catchment areas with better data integration within and outside the EHR. Our population health approach follows a similar definition to the Center for Disease Control and Prevention on Population Health and is different from traditional public health research.[13] The SMDC includes adult (18 years or older) patients with ED visits or inpatient hospitalizations and at least 1 encounter with a substance use-related diagnosis code at 2 University of Wisconsin (UW) hospitals in Dane County, WI, between 2008 and 2022. Substance-related International Classification of Diseases (ICD)-9/10 codes from existing UW Hospital EHR data were used to identify patients with substance misuse.[14] A total of 342 International Classification of Diseases codes for misuse across alcohol, opioids, stimulants, depressants, and illicit drugs were included (Table S1). EHR data corresponding to all encounters were extracted and linked with non-EHR data sources at the patient level.

SMDC data

Data use agreements were reached between the study team and the data governance board for each data owner (see Supplementary Material for details). The UW-Madison Health Sciences Internal Review Board (IRB) approved the use of a Health Insurance Portability and Accountability Act (HIPAA)-limited (ie, with patient identifiers removed but with dates and timestamps retained to allow for longitudinal analysis, and census block group IDs for geolocation analysis) dataset (IRB No. 2021-0553). The IRB and each contract were carefully crafted to allow future users to be added for related projects with opt-out options by the data owners on IRB-approved projects. The establishment of our data commons and each contract allows UW researchers and collaborators to get credentialed and authenticated through UW's School of Medicine and Public Health (SMPH) to access the dataset permitted to them via separate IRB approvals, without further contracting with data owners. Table 1 describes each data modality within the SMDC. Details regarding data governance and information within each data modality are provided in the Supplementary Material.

Data linkage across entities

We utilized a combination of Privacy-Preserving Record Linkage (PPRL) and Linkage Honest Broker Services to link disparate datasets across multiple entities. The PPRL was licensed and subscribed through Datavant software for all data-contributing sites. Each data-contributing entity generated deidentified universal patient keys (tokens) by running the software on-premises behind firewalls for each data owner. The token generation process was accomplished by processing identifiable demographic attributes (eg, first name, last name, date of birth) through a cryptographic hashing process that produces a series of encrypted keys certified as deidentified data under the HIPAA Expert Determination standard, which remains with each data owner. The output hash was then encrypted using a site-specific key, ensuring each data owner's tokens were unique with a subsequent step to hash again, and 1-way transit tokens were sent to the SMDC honest broker (ie, the Office of the Honest Broker in the School of Medicine and Public Health at UW-Madison) for linkage. The honest broker is a neutral entity located outside of any research team or data-contributing site, serving as an escrow for the tokens and utilizing the Datavant software

