frontiers | Frontiers in Artificial Intelligence

TYPE Systematic Review
PUBLISHED 05 April 2024
DOI 10.3389/frai.2024.1208874

Check for updates

OPEN ACCESS

EDITED BY
Ramesh Kandimalla,
Indian Institute of Chemical Technology (CSIR), India

REVIEWED BY
Katja Mayer,
University of Vienna, Austria
Shawn Dolley,
Open Global Health, United States

*CORRESPONDENCE
Olga Scrivner
✉ obscrivn@indiana.edu

RECEIVED 21 May 2023
ACCEPTED 29 January 2024
PUBLISHED 05 April 2024

CITATION
Scrivner O, Nguyen T, Ginda M, Simon K and Börner K (2024) Interactive network visualization of opioid crisis research: a tool for reinforcing data linkage skills for public health policy researchers.
*Front. Artif. Intell.* 7:1208874.
doi: 10.3389/frai.2024.1208874

# Interactive network visualization of opioid crisis research: a tool for reinforcing data linkage skills for public health policy researchers

Olga Scrivner[1,2]*, Thuy Nguyen[3], Michael Ginda[1], Kosali Simon[4,5] and Katy Börner[1]

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, United States, [2]Rose-Hulman Institute of Technology, Terre Haute, IN, United States, [3]School of Public Health, University of Michigan, Ann Arbor, MI, United States, [4]O'Neill School of Public and Environmental Affairs, Indiana University, Bloomington, IN, United States, [5]National Bureau of Economic Research, Cambridge, MA, United States

**Background:** Public health policy researchers face a persistent challenge in identifying and integrating relevant data, particularly in the context of the U.S. opioid crisis, where a comprehensive approach is crucial.

**Purpose:** To meet this new workforce demand health policy and health economics programs are increasingly introducing data analysis and data visualization skills. Such skills facilitate data integration and discovery by linking multiple resources. Common linking strategies include individual or aggregate level linking (e.g., patient identifiers) in primary clinical data and conceptual linking (e.g., healthcare workforce, state funding, burnout rates) in secondary data. Often, the combination of primary and secondary datasets is sought, requiring additional skills, for example, understanding metadata and constructing interlinkages.

**Methods:** To help improve those skills, we developed a 2-step process using a scoping method to discover data and network visualization to interlink metadata. Results: We show how these new skills enable the discovery of relationships among data sources pertinent to public policy research related to the opioid overdose crisis and facilitate inquiry across heterogeneous data resources. In addition, our interactive network visualization introduces (1) a conceptual approach, drawing from recent systematic review studies and linked by the publications, and (2) an aggregate approach, constructed using publicly available datasets and linked through crosswalks.

**Conclusions:** These novel metadata visualization techniques can be used as a teaching tool or a discovery method and can also be extended to other public policy domains.

KEYWORDS

**public health policy, network visualization, data exploration, data linkage, data visualization, skills**

## 1 Introduction

The U.S. opioid epidemic is a major national concern. As of April 2023, the 12-month counts of reported deaths from drug overdose have increased by an estimated 6.5% compared with the year 2021—rising from 99,782 to 106,275 deaths (Ahmad et al., 2023). Particularly, the overdose involving synthetic opioids increased from 17 to 21.8%

from 2020 through 2021, according to the National Center for Statistics (Spencer et al., 2022). Furthermore, among the 40.7 million people with a substance use disorder, only 1.1% (447,000) received treatment, 2.1% (837,000) felt the need but did not get treatment, and 96.8% felt no need for treatment (Substance Abuse and Mental Health Services Administration (SAMHSA), 2022). Some of the commonly reported reasons for not seeking treatment were (1) no health coverage, (2) not finding the program, (3) the perception of negative effects on their jobs, and (4) not knowing where to go.

To address the current opioid crisis, the Department of Health and Human Service (HHS)'s strategic priorities includes improvements in (1) pain management, (2) prevention, treatment, and recovery, (3) data and research related to the opioid crisis, and (4) overdose-reversing drugs (Price, 2017). It is crucial to integrate a holistic approach across multiple data resources, covering drug policy, pharmacy claims, treatment workforce, and opioid-related harms, among other research topics. Identifying trends and insights in this complex data presents a challenge and requires data analysis skills. Network visualization has been shown as a useful technique for analyzing complex relations in medical care services, clinical data, and physician network (Niyirora and Aragones, 2020). Many network models have included co-occurrences of records (e.g., patient diagnoses), chronological sequential occurrences (patient's admissions or discharge), and source and target occurrences (e.g., patient transfers). A recent study has extended network visualization to opioid prescription data, providing insights for healthcare professionals on the inappropriate use of drugs (Hu et al., 2020). While data analytical skills have been increasingly introduced to health policy programs, there is still a gap in learning network analysis skills (Payán, 2021).

Similarly, there is a growing body of systematic reviews and scientometric meta-analyses focusing on relationships between opioid use disorder and various factors, such as chronic pain, intervention, mitigation strategies, and policies (Chou et al., 2021; Gamage et al., 2023). Furthermore, a systematic meta-analysis of datasets unveils additional insights into available resources and their interlinkages. Stakeholders and practitioners are often challenged by the large number, complexity, and peculiarities of the existing data. Researchers may also not be aware of available resources as they are provided by many different organizations and have varying data quality and coverage. Some datasets are freely available, while others require the signing of legal documents or payment of fees for additional fields. Furthermore, some datasets are massive in size, requiring database expertise to run queries; others exist only as textual data in a PDF format and require pre-processing skills before usage. Providing data meta-analysis has become a new "informational asset" transforming how we observe and analyze data (Weber et al., 2014). In addition, linking resources together (or crosswalks) allows researchers and stakeholders to identify new areas for public or health interventions and provide evidence-based guidelines for practitioners and patients (Smart et al., 2018).

There is a growing need for metadata skills to help develop data strategies for identifying and linking resources. We propose a two-step framework to facilitate metadata discovery and relationships between datasets using metadata network visualization. In the first step, we show how to gather relevant metadata using the modified systematic review method. In the second step, we use a network design to represent linked datasets that communicate temporal, geospatial, and topical coverage via metadata nodes. This metadata visualization provides an alternative way to identify and integrate opioid-related datasets.

## 2 Background

The causes, consequences, and manifestations of the U.S. opioid crisis have been studied from many different angles, including prevention, treatment, drug prescription, law enforcement, criminal justice, and overdose reversal. Treatment expansions and prescription reductions are two essential steps in reducing mortality and improving safety for patients with chronic pain. Monitoring and regulatory policies play an equally important role in balancing between harms, cost, availability, and benefits of opioid use, as seen in policies such as prescription drug monitoring programs (PDMPs), health insurance expansions, and comprehensive federal legislation (e.g., the Comprehensive Addiction and Care Act) (Poitras, 2018; Scrivner et al., 2020). These efforts have led to a decrease in the overall U.S. drug prescription rate, from 81.3 per 100 people in 2012 to 46.7 in 2019 (CDC, 2019). But while the U.S. has had success in implementing these preventative measures, there has been an increase in harm from illicit drug sources, and there has been a challenge in improving treatment access for those suffering from addiction disorders. A major gap remains between service demand and supply: 94% of people aged 12 or older with a substance use disorder did not receive any treatment, according to the 2021 National Survey on Drug Use and Health data (NSDUH). The 2020 report on admissions to substance use treatment facilities (TEDs) has also reported a decrease in opioid-related admissions (381,040), as compared to 677,296 admissions in 2018. In terms of the number of facilities, only 1,754 out of 16,066 treatment facilities are Opioid Treatment Program (OTP) certified (Substance Abuse and Mental Health Services Administration, 2021). The 2021 County Business Patterns data (CBP) identifies 14,461 Substance Use Disorder Treatment (SUDT) outpatient centers, 44,731 Residential SUDT facilities, and 795 SUDT hospitals (API link—https://data.census.gov/table?q=CBP2021.CB2100CBP&n=62142:6222:6232&tid=CBP2021.CB2100CBP). Despite the high priority for training expressed by the U.S. Department for Health and Human Service and high job demand, the historical behavioral health (integrated mental and substance use disorder) workforce shortage has been a major roadblock (Skillman et al., 2016; McNeely et al., 2021).

