JAMIA Open, 2023, **6(2)**, ooad036
https://doi.org/10.1093/jamiaopen/ooad036
Brief Communications

Brief Communications

Common data model for sickle cell disease surveillance: considerations and implications

Matthew P. Smeltzer [1], Sarah L. Reeves [2], William O. Cooper [3,4], Brandon K. Attell [5], John J. Strouse [6], Clifford M. Takemoto [7], Julie Kanter [8], Krista Latta [2], Allison P. Plaxco [1], Robert L. Davis [9], Daniel Hatch [10], Camila Reyes [11], Kevin Dombkowski [2], Angela Snyder [5], Susan Paulukonis [12], Ashima Singh [13], and Mariam Kayle [10]

[1]Division of Epidemiology, Biostatistics, and Environmental Health School of Public Health, University of Memphis, Memphis, Tennessee, USA
[2]Department of Pediatrics, Susan B Meister Child Health Evaluation and Research (CHEAR) Center, University of Michigan, Ann Arbor, Michigan, USA
[3]Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[4]Department of Health Policy, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[5]Georgia Health Policy Center, Georgia State University, Atlanta, Georgia, USA
[6]Department of Hematology, Duke University, Durham, North Carolina, USA
[7]Department of Hematology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA
[8]Division of Hematology-Oncology, University of Alabama Birmingham, Birmingham