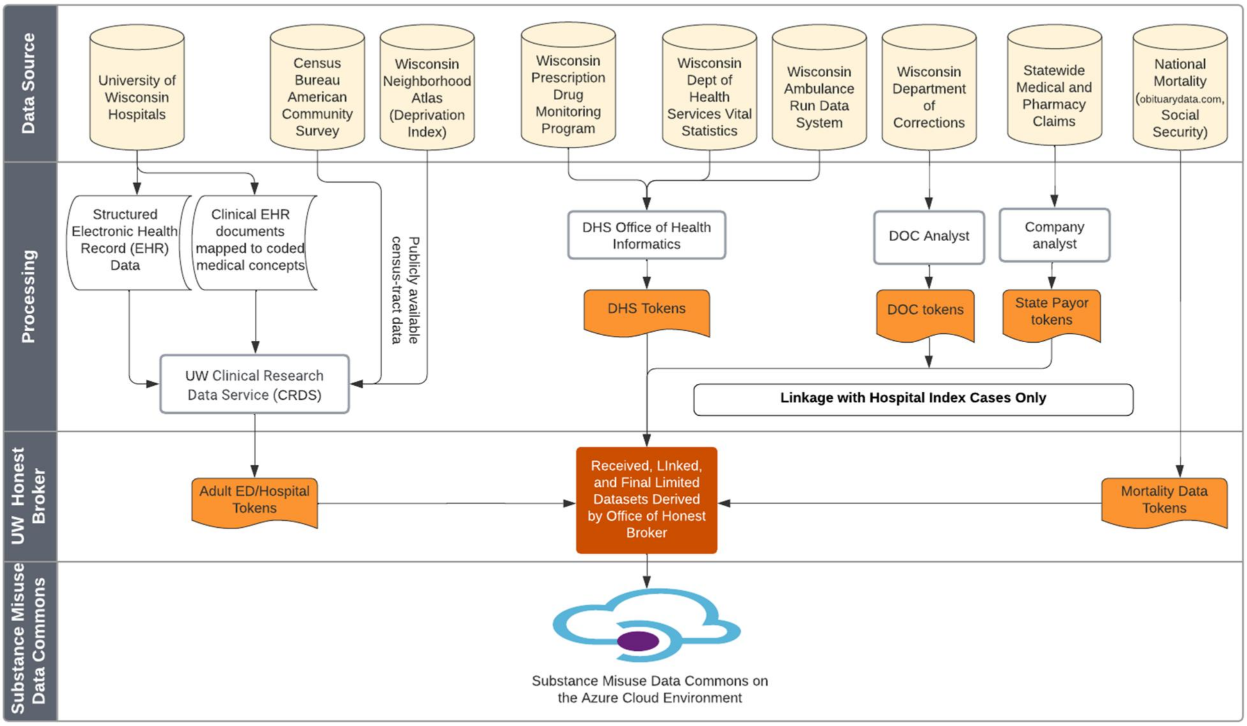
Data Source
University of Wisconsin Hospitals
Census Bureau American Community Survey
Wisconsin Neighborhood Atlas (Deprivation Index)
Wisconsin Prescription Drug Monitoring Program
Wisconsin Dept of Health Services Vital Statistics
Wisconsin Ambulance Run Data System
Wisconsin Department of Corrections
Statewide Medical and Pharmacy Claims
National Mortality (obituarydata.com, Social Security)
Processing
Structured Electronic Health Record (EHR) Data
Clinical EHR documents mapped to coded medical concepts
Publicly available census-tract data
UW Clinical Research Data Service (CRDS)
DHS Office of Health Informatics
DHS Tokens
DOC Analyst
DOC tokens
Company analyst
State Payor tokens
Linkage with Hospital Index Cases Only
UW Honest Broker
Adult ED/Hospital Tokens
Received, LInked, and Final Limited Datasets Derived by Office of Honest Broker
Mortality Data Tokens
Substance Misuse Data Commons
Substance Misuse Data Commons on the Azure Cloud Environment

Figure 1. Data flow for substance misuse data commons.

Table 1. Data modalities collected in the substance misuse data commons.

Structured Electronic Health Record (EHR) Data	*Characteristics and physiology-level*: Demographics, diagnoses, comorbidities, vital signs, pain scores, behavioral scores (depression, substance use), medications, procedures, laboratory results *Hospital-level*: Discharge disposition including mortality, length of stay, hospital service, acuity level
Unstructured EHR documents	Mapped concept unique identifiers (CUIs as deidentified version of notes) from clinical notes and imaging reports collected during a hospital encounter
Census Bureau American Community Survey (ACS)	Unemployment, dependency, education, crowded housing, per capita income, and poverty (economic hardship index). Additional factors including median household income/gross rent, household size, citizenship, food stamps, terrestrial development index, disability status, marital status, preferred language, and insurance status
Wisconsin Neighborhood Atlas Area Deprivation Index (ADI)	Census block group with a ranking between 1 and 100 for national- and state-level data. Group 1 is the lowest ADI (least disadvantaged neighborhood) and 10 is the highest ADI (most disadvantaged neighborhood)
Wisconsin Department of Safety and Professional Services Prescription Drug Monitoring Program (PDMP)	Dispensing details, drug details, pharmacy, and provider pseudoIDs, and timestamps for the drug prescription
Wisconsin Department of Health (DHS) Vital Statistics	Death, cause of death, flag for deaths related to alcohol use
DHS Ambulance Run Data System	Patient complaints and provider impressions (ie, agitated, hallucination, hostile, suicidal, and violent). Cardiac arrest events, dispatch text, medications, vital signs, residential and public facility locations of the event, encounter time
Department of Corrections	Incarceration details (prison admission date, release type/date, etc.), facilities resided in, offenses and prior felony convictions, substance use, substance use programming available in prison and program type
Medical and Pharmacy Claims	Open and closed claims for medical hospital payer, patient, provider, diagnosis, procedure, enrollment month, remittance, and pharmacy transactions. Medicare, Medicaid, and Commercial payors included
National Mortality	Social Security Administration death master file, publicly available newspaper obituary feeds

Abbreviations: CUI, concept unique identifiers mapped from the National Library of Medicine Unified Medical Language System.

as a central hub.[15] Notably, the honest broker did not receive, store, or process any protected health information (PHI). The PHI from tokens remained with the data owners, and only the related data attributes, ie, the non-PHI data variables from each owner, were sent along with the tokens. Additionally, no transit token could be linked back to the site token, thereby protecting against the reidentification of any individual.