The interdependence of these social, health, economic, and public policy factors calls for an interdisciplinary holistic and systematic approach where researchers and practitioners can zoom out and examine the problem as a whole and then zoom in to solve the most pressing issues that have the highest positive impact on improving health and services while decreasing crime and addictions-related disorders. One of the approaches is to provide a systematic review of opioid-related studies along with the secondary data relevant to the research (Leece et al., 2019; Maclean et al., 2020; Smart et al., 2020). This perspective enables researchers to discover data, identify new connections (linkages) between existing data, and learn about data accessibility and coverage. These systematic review studies also offer various perspectives on grouping datasets. From the economic perspective, data can be classified into several data categories relevant to understanding the opioid crisis: (1) pharmaceutical industries and medication prescriptions, (2) healthcare providers and labor market, (3) harms and crime, and (4) policies (Maclean et al., 2020). From the treatment perspective, data can be grouped by (1) intervention variables (e.g., prevention, treatment, and harm reductions) and (2) enabling variables (e.g., surveillance and stigma) (Leece et al., 2019). From the strategic perspective, data can be categorized according to the Health and Human Service strategic priorities: (1) better pain management, (2) addiction prevention, treatment, and recovery service, and (3) better targeting of overdose-reversing drugs (Smart et al., 2018, 2020). In addition, data can be classified based on type and format: national surveys, electronic health records (EHR), claims data, mortality records, prescription monitoring data, contextual and policy data, and others (national, state, local) (Smart et al., 2018, 2020).

In recent years, it also became common to share data with metadata, including data description, coverage, and attributes (Wu et al., 2023). As each dataset provides its unique identifier (e.g., geographical units, drug names, occupation, or billing codes), it is essential to (1) identify a crosswalk, an identifier that can link records to other datasets; (2) data coverage, such as number of records, frequency of updates, and data interval or data units (e.g., monthly and quarterly); and (3) data accessibility (e.g., open data or contract data). Understanding linkages and metadata becomes even more critical as many new datasets are released (Shlomo, 2018; Blanco et al., 2021). There are various ways to represent metadata, e.g., a tabular format or a dictionary schema. This representation, however, does not include the assessment of data coverage, its weaknesses, or its strengths. Novel solutions are offered by recent systematic dataset overviews: (1) each variable is provided with its relative frequency of occurrence in the reviewed literature (Leece et al., 2019); (2) a plus/minus sign is used to indicate strengths and weaknesses for each dataset (Smart et al., 2020); and (3) a "probabilistic linkage", focusing on a visual representation of potential biomedical sources and the values of their linkages (Weber et al., 2014). The latter approach involves the use of a tabular form with sizes, shapes, colors, and positions to indicate data quality, data linkage, types of data (e.g., pharma, claims, EHR, and non-clinical data), data coverage, and even the probabilities for obtaining new data or linking existing data.

Furthermore, recent work on data integration and federation demonstrates advances in ontology and knowledge graph-based approaches allowing for integration, querying, analysis, and visualization across heterogeneous data sources (Sima et al., 2019; SN SciGraph, 2019; Cox et al., 2020; Amer-Yahia et al., 2021; Morris et al., 2023). For example, SPOKE (Morris et al., 2023) and Springer Nature SciGraph (SN SciGraph, 2019) use a knowledge graph (KG) to interlink and query different datasets. The SPOKE KG interlinks more than 30 publicly available biomedical databases, whereas SciGraph interlinks funders, projects, publications, citations, and scholarly metadata in support of data exploration. In addition, a natural language querying and visualization tool for biological knowledge is implemented for heterogeneous data sources (Sima et al., 2019). The INODE project (Amer-Yahia et al., 2021) incorporates machine learning techniques in support of guided, natural language querying and visualization of semantically integrated data sources in bio-medicine, astrophysics, and public policy.

Therefore, the deployment of visualization techniques emerges as a powerful data discovery tool and can be used to communicate metadata (data coverage, quality, and linkage). In addition, the graph representation not only facilitates a more intuitive understanding of complex datasets but also provides a unique resource to illustrate interlinkages between heterogeneous datasets, offering a more insightful perspective for data analysis than traditional data repositories.

## 3 Methods

To build metadata network visualization, we designed a two-step process enabling data discovery and synthesis in the first step and network graph implementation in the second step. This methodical process is not only reproducible but also adaptable, allowing for its application across various datasets and topics, thereby extending its utility and scope in diverse research contexts.

### 3.1 Data collection

Several recent systematic reviews on opioid-related studies include primary and secondary data. To collect metadata, we applied the modified scoping method by extracting and filtering studies from these systematic reviews. To represent a diverse collection of existing datasets, we included the following perspectives: (1) an economic systematic review with opioid-related datasets (Maclean et al., 2020), (2) pain management (Phillips et al., 2017), and (3) data sources for research and evaluation to address the Department of Health and Human Services (HHS) strategy combating opioid crisis (Smart et al., 2018, 2020). In addition, we parsed the scoping review references describing over 100 major economic studies on the U.S. opioid crisis (Maclean et al., 2020).

Using the scoping review protocol (Arksey and O'Malley, 2005), we identified 176 cited papers (see Figure 1). Specifically, we established the following pipeline: (1) Importing—we imported the 176 cited articles ranging from 1986 to 2020 to the bibliographical software Mendeley. (2) Scanning—each article was scanned for datasets mentioned in the methodology section and articles without datasets were discarded. (3) Tagging—the remaining set (107 articles) was tagged with dataset names as they were used in the studies. As a result, we identified 283 unique name tags. Across the 107 studies, there were many inconsistencies in naming and spelling, for instance, "nvss," "nvss multiple cause of death," and "nvss multiple cause-of-death mortality" all referred to the U.S. mortality data from death certificates, produced by the National Center for Health Statistics. We normalized labels using OpenRefine and the Nearest Neighbor algorithm with Prediction by Partial Matching (PPM) distance (Stephens, 2018). The