The honest broker at UW's central hub facilitated study-specific central mapping that represented records linked using the transit tokens from the various data owners (see Table S2 for the list of tokens). The "Net Tokens" matching algorithm was used for the linkages. In this algorithm, a match is identified when comparing all available tokens when the number of matching tokens across 5 attributes (first/last name, address, gender, social security number, and birth date) exceeded the number of nonmatches (ie, majority rules).[16] The algorithm was robust to varying linkage accuracies of the underlying information (ie, when different tokens are available across data contributors) with linkage performance as successful as 97.9% precision and 90.3% recall.[16] The honest broker performed a final crosswalk across datasets to provide the research team with a set of deidentified patient identification numbers as the relational key across the modalities of data.

The final linked dataset underwent quality checks and was organized according to the Kahn Framework.[17,18] A mirrored approach was also followed for state agency data. The framework provided quality checks on conformance, consistency, completeness, and plausibility. These elements were then verified within the site and validated with external benchmarks across sites.

Cloud computing environment

The SMDC is hosted within a HIPAA-secure Microsoft Azure cloud computing environment with the necessary data science libraries and Integrated Development Environments for Python and R software tools. The Azure cloud data environment facilitates the scaling of storage and computing resources as needed by users. The environment is maintained by the Institute for Clinical and Translational Research at UW-Madison and is available to researchers via credentialing and security authorization. The cloud is secure and compliant with the US Department of Health and Human Services, National Institute of Standards and Technology (NIST) 800-53A compliance framework, Federal Information Security Modernization Act moderate, and HIPAA standards. The security environment includes multiple firewalls, intrusion detection, logging, monitoring, and alerting. The system security plan includes over 140 "moderate" NIST 800-53A Security & Privacy controls. Hard-coding firewall rules only allow access to trusted IP addresses for data science software dependencies. Users of the SMDC were required to have credentials through the central hub at UW with 2-factor authorization. All data transfer into and out of the cloud followed a secure protocol using the Globus.[19,20]

Analysis plan

We first examined the study population characteristics and match rates across datasets. We then analyzed the sensitivity of all-cause mortality across the different mortality data sources. In-hospital and out-of-hospital deaths were identified using 3 sources of data: (1) EHR; (2) state Vital Statistics, which collects death certificates with recorded causes of death filed with the state; and (3) a commercial national data source provided by Datavant, which includes deaths from the Social Security Administration's Death Master File augmented with deaths sourced from funeral homes and newspaper feeds, to construct an individual-level database of more than 80% of the US deaths annually.[21,22] Deaths between 2015 and 2020 were analyzed because they were available in all 3 mortality data sources. We used the EHR-deaths as the reference standard to compare the sensitivity and recall measures between the Vital Statistics and the commercial national mortality data source. The underlying cause of death was derived from the death certificates with the state Vital Statistics and the categories of death were modeled from the Center for Disease Control underlying cause of death recodes.[23] Comparisons were made between the 3 data sources across demographics and Area Deprivation Indices (ADI).[24]

Results

Study population

We successfully linked our patients across all data sources in our proof of concept from a single health system. The SMDC includes data linked across 36 522 patients and their 62 594 encounters across 2 UW hospitals. Alcohol misuse was involved in 64.7% ($n = 40$ 471) of the encounter followed by opioid misuse (23.7%, $n = 14$ 838). Our study population had a median age of 48 years (interquartile range [IQR] 33-59), were 38% female ($n = 23$ 596), 3% Hispanic ($n = 2098$), and had a median national ADI of 46 (IQR 32-59) (see Table 2 for detailed encounter-specific characteristics).

Data linkage rates

The match statistics across each dataset are provided in Table S3. Neighborhood-level information from census tract data sources was derived for all patients. We matched a total of 26 425 (72.4% of our cohort) to the PDMP database and 22 691 (62.1%) to the Claims data. For the state's ambulance-run data system, missing data were less than 3% on identifiers with first/last name, gender, date of birth, and ZIP. A total of 20 061 (54.9%) of our cohort were linked to the statewide ambulance database, and 6.6% had duplicate linkages.

Overlap in mortality reporting across data sources

The SMDC study cohort of 35 522 patients had 5640 deaths identified across all 3 death data sources between 2015 and 2020. The UW EHR had the lowest counts of deaths with 3812 (67.6%) followed by the National Mortality database with 4540 (80.5%) and the state Vital Statistics with 4698 (83.3%). The overlap in deaths across all 3 sources was 2770 (49.1%) (Figure 2). The National Mortality dataset had the highest number of deaths that were not captured by the other 2 sources (555, 9.8%), followed by the Vital Statistics (346, 6.1%) and the UW EHR (99, 1.8%) databases.

We matched 3058 deaths between our EHR and National Mortality data. Of these, 99.9% of the differences between the EHR and National Mortality recorded date of death were within 30 days. There were 754 deaths recorded in the EHR data that could not be found in the National Mortality files, which provides 80.2% sensitivity/recall if EHR deaths were the reference standard. National Mortality had an additional 1482 deaths that were outside EHR deaths. Among these, 59

Table 2. Encounter-specific characteristics for the SMDC study population.

Variables	Total number of encounters ($n = 62\ 594$)
Age, years, median (IQR)	48 (33-59)
Female sex, *n* (%)	23 596 (37.7)
Race, *n* (%)	
White or Caucasian	53 393 (85.3)
Black or African American	6891 (11.0)
American Indian or Alaska Native	875 (1.4)
Asian or Middle Eastern	591 (0.9)
Pacific Islander or Hawaiian Native	105 (0.2)
Other	739 (1.2)
Hispanic/Latino, *n* (%)	2098 (3.4)
ADI, median (IQR)	
National rank	46 (32-59)
State rank	4 (2-6)
Substance misuse type, *n* (%)	
Alcohol	40 471 (64.7)
Cannabis	6945 (11.1)
Cocaine	4636 (7.4)
Hallucinogens	302 (0.5)
Opioid	14 838 (23.7)
Psychoactive/other	10 298 (16.5)
Sedative/hypnotic	993 (1.6)
Stimulant	1750 (2.8)
Encounter type, *n* (%)	
Emergency department visit or hospitalization	46 942 (75.0)
Inpatient hospitalization	39 180 (62.6)
Length of stay, days, median (IQR)	2(1-5)
Admission service, *n* (%)	
Emergency service	17 220 (27.5)
General medicine	15 648 (25.0)
Surgery (trauma/specialty/general)	11 840 (18.9)
Other/unknown	17 886 (28.6)
Critical care	2955 (4.7)
Discharge disposition, *n* (%)	
Died in hospital	1387 (2.2)
Discharged alive	59 793 (95.5)
Left against medical advice	1414 (2.3)

(1.3% of total deaths) were admitted to UW Hospital after the date of death recorded within the National Mortality data and were thus corrected in the SMDC. UW Hospital admission dates that were >30 days from the National Mortality date accounted for 85.1% of the patient encounters that had a match to the EHR data.

We matched 3425 deaths between our EHR and Vital Statistics. There were 387 EHR deaths that could not be found in the Vital Statistics dataset, which provides 89.9% sensitivity/recall if EHR deaths were the reference standard. On the matched death dates ($n = 3697$), 99.7% of the differences in timestamps between the EHR and Vital Statistics were within 30 days. Vital Statistics had an additional 1273 deaths that were outside EHR deaths. Of these, 2 were admitted to UW Hospital after the date of death was recorded within Vital Statistics and was corrected in the SMDC. UW Hospital admission dates that were >30 days from the Vital Statistics date accounted for 84.1% of the patient encounters that had a match to EHR data. Vital Statistics, the only data source that provides the underlying cause of death, reported the top 3 causes of death as malignant neoplasm ($n = 1196$, 25.4%), major cardiovascular disease ($n = 766$, 16.3%), and liver disease ($n = 583$, 12.4%). Notably, only 9.2% ($n = 431$) of reported deaths were drug-related overdoses. The deaths recorded in Vital Statistics but not found in the EHR had a higher median national ADI ranking (56 [IQR 40-75] vs 47 [IQR 33-62], $P < .01$).

We matched 3697 deaths between our National Mortality and Vital Statistics datasets. There were 1001 deaths that occurred in Vital Statistics that could not be found in the National Mortality files. There were 843 National Mortality deaths that could not be found in the Vital Statistics files. On the matched death ($n = 3697$), 99.7% of the differences between the National Mortality and Vital Statistics recorded date of death were within 30 days. The distribution of state locations for deaths unique to the obituary feeds within the National Mortality dataset ($n = 759$, out-of-state deaths were available only from the obituary feeds) is shown in Figure S1.