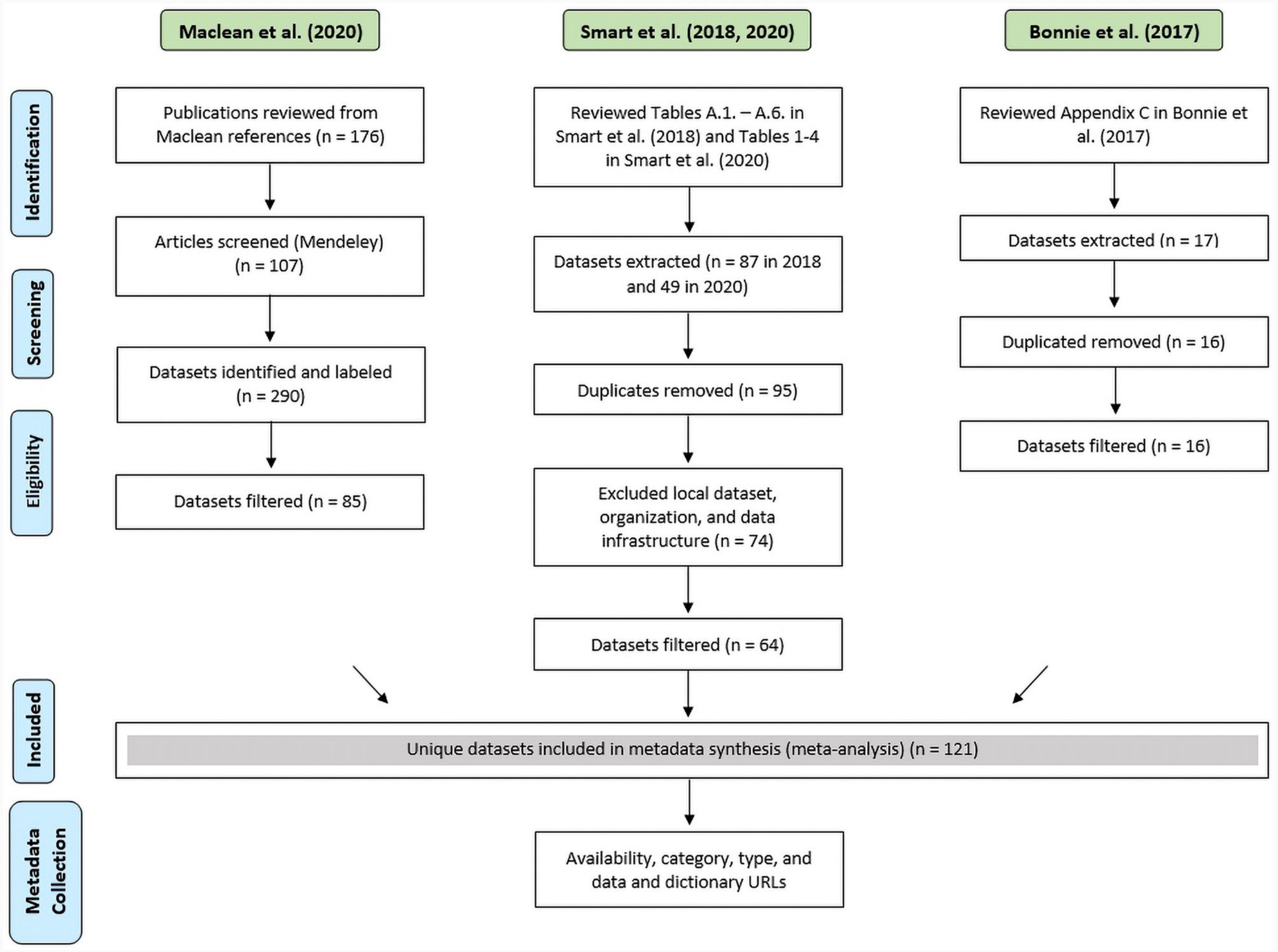

Maclean et al. (2020)
Smart et al. (2018, 2020)
Bonnie et al. (2017)
Identification
Screening
Eligibility
Included
Metadata Collection
Publications reviewed from Maclean references (n = 176)
Articles screened (Mendeley) (n = 107)
Datasets identified and labeled (n = 290)
Datasets filtered (n = 85)
Reviewed Tables A.1. – A.6. in Smart et al. (2018) and Tables 1-4 in Smart et al. (2020)
Datasets extracted (n = 87 in 2018 and 49 in 2020)
Duplicates removed (n = 95)
Excluded local dataset, organization, and data infrastructure (n = 74)
Datasets filtered (n = 64)
Reviewed Appendix C in Bonnie et al. (2017)
Datasets extracted (n = 17)
Duplicated removed (n = 16)
Datasets filtered (n = 16)
Unique datasets included in metadata synthesis (meta-analysis) (n = 121)
Availability, category, type, and data and dictionary URLs

FIGURE 1
PRISMA flow diagram of the reference review process to identify datasets and data synthesis. Green labels are review studies, and blue labels are data collection steps.

algorithm detected 61 clusters that were merged, resulting in 230 normalized labels. We manually inspected all labels and removed datasets that did not fit our eligibility criteria. The following filters were applied: (1) organization/agency sources without a reference to data (e.g., Bureau of Labor Statistics) [10 datasets removed], (2) only local data sources (e.g., Massachusetts All-Payer Claims Database) [21 datasets removed], (3) duplicate sources (e.g., Center for Disease Control and Prevention WONDER Multiple Cause of Death and National Vital Statistics System Multiple-Cause-of-Death files) [two datasets removed]. We classified the obtained datasets into the following categories as suggested by the policymakers experts we consulted: (1) pharmaceutical data–related to opioid prescription, (2) policy data–related to state drug laws, (3) opioid overdose data–related to treatment and treatment results, and (4) employment data–related to training and hiring in the substance use disorder treatment industry (SUDT). These datasets were then combined with the sources provided in Phillips et al. (2017) and Smart et al. (2020) summary tables. As a result, we identified 121 unique datasets extracted from prior scoping reviews for synthesis and data linkage exploration (see Table 1).

## 3.2 Data synthesis

The data synthesis stage included gathering specific information from each identified dataset: (1) data description (dictionary, availability, category, and format) and (2) publication linkages. For each dataset, we assigned a format type, namely, national surveys, contextual data, and claims (Smart et al., 2018). National surveys are datasets that come from surveys conducted on a national sample. Contextual and policy data are datasets collected to analyze policy and policy changes. Claims and EHRs are datasets that include information on patient-level claims data for reimbursement and patients' health records. All remaining categories were grouped into *others*. The second taxonomy was the category, which is based on whether the dataset contains policy, pharmaceutical, opioid, or job-related data (see Tables 1, 2).

In addition, for each dataset, we searched for a data download link and a dictionary, which provides valuable information about data content and format. For some datasets, one or both of the URLs were not available. As a result, we provide at least one URL for 113 datasets and both URLs for only 33 of the 121 datasets. We were also unable to compile the crosswalks between all 121

TABLE 1 Dataset format and category by availability.

| | Public | Non-public | Totals |
|---|---|---|---|
| **Panel 1: dataset format** | | | |
| National surveys | 17 (74%) | 6 (26%) | 23 |
| Contextual and policy data | 15 (94%) | 1 (6%) | 16 |
| Claims and EHRs | 8 (38%) | 13 (62%) | 21 |
| Other | 36 (59%) | 25 (41%) | 61 |
| Totals | 76 (63%) | 45 (37%) | 121 |
| **Panel 2: dataset category** | | | |
| Harms | 29 (62%) | 18 (38%) | 47 |
| Jobs | 22 (71%) | 9 (29%) | 31 |
| Pharma | 11 (41%) | 16 (59%) | 27 |
| Policy | 14 (88%) | 2 (12%) | 16 |
| Totals | 76 (63%) | 45 (37%) | 121 |

The non-public category includes private and data available by request.