Discussion

In this study, we describe the creation of the SMDC, a first-of-its-kind informatics platform that captures a comprehensive picture of patients at-risk for substance misuse and their encounters with hospitals and health systems, emergency medical services, and pharmacies, as well as the attributes of the neighborhoods where they live. Our approach uses PPRL to link state-level public health data with longitudinal EHR and national mortality datasets, providing an opportunity to fully grasp the incidence and response for *all* conditions related to substance misuse with a focus on all-cause mortality. To the best of our knowledge, the SMDC is the first public–private–academic partnership to adhere to the regulatory, legal, and privacy needs of vulnerable populations and provide a unique environment for addiction research using longitudinal data that span multiple data modalities. Importantly, our proof-of-concept serves as a roadmap for developing a population health-driven framework to foster better engagement and data sharing between health systems and surrounding county and state agencies.

Since 2019, mortality from drug overdoses and alcohol-related causes in the United States has risen 10% per year.[1,2] Furthermore, substance misuse is a leading cause of hospital readmissions.[25] Public health agencies tasked with designing local treatments and interventions have primarily focused on fatal events. Despite this, the substance misuse epidemic has continued unabated, prompting the Office of National Drug Control Policy to issue a recommendation to shift focus to warning signs—ED visits, hospitalizations, or encounters with the legal system—that lie on the path to more severe events.[26] In practice, however, clinical providers and policymakers do not have access to a unified record of information critical to enforce these recommendations.

Other states have executed different models of data linkage to study substance misuse conditions but have inherent limitations. For example, Massachusetts passed new state-level legislation to link data across state agencies to enable data sharing for opiate-related overdose events.[27,28] However, EHR data are only used to record overdose events, and the full breadth of data available in the EHR that captures rich information related to nonoverdose conditions remains unused. Additionally, the data are linked deterministically, ie, only exact matches, and thus could be missing key events. In another study, North Carolina used a combination of probabilistic and deterministic linkage to link multiple data sources for tracking fatal or nonfatal (eg, received naloxone) overdose events but do not consider other substance misuse conditions.[29] Similarly, Maryland

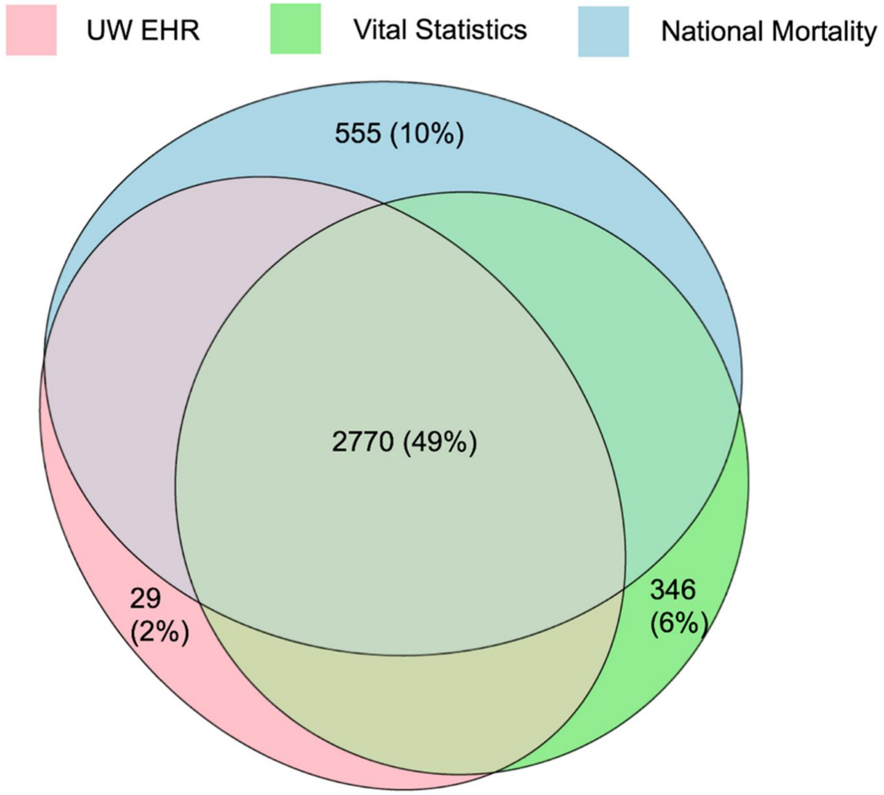

Figure 2. Venn diagram of death data sources with nonoverlap contributions.

studied unintentional overdose deaths (derived from the Office of the Chief Medical Examiner) and used probabilistic methods to link data with incarceration and parole records, ED visits, and hospitalizations for overdose from hospital discharge records, social services, and cases from juvenile services.[30] However, their focus remains on linking data for overdose deaths only and cannot be used to study substance misuse-related events leading up to death. Outside of the United States, a study set in New South Wales, Australia linked hospital and ED data with ambulance, cancer registry, court and incarceration records, and a national death database in a probabilistic manner to describe individuals with problematic alcohol use.[31] Our study adds to this body of work by creating a broad repository with the inclusion of more EHR data on all physical ailments and conditions for patients with recorded misuse of alcohol, opioids, stimulants, depressants, and illicit drugs. Our population health systems approach to addressing the substance misuse epidemic offers regional hospitals and public health departments leverage and information to implement evidence-based strategies to reduce preventable substance use-related hospitalizations and deaths. Our approaches can be translated to other regions and states where the substance misuse epidemic continues unabated. Notably, using our codes and data linkage methods can reduce the time to establish a data commons to improve the strategies for region-specific prevention, treatment, and harm reduction.

Our study also highlights methods to create a new framework for secure and private data sharing. Ethical, legal, and organizational reasons often prevent the sharing of sensitive data by medical and service providers with outside collaborators and peers. Data security is becoming more crucial for patients with substance misuse as extra protection is needed for this vulnerable population. The PPRL approach addresses these challenges and enables secure, large-scale data sharing.[15] Similar approaches have been used recently to democratize EHR data in national data networks (eg, PCORnet and the National Covid Cohort Collaborative).[15,32] We exercise further precautions by using the services of an honest broker who is positioned independently in the university and is not part of the commercial data linkage service company. Overall, we were able to acquire non-EHR information related to substance misuse for over half of our EHR-based cohort. Our study is the first to utilize PPRL and honest-broker services for linking a health system with state agencies to build a data commons for substance misuse research.

Our approach also highlights the utility of using multiple mortality datasets to capture both in-hospital and out-of-hospital deaths in hospitalized patients with substance misuse. The SMDC allows health systems, state health departments, and public policymakers to follow the NDCC recommendations to focus on nonoverdose-related events for substance misuse prevention and treatment. Mortality was high among hospitalized patients with substance misuse, and chronic conditions were the top recorded causes of death. With less than 10% of deaths attributable to overdose, more focus is needed to prevent hospitalizations for physical ailments related to substance use. Further, our multimodal approach to capturing death data enriched the capture of fatal events and highlighted important disparities between data sources. The higher ADI in the state vital records of deaths not found in the EHR likely represent the uninsured and impoverished population, which are major risk factors for substance misuse.[33] The use of obituary feeds also helped identify out-of-state deaths not routinely recorded by state Vital Statistics.

Our open multiuser data user agreement was designed to foster new research on a sensitive population. The data science service capabilities within our secure cloud infrastructure enable data analytics and machine learning to be conducted by researchers. We are acquiring data from the Department of Corrections (the data use agreement is complete), bringing important criminal justice information to our population. Additionally, our hospital has built a high-throughput pipeline to convert all clinical documents into standardized medical vocabularies for creating deidentified features from text within all clinical notes and imaging reports for each hospitalization.[34,35] The availability of text-based features further expands the utility of the SMDC for cohort identification.[36–39]

Our efforts in creating the SMDC have several limitations. First, the SMDC only contains information from one health system. Expanding to other urban and rural regional health systems will increase the size and diversity of our patient population. Second, all data sources do not contribute data for the same period. For example, the EHR data covers the years 2008-2022, whereas the ambulance data begins in 2016. However, substance misuse rates tend to depend on more recent longitudinal trends in predictors and thus have a short time lag. Finally, while more comprehensive than any existing data repository, the SMDC is missing data elements, such as medical examiner data and toxicology reports, or is limited to available information within each data source, such as medical and pharmacy insurance payors.

Conclusion

The SMDC is a comprehensive data resource that combines data from hospitals, public health agencies, and first responder agencies, allowing for better identification and prioritization of care for Wisconsin's most vulnerable residents. The SMDC will support collaborations that bring together expertise in biomedicine, public health and epidemiology, data management, and data science to make patient data useful and usable for curbing the substance misuse epidemic. Further, the SMDC offers key insights into the integration of multiple data sources to improve region-specific prevention, treatment, and harm reduction of substance misuse through a privacy-preserving approach.