TABLE 2 Dataset category aggregated by three authors.

| Dataset category | Smart | Maclean | Bonnie | Totals |
|---|---|---|---|---|
| Harms | 28 (39%) | 31 (44%) | 12 (17%) | 71 |
| Jobs | 5 (15%) | 29 (85%) | 0 (0%) | 34 |
| Pharma | 18 (51%) | 14 (40%) | 3 (9%) | 35 |
| Policy | 11 (48%) | 11 (48%) | 1 (4%) | 23 |
| Totals | 62 (38%) | 85 (52%) | 16 (10%) | 163 |

datasets as some data are not available publicly. As a result, we manually assigned the following attributes to datasets: (1) data description (dictionary, size, category, and time coverage), (2) data linkages, and (3) scholarly metadata (relevant publications). Size was determined as the number of records based on the most recent year and split into three commonly used set sizes: (1) small (<10,000), (2) medium-sized (between 10,000 and 1,000,000), and (3) large (1,000,000 or greater). Note that the choice of the split threshold was arbitrary and based on the row number instead of the storage size as we did not have access to physical copies for each dataset. Time coverage provides information on the year when the dataset became available and the most recent data available for download. Several data attributes are used to identify data linkages: geographical units (e.g., state and county) and standard crosswalks [e.g., the North American Industry Classification System (NAICS), Drug Name]. Finally, for each dataset, we identified three recent publications using the Web of Science to illustrate research results derived from that data—this is not meant to be exhaustive but rather to show a starting point for a researcher looking into a new dataset. In total, 16 variables are provided for each dataset: common abbreviation, full name, data description, dataset category, source URL, dictionary URL, the number of records per year (most recent), size, time coverage (year-start and year-end), size, geo units, crosswalks, and three publications. Given the data accessibility restrictions, we were unable to assign these attributes to private or restricted datasets. As a result and after consulting with health policy experts, we created a subset providing all 16 variables for each dataset (see Table 3).

TABLE 3 Datasets to support research on the opioid crisis.

| Dataset | Description | Category |
|---|---|---|
| CDC mortality | CDC Opioid Overdose Rate | Harms |
| TEDS-A | Treatment Episode Dataset: Admissions | Harms |
| NHIS | National Health Interview Survey | Harms |
| NSDUH | National Survey on Drug Use and Health | Harms |
| NAMCS | National Ambulatory Medical Care Survey | Harms |
| AHRF | Area Health Resources Files | Jobs |
| N-SSATs | National Survey of Substance Abuse Treatment Services | Jobs |
| QCEW | Quarterly Census of Employment and Wages | Jobs |
| CPS | Current Population Survey | Jobs |
| CBP | County Business Patterns | Jobs |
| ACS | American Community Survey | Jobs |
| MEPS | Medical Expenditure Panel Survey | Pharma |
| Sunshine Act | Open Payments | Pharma |
| SDUD (Medicaid) | State Drug Utilization Data | Pharma |
| Medicare* | Medicare Part D Prescription Drug Event | Pharma |
| ARCOS* | Automated Reports and Consolidated Ordering System | Pharma |
| CDC Prescription | CDC Drug Prescription | Pharma |
| PDMP | Prescription Drug Monitoring Program | Pharma |
| PDAPS | Prescription Drug Abuse Policy System | Policy |
| NAMSDL | National Alliance for Model State Drug Laws | Policy |

Twenty datasets are extracted from the review study (Maclean et al., 2020) datasets marked with * require a request submission prior to download.

## 3.3 Network visualization

Network visualizations are widely used to capture the relationship between entities (e.g., co-authorship networks or gene-disease networks). These visualizations represent entities as nodes and their connections as edges, arranged in layouts that depict the overall connectivity structure and clusters while minimizing edge crossings. Networks can be derived from tabular data, such as the creation of a co-author network from a dataset containing the information on papers and the respective authors per paper. Co-author links connect all authors who appear together in a paper, creating an undirected weighted network (Emmert-Streib et al., 2018). Furthermore, nodes and edges within these networks can be enriched with additional visual cues, such as color or size coding, to

TABLE 4 Nodelist used in the presented network visualization, partial—only four of 13 attributes are shown.

| ID | Format | Category | Availability |
|---|---|---|---|
| ACS | National surveys | Jobs | Public |
| Maclean2020 | Author | Author | (blank) |
| Smart2020 | Author | Author | (blank) |

TABLE 5 Edgelist used in the presented network visualization.

| Source | Target | Network type | ID | Weight |
|---|---|---|---|---|
| Maclean2020 | ACS | Undirected | 17 | 1 |
| Smart2020 | ACS | Undirected | 82 | 1 |

highlight supplementary attributes. This can include characteristics such as the number of papers, the number of citations, the year of first publication, publication sources, and thematic categorization, enhancing the information conveyed by the visualization.

We developed two network visualization prototypes offering various metadata discovery perspectives: (1) citation network, offering potential insights into the frequency of data usage, and (2) conceptual network, providing insights into data relevancy. The design included the following pipeline: (1) data transformation, (2) network layout, and (3) interactive deployment. First, we transformed the CSV file with 121 datasets (rows) and 13 attributes (columns) into two distinct files: a "nodelist" and an "edgelist." The nodelist included an additional identifier for each dataset, which was utilized in the edgelist to illustrate the connections between datasets. For instance, the ACS dataset was assigned the ID "ACS," while the papers were labeled "Maclean2020" and "Smart2020" (refer to Table 4 for details). Since the ACS dataset was referenced in both papers, we established linkages from ACS (as the source) to Maclean2020 (as the target) and vice versa. Similarly, we established linkages from ACS (as the source) to Smart2020 (as the target) and vice versa, given that the network is an undirected graph. The resulting network consisted of 125 nodes categorized into five distinct groups and featured 203 edges (the example is outlined in Table 5).

Next, we used the Force Atlas 2 algorithm in Gephi (Bastian et al., 2009) (see Figure 2). Datasets were then grouped by paper and category with the 34 datasets mentioned in more than one paper being grouped in the middle. Datasets are also color-coded to visually render five categories: harms, jobs, prescription, policy, and author name. The workflow for creating this network in Gephi is available at GitHub (https://github.com/cns/iu/agc2/jobs). The interactive visualization was created using JavaScript GEXF viewer package (Velt, 2019). The Gephi network was exported from Gephi into a gexf format (.gexf), a native XML format suitable for JavaScript (js) interactive visualization frameworks. Then, the gefx.js code was updated and uploaded to GitHub. The interactive network is available at https://cns/iu.github.io/agc2/jobs/all_opioid_datasets/main/index.html and it supports search, filter, and details on demand (Shneiderman, 1996), as illustrated in Figure 3.

The second network prototype represents a conceptual linkage, following similar steps to create nodes and edges using geolocation and standard crosswalks as linkages. For instance, the CDC Mortality and TEDS share the same linkage attribute "State." Thus, we can build their linkage from CDC Mortality with the ID "0" (source) to TEDs admission with the ID "1" (target) and vice versa since the network is undirected (see Tables 6, 7). The resulting network has 20 nodes of four categories and 146 edges of nine different types.

The second network is also color-coded to visually render four categories: harms, jobs, pharmaceuticals, and policy (see Figure 4). The workflow for creating this network in Gephi is available at GitHub (https://github.com/cns-iu/agc2-jobs). We followed the same steps as described earlier. The interactive network is available at https://cns-iu.github.io/agc2-jobs/20_datasets-main/index.html.

## 4 Discussion

The interactive visualization enables researchers to explore the relationships between data sources in the diverse context of public policy, economy, and treatment research related to the U.S. opioid crisis. In addition, it has the potential to help with data strategies and decision-making. For example, the user can search for a dataset AMA Physician by typing the name of the dataset in the search window or clicking directly on the node (see Figure 5). The information panel on the left displays the metadata information, showing that the dataset is private and related to the jobs category. It also provides the full dataset name, link, and the citation source. The researcher can also learn about other secondary datasets from the same job category color-coded in yellow, such as the NPPES dataset, which is publicly available. This information could help the researcher to make an informed decision on which dataset to use or to brainstorm new perspectives on opioid-related policy research.

Furthermore, the network representation demonstrates the potential for data discovery using metadata and data interlinkage skills as well as the discovery of the scholarly literature and case studies that use these datasets, which can assist researchers in identifying publicly available datasets, determining how these data can be combined in analysis, and surfacing relevant information about the provenance, availability, and definitions of data sources and variables. Figure 6 illustrates the discovery of the dataset ARCOS.

The node's dark blue color specifies that this dataset belongs to a pharmaceutical category. Clicking on the ARCOS node reveals links to the data dictionary and data source. In addition, the researcher is provided with a starter kit of the three most cited publications with this dataset. The size of the dataset is another useful attribute that allows the researcher to make decisions on storage space. Furthermore, this dataset could be linked at the state level with the CDC Mortality and TED Admissions data, using the drug name it can be connected with Open Payment and Medicare. This interactive network visualization of data relevant to public policy analysis may also be used as an instructional tool to help develop novel questions and assist in research. For example,

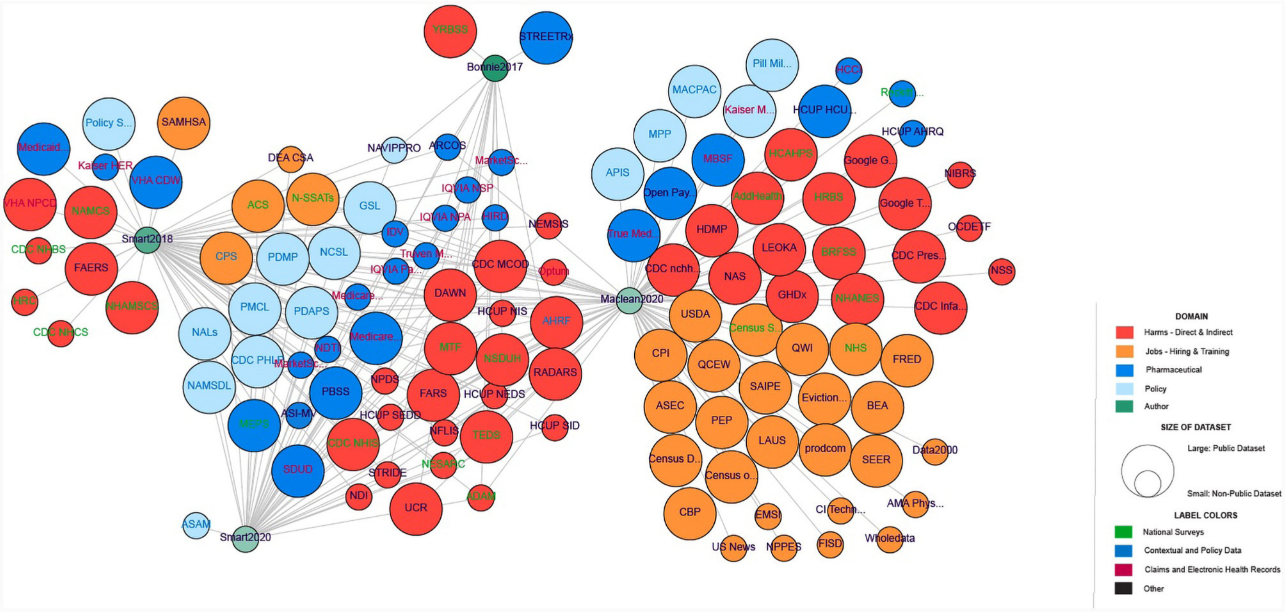
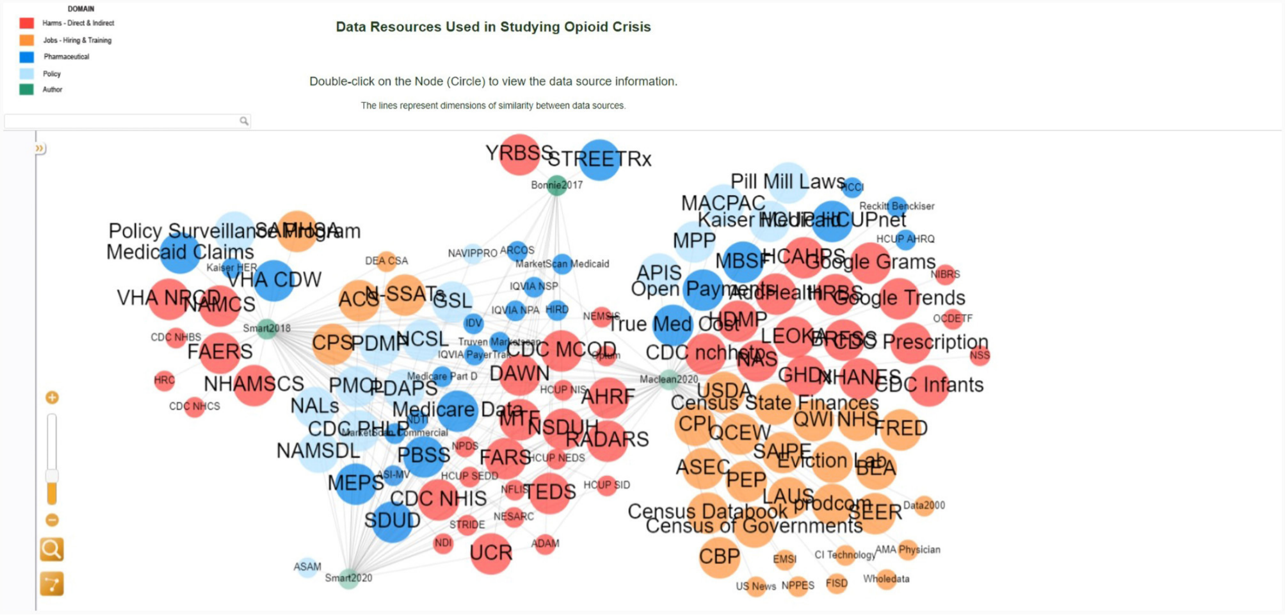


FIGURE 2
Network representation of the 121 datasets with policy data (in light blue), pharmaceutical data (dark blue), harms (red), jobs (hiring/training) data (orange), and authors (green). Transparency on the author nodes denotes the year of the publication (more transparency means a less recent year). Circle size corresponds to the availability (public or non-public) of the dataset. Label color denotes the type of dataset with national surveys (green), contextual and policy data (blue), claims and EHRs (maroon), and others (black).


Data Resources Used in Studying Opioid Crisis

Double-click on the Node (Circle) to view the data source information.

The lines represent dimensions of similarity between data sources.

FIGURE 3
Interactive network visualization with the legend in the top left explaining color and size coding; details on demand in the lower left; interactive network layout on right.

by discovering the state linkage between ARCOS, CDC Mortality, and TED Admissions, the new question could be what states have the highest mortality rate as well as the highest pharmaceutical sales.