Acknowledgments

The authors would like to specially thank Gabriel McMahan and the University of Wisconsin-Madison Clinical and Health Informatics Institute (CHI^2) and the Wisconsin Department of Health Services. The authors would also like to specially thank UW team members, Lisa Wilson, Andy Chen, and Ryan Moze, for their council on regulatory, legal, contractual, and privacy strategy that helped shape our data sharing agreements and IRB protocols, and protect our SMDC data.

Author contributions

M.A., A.M., and M.O. contributed to the study's conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, and/or writing of this manuscript. T.A.C., J.M., C.M., E.N.-N., J.C., and H.N.-H. contributed to data curation, methodology, project administration, and software. T.G., P.G., C.S., and A.S.A. contributed to data curation, investigation, methodology, validation, and visualization. M.M.C., R.W., E.S.-A., M.G., and B.P. contributed to investigation, visualization, and writing of this manuscript.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work and M.A. and M.G., T.G., and M.O. were supported by NIH/NIDA (grant number R01DA051464-S1). P.G. was supported by NIH/NHLBI (grant number 5T32HL144909-04). A.M. was supported by NIH/NHLBI (grant number K01HL14 8390).

Conflicts of interest

None declared.

Data availability

The data underlying this article cannot be shared publicly to protect the privacy of the patients included in this study, and due to ethical and legal restrictions imposed by our data contributors. This project contains very restricted, sensitive (health, substance use) data that was provided by UW Health EHR, WI Department of Health Services Emergency Medical Service data, WI Department of Health Services Vital Records Data, Datavant Incorporated Mortality data, and the WI Department of Safety and Professional Services Prescription Drug Monitoring Program. Our data providing entities have data transfer and use agreements (DTUAs) executed with UW-Madison and the DTUAs restrict the data to UW's environment and have very specific restrictions on data sharing/reporting of use. Please contact the corresponding author M.A. (majid.afshar@wisc.edu) or M.O. (mkoguss@medicine.wisc.edu) for further details about accessing the data.

References

1. Spencer MR, Minino AM, Warner M. Drug overdose deaths in the United States, 2001-2021. *NCHS Data Brief*. 2022;(457):1-8.
2. Yeo YH, He X, Ting PS, *et al*. Evaluation of trends in alcohol use disorder-related mortality in the US before and during the COVID-19 pandemic. *JAMA Netw Open*. 2022;5(5):e2210259.
3. Townsend T, Kline D, Rivera-Aguirre A, *et al*. Racial/ethnic and geographic trends in combined stimulant/opioid overdoses, 2007-2019. *Am J Epidemiol*. 2022;191(4):599-612.
4. US Department of Health and Human Services. Drug Abuse Warning Network, 2004: National Estimates of Drug-Related Emergency Department Visits. *PsycEXTRA Dataset*. 2006.
5. Larney S, Tran LT, Leung J, *et al*. All-cause and cause-specific mortality among people using extramedical opioids: a systematic review and meta-analysis. *JAMA Psychiatry*. 2020;77(5):493-502.
6. Lewer D, Tweed EJ, Aldridge RW, Morley KI. Causes of hospital admission and mortality among 6683 people who use heroin: a

cohort study comparing relative and absolute risks. *Drug Alcohol Depend*. 2019;204:107525.