## 5 Conclusion

In alignment with the key priority set forth by the Department of Health and Human Services (HHS) to address the opioid

TABLE 6 Nodelist, partial—only six of 16 attributes are shown.

| ID | Label | Category | Size | Start | End |
|---|---|---|---|---|---|
| 0 | CDC Opioid Mortality | Harms | S | 1999 | 2018 |
| 1 | Treatment Episode Dataset: Admissions | Harms | L | 1992 | 2017 |
| 2 | National Health Interview Survey | Harms | M | 1963 | 2019 |

TABLE 7 Edgelist.

| Source | Target | Network type | ID | Label | Weight | Relation type |
|---|---|---|---|---|---|---|
| 0 | 1 | Undirected | 0 | State | 1 | Geo unit |
| 0 | 3 | Undirected | 1 | County | 1 | Geo unit |
| 0 | 5 | Undirected | 2 | State | 1 | Geo unit |

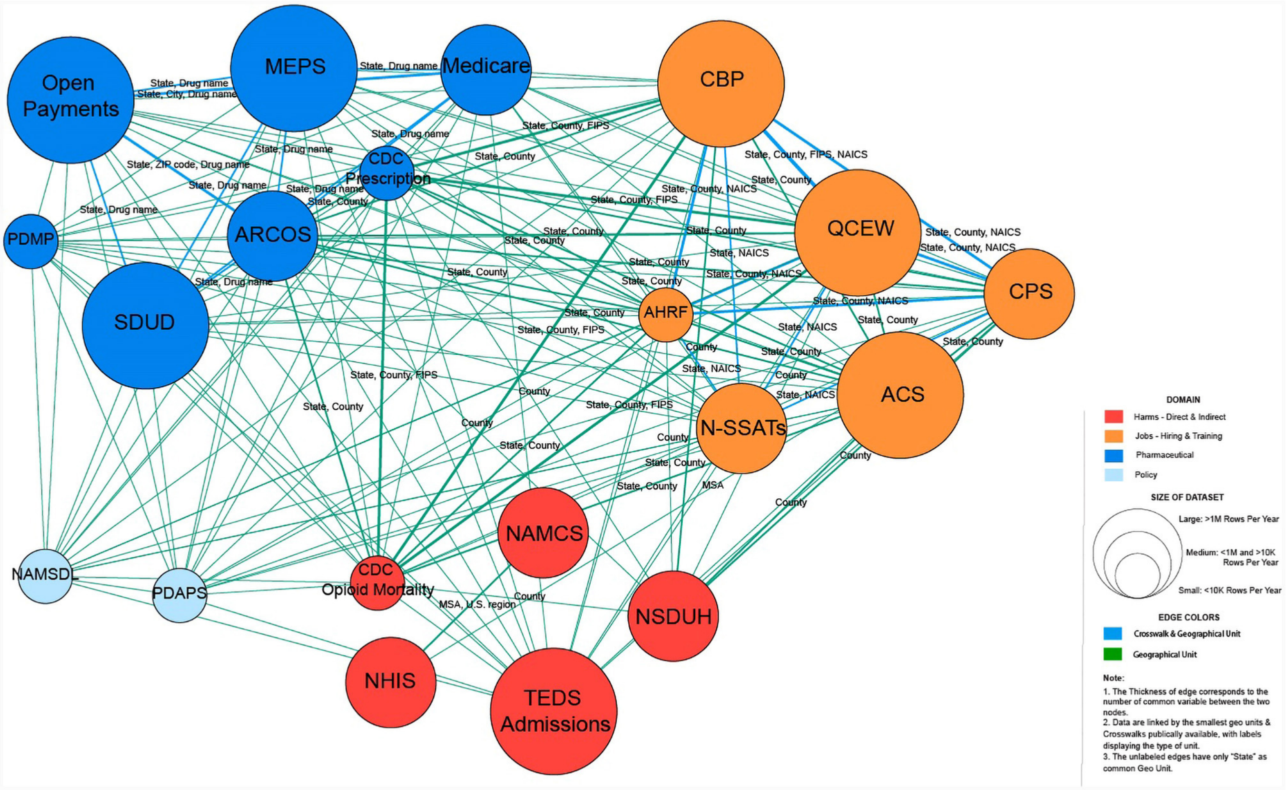
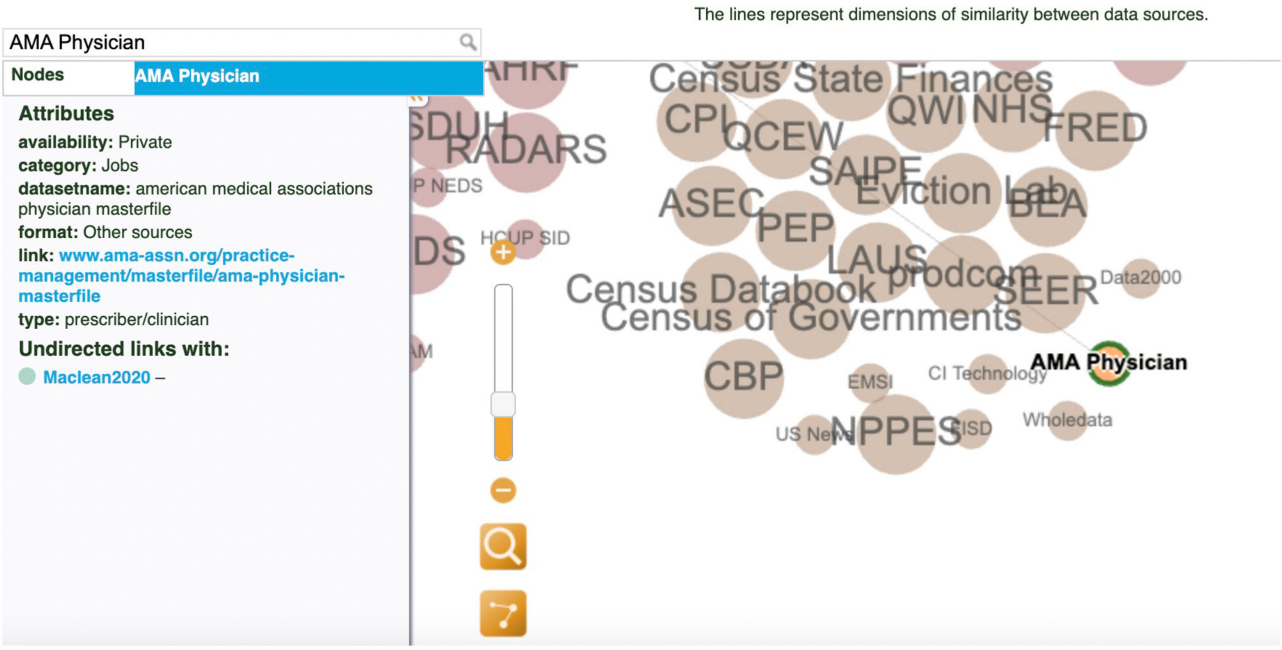

Open Payments
MEPS
Medicare
CBP
CDC Prescription
PDMP
ARCOS
QCEW
CPS
SDUD
AHRF
ACS
N-SSATs
NAMCS
NAMSDL
PDAPS
CDC Opioid Mortality
NSDUH
NHIS
TEDS Admissions
State, Drug name
State, City, Drug name
State, ZIP code, Drug name
State, County, FIPS
State, County, FIPS, NAICS
State, County, NAICS
State, County
State, NAICS
County
MSA
MSA, U.S. region
DOMAIN
Harms - Direct & Indirect
Jobs - Hiring & Training
Pharmaceutical
Policy
SIZE OF DATASET
Large: >1M Rows Per Year
Medium: <1M and >10K Rows Per Year
Small: <10K Rows Per Year
EDGE COLORS
Crosswalk & Geographical Unit
Geographical Unit
Note:
1. The Thickness of edge corresponds to the number of common variable between the two nodes.
2. Data are linked by the smallest geo units & Crosswalks publically available, with labels displaying the type of unit.
3. The unlabeled edges have only "State" as common Geo Unit.

FIGURE 4
Network representation of the subset datasets with policy data (in light blue), pharmaceutical data (dark blue), opioid data (red), and jobs (hiring/training) data (orange). Circle size corresponds to the size of the dataset. The edge color denotes the type of linkage.

crisis–enhancing information accessibility and promoting data-driven policy-making– our efforts have centered on metadata skills to discover and interconnect existing datasets. We developed a two-step process showing how to collect datasets using the scoping review method and transform data into network graphs. We curated 121 datasets, drawing from recent systematic reviews related to policy and opioid research. Furthermore, we have designed innovative visualization tool prototypes to assist researchers in data exploration. The interactive network visualization allows potential data users to navigate each dataset via data linkages, embedded data dictionaries, and recent publications using the selected dataset. A dataset can be identified as a complementary dataset to their current datasets to conduct relevant health services research studies or policy evaluations. Another data source can be used as an alternative dataset to validate their current data analysis. In addition, we provided the protocol for metadata collection and guidelines for network visualization and made it available for researchers to develop their dataset linkage networks.

This study has several limitations. First, we included datasets from four systematic review studies, potentially overlooking less common datasets. Second, we designed the linkage network between datasets only for 20 publicly available datasets, and



The lines represent dimensions of similarity between data sources.

AMA Physician

Nodes AMA Physician

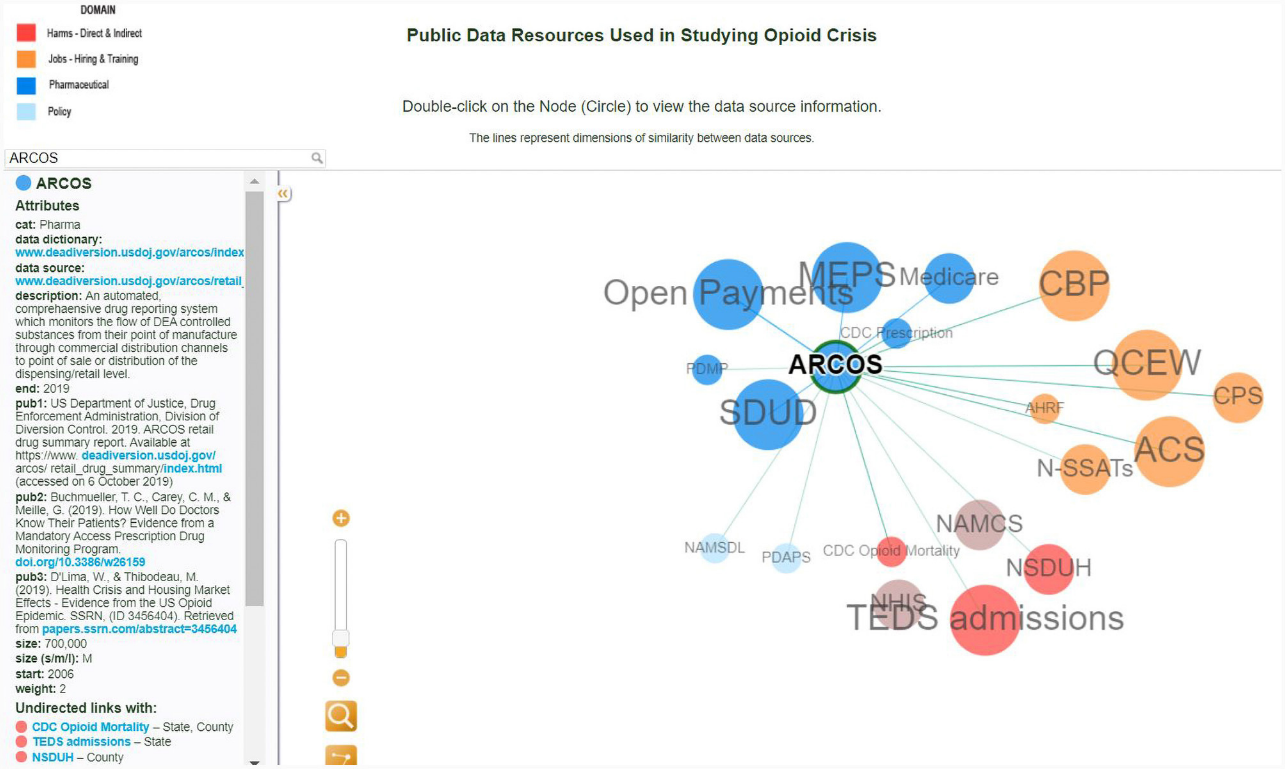
**Attributes**

**availability:** Private
**category:** Jobs
**datasetname:** american medical associations physician masterfile
**format:** Other sources
**link:** www.ama-assn.org/practice-management/masterfile/ama-physician-masterfile
**type:** prescriber/clinician

**Undirected links with:**

● Maclean2020 –

FIGURE 5
Search filter and information panel: The AMA Physician dataset.



DOMAIN
🟥 Harms - Direct & Indirect
🟧 Jobs - Hiring & Training
🟦 Pharmaceutical
▢ Policy

**Public Data Resources Used in Studying Opioid Crisis**

Double-click on the Node (Circle) to view the data source information.

The lines represent dimensions of similarity between data sources.

ARCOS

● **ARCOS**

**Attributes**

**cat:** Pharma
**data dictionary:** www.deadiversion.usdoj.gov/arcos/index
**data source:** www.deadiversion.usdoj.gov/arcos/retail_
**description:** An automated, comprehaensive drug reporting system which monitors the flow of DEA controlled substances from their point of manufacture through commercial distribution channels to point of sale or distribution of the dispensing/retail level.
**end:** 2019
**pub1:** US Department of Justice, Drug Enforcement Administration, Division of Diversion Control. 2019. ARCOS retail drug summary report. Available at https://www. deadiversion.usdoj.gov/ arcos/ retail_drug_summary/index.html (accessed on 6 October 2019)
**pub2:** Buchmueller, T. C., Carey, C. M., & Meille, G. (2019). How Well Do Doctors Know Their Patients? Evidence from a Mandatory Access Prescription Drug Monitoring Program. doi.org/10.3386/w26159
**pub3:** D'Lima, W., & Thibodeau, M. (2019). Health Crisis and Housing Market Effects - Evidence from the US Opioid Epidemic. SSRN, (ID 3456404). Retrieved from papers.ssrn.com/abstract=3456404
**size:** 700,000
**size (s/m/l):** M
**start:** 2006
**weight:** 2

**Undirected links with:**

● CDC Opioid Mortality – State, County
● TEDS admissions – State
● NSDUH – County

FIGURE 6
Interactive conceptual network visualization: the interlinkages for the ARCOS dataset.

we faced difficulties accessing metadata for non-publicly available datasets. This study also focuses on linkages at the concept level, rather than individual-level linking of data across datasets. This is an area of future development as it involves identifiable data and requires special data privacy considerations. Going forward, the same methodology can be applied to individual-level linked data and non-public resources. Another important area for future work is conducting user studies to identify how to best improve the visualization for different stakeholder groups and what additional datasets should be added. Finally, due to the limitation of dataset coverage, the current study assigned datasets only into four categories (policy, pharmaceutical, opioid, and jobs/training), excluding other important topics. Future research should use a more comprehensive view based on the opioid ecosystem approach (Stein et al., 2023).

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## Funding

The Indiana University Responding to the Addictions Crisis Grand Challenge Grant url—https://addictions.iu.edu/responding-to-crisis/grand-challenge.html.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Ahmad, F., Rossen, L., and Sutton, P. (2023). *Provisional Drug Overdose Death Counts. National Center for Health Statistics.*

Amer-Yahia, S., Koutrika, G., Bastian, F., Belmpas, T., Braschler, M., Brunner, U., et al. (2021). INODE: building an end-to-end data exploration system in practice [extended vision]. *arXiv*. [Preprint]. dpo: 10.48550/arXiv.2104.04194

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi: 10.1080/1364557032000119616

Bastian, M., Heymann, S., and Jacomy, M. (2009). "Gephi: an open source software for exploring and manipulating networks," in *Proceedings of the Third International Conference on Weblogs and Social Media*, eds E. Adar, M. Hurst, T. Finin, N. Glance, N. Nicolov, and B. Tseng (Menlo Park, CA: The AAAI Press), 361–362. doi: 10.1609/icwsm.v3i1.13937

Blanco, C., Wall, M. M., and Olfson, M. (2021). Data needs and models for the opioid epidemic. *Mol. Psychiatry* 27, 787–792. doi: 10.1038/s41380-021-01356-y

CDC (2019). *U.S. Opioid Dispensing Rate Maps*. Atlanta, GA: CDC.