7. Larney S, Randall D, Gibson A, Degenhardt L. The contributions of viral hepatitis and alcohol to liver-related deaths in opioid-dependent people. *Drug Alcohol Depend*. 2013;131(3):252-257.
8. Schranz AJ, Fleischauer A, Chu VH, Wu LT, Rosen DL. Trends in drug use-associated infective endocarditis and heart valve surgery, 2007 to 2017: a study of statewide discharge data. *Ann Intern Med*. 2019;170(1):31-40.
9. Gupta R, Holtgrave DR. US tracking system for nonfatal drug overdoses—reply. *JAMA* 2022;328(20):2068-2069.
10. Grossman RL, Heath A, Murphy M, Patterson M, Wells W. A case for data commons: toward data science as a service. *Comput Sci Eng*. 2016;18(5):10-20.
11. Wilson S, Fitzsimons M, Ferguson M, *et al.*; GDC Project. Developing cancer informatics applications and tools using the NCI genomic data commons API. *Cancer Res*. 2017;77(21):e15-e18.
12. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
13. Kindig D, Stoddart G. What is population health? *Am J Public Health*. 2003;93(3):380-383.
14. Owens PF, McDermott KW, Muhuri PK, Heslin KC. Inpatient stays involving mental and substance use disorders, 2016. In: *HCUP Statistical Brief #249*. Rockville, MD: Agency for Healthcare Research and Quality; 2019.
15. Kiernan D, Carton T, Toh S, *et al.* Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet((R)), the National Patient-Centered Clinical Research Network. *BMC Res Notes*. 2022;15(1):337.
16. Bernstam EV, Applegate RJ, Yu A, *et al.* Real-world matching performance of deidentified record-linking tokens. *Appl Clin Inform*. 2022;13(4):865-873.
17. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care*. 2012;50 Suppl(0):S21-S29.
18. Kahn MG, Callahan TJ, Barnard J, *et al.* A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)*. 2016;4 (1):1244.
19. Foster I. Globus online: accelerating and democratizing science through cloud-based services. *IEEE Internet Comput*. 2011;15 (3):70-73.
20. B Allen JB, Childers L, Foster I, *et al.* Software as a service for data scientists. *Commun ACM*. 2012;55(2):81-88.
21. Wallace J, Lollo A, Ndumele CD. Evaluation of the association between medicare eligibility and excess deaths during the COVID-19 pandemic in the US. *JAMA Health Forum*. 2021;2(9):e212861.
22. Zhang Q, Gossai A, Monroe S, Nussbaum NC, Parrinello CM. Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. *Health Serv Res*. 2021;56(6):1281-1287.
23. Anderson RN, Minino AM, Hoyert DL, Rosenberg HM. Comparability of cause of death between ICD-9 and ICD-10: preliminary estimates. *Natl Vital Stat Rep*. 2001;49:1-32.
24. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible – the neighborhood atlas. *N Engl J Med*. 2018;378(26):2456-2458.
25. Owens PL, Fingar KR, McDermott KW, Muhuri PK, Heslin KC. Inpatient stays involving mental and substance use disorders, 2016: statistical brief #249. In: *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*. Rockville, MD: Agency for Healthcare Research and Quality; 2019.
26. Gupta R, Holtgrave DR. A national tracking system for nonfatal drug overdoses. *JAMA* 2022;328(3):239-240.
27. Evans EA, Delorme E, Cyr KD, Geissler KH. The Massachusetts public health data warehouse and the opioid epidemic: a qualitative study of perceived strengths and limitations for advancing research. *Prev Med Rep*. 2022;28:101847.
28. Larochelle MR, Bernson D, Land T, *et al.* Medication for opioid use disorder after nonfatal opioid overdose and association with mortality: a cohort study. *Ann Intern Med*. 2018;169(3):137-145.
29. Becker WC, Heimer R, Dormitzer CM, *et al.* Merging statewide data in a public/university collaboration to address opioid use disorder and overdose. *Addict Sci Clin Pract*. 2021;16(1):1.
30. Cherico-Hsii S, Bankoski A, Singal P, *et al.* Sharing overdose data across state agencies to inform public health strategies: a case study. *Public Health Rep*. 2016;131(2):258-263.
31. Peacock A, Chiu V, Leung J, *et al.* Protocol for the Data-Linkage Alcohol Cohort Study (DACS): investigating mortality, morbidity and offending among people with an alcohol-related problem using linked administrative data. *BMJ Open* 2019;9(8):e030605.
32. Haendel MA, Chute CG, Bennett TD, *et al.*; N3C Consortium. The national COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021;28(3):427-443.
33. Doran KM, Rahai N, McCormack RP, *et al.* Substance use and homelessness among emergency department patients. *Drug Alcohol Depend*. 2018;188:328-333.
34. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis And Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513.
35. Afshar M, Dligach D, Sharma B, *et al.* Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. *J Am Med Inform Assoc*. 2019;26(11):1364-1369.
36. Afshar M, Joyce C, Dligach D, *et al.* Subtypes in patients with opioid misuse: a prognostic enrichment strategy using electronic health record data in hospitalized patients. *PLoS One* 2019;14(7): e0219717.
37. Afshar M, Phillips A, Karnik N, *et al.* Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *J Am Med Inform Assoc*. 2019;26(3):254-261.
38. Afshar M, Sharma B, Bhalla S, *et al.* External validation of an opioid misuse machine learning classifier in hospitalized adult patients. *Addict Sci Clin Pract*. 2021;16(1):19.
39. Afshar M, Sharma B, Dligach D, *et al.* Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI): a retrospective deep learning study. *Lancet Digit Health*. 2022;4(6):e426-e435.