Chou, R., Selph, S., Wagner, J., Ahmed, A. Y., Jungbauer, R., Mauer, K., et al. (2021). "Systematic Review on Opioid Treatments for Chronic Pain: Surveillance Report 1: Literature Update Period: August 2019 through September 2021," in *Systematic Review on Opioid Treatments for Chronic Pain: Surveillance Reports*. Rockville, MD.

Cox, S., Ahalt, S. C., Balhoff, J., Bizon, C., Fecho, K., Kebede, Y., et al. (2020). Visualization environment for federated knowledge graphs: development of an interactive biomedical query language and web application interface. *JMIR Med. Inform.* 8:e17964. doi: 10.2196/17964

Emmert-Streib, F., Tripathi, S., Yli-Harja, O., and Dehmer, M. (2018). Understanding the world economy in terms of networks: a survey of data-based network science approaches on economic networks. *Front. Appl. Math. Stat.* 4:37. doi: 10.3389/fams.2018.00037

Gamage, K. K., Chiu, K., Ryk, J., Grundy, Q., and Sud, A. (2023). How policy problems and solutions travel in the scientific literature: an international scientometric analysis of the french model of opioid use disorder care. *J. Eval. Clin. Pract.* 29, 576–590. doi: 10.1111/jep.13822

Hu, X., Gallagher, M., Loveday, W., Dev, A., and Connor, J. P. (2020). Network analysis and visualisation of opioid prescribing data. *IEEE J. Biomed. Health Inform.* 24, 1447–1455. doi: 10.1109/JBHI.2019.2939028

Leece, P., Khorasheh, T., Paul, N., Keller-Olaman, S., Massarella, S., Caldwell, J., et al. (2019). 'Communities are attempting to tackle the crisis': a scoping review on community plans to prevent and reduce opioid-related harms. *BMJ Open* 9:e028583. doi: 10.1136/bmjopen-2018-028583

Maclean, J. C., Mallatt, J., Ruhm, C. J., and Simon, K. (2020). *Economic Studies on the Opioid Crisis: A Review. NBER Working Papers*. Cambridge, MA: National Bureau of Economic Research. doi: 10.3386/w28067

McNeely, J., Schatz, D., Olfson, M., Appleton, N., and Williams, A. R. (2021). How physician workforce shortages are hampering the response to the opioid crisis. *Psychiatr. Serv.* 73, 547–554. doi: 10.1176/appi.ps.202000565

Morris, J. H., Soman, K., Akbas, R. E., Zhou, X., Smith, B., Meng, E. C., et al. (2023). The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics* 39:btad080.

Niyirora, J., and Aragones, O. (2020). Network analysis of medical care services. *Health Inform. J.* 26, 1631–1658. doi: 10.1177/1460458219887047

Payán, D. D. (2021). Cultivating health policy analysis and communication skills in undergraduate public health education: an active learning approach. *Pedagogy Health Promot.* 7, 235–241. doi: 10.1177/23733799211003248

Phillips, J. K., Ford, M. A., and Bonnie, R. J. (Eds.). (2017). "Appendix C: existing data sources on opioid use, misuse, overdose, and other harms," in *Pain Management and the Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription Opioid Use* (Washington, DC: The National Academies Press), 447–455.

Poitras, G. (2018). The prescription opioid epidemic: an update. *Medicoleg. Bioeth.* 8, 21–32. doi: 10.2147/MB.S170220

Price, T. E. (2017). *Secretary Price Announces HHS Strategy for fighting Opioid Crisis.*

Scrivner, O., Nguyen, T., Simon, K., Middaugh, E., and Taska, B. Börner, K. (2020). Job postings in the substance use disorder treatment related sector during the first five years of Medicaid expansion. *PLoS ONE* 15:e0228394. doi: 10.1371/journal.pone.0228394

Shlomo, N. (2018). "Overview of data linkage methods for policy design and evaluation," in *Data-Driven Policy Impact Evaluation*, eds N. Crato, and P. Paruolo (Berlin: Springer International Publishing), 47–65. doi: 10.1007/978-3-319-78461-8_4

Shneiderman, B. (1996). "The eyes have it: a task by data type taxonomy for information visualizations," in *Proceedings 1996 IEEE Symposium on Visual Languages* (Boulder, CO: IEEE Comput. Soc. Press), 336–343. doi: 10.1109/VL.1996.545307

Sima, A. C., Mendes de Farias, T., Zbinden, E., Anisimova, M., Gil, M., Stockinger, H., et al. (2019). Enabling semantic queries across federated bioinformatics databases. *Database* 2019:1–11. doi: 10.1093/database/baz106

Skillman, S. M., Snyder, C. R., Frogner, B. K., and Patterson, D. G. (2016). *The Behavioral Health Workforce Needed for Integration with Primary Care: Information for Health Workforce Planning*. Technical report. Washington, DC: Center for Health Workforce Studies, University of Washington.

Smart, R., Kase, C. A., Meyer, A., and Stein, B. (2018). *Data Sources and Data-Linking Strategies to Support Research to Address the Opioid Crisis*. Technical Report September. Washington, DC: U.S. Department of Health and Human Services.

Smart, R., Kase, C. A., Taylor, E. A., Lumsden, S., Smith, S. R., Stein, B. D., et al. (2020). Strengths and weaknesses of existing data sources to support research to address the opioids crisis. *Prev. Med. Rep*. 17:101015. doi: 10.1016/j.pmedr.2019.101015

SN SciGraph. (2019). *SN SciGraph Latest Release: Patents, Clinical Trials and Many New Features*. Springer Nature.

Spencer, M. R., Minio, A. M., and Warner, M. (2022). *Drug Overdose Deaths in the United States, 2001–2021. NCHS Data Brief*. Atlanta, GA: CDC. doi: 10.15620/cdc:122556

Stein, B. D., Kilmer, B., Taylor, J., Vaiana, M. E., Barnes-Proby, D., Caulkins, J. P., et al. (2023). America's opioid ecosystem: how leveraging system interactions can help curb addiction, overdose, and other harms. *Rand Health Q*. 10.

Stephens, O. (2018). *OpenRefine User Manual: Clustering in Depth. Methods and Theory Behind the Clustering Functionality in OpenRefine. [Version 3.7.9]*.

Substance Abuse and Mental Health Services Administration (2021). *National Survey of Substance Abuse Treatment Services (N-SSATS): 2020. Data on Substance Abuse Treatment Facilities*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2022). *Key substance use and mental health indicators in the United States: Results from the 2021 National Survey on Drug Use and Health (HHS Publication No. PEP22-07-01-005, NSDUH Series H-57)*. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Available online at: https://www.samhsa.gov/data/report/2021-nsduh-annual-national-report

Velt, R. (2019). *A JavaScript GEXF Viewer [Application]*.

Weber, G. M., Mandl, K. D., and Kohane, I. S. (2014). Finding the missing link for big biomedical data. *JAMA* 311, 2479–2480. doi: 10.1001/jama.2014.4228

Wu, M. F., Richard, S. M., and Verhey, C. (2023). An analysis of crosswalks from research data schemas to schema.orh. *Data Intell.* 5, 100–121. doi: 10.1162/dint_a_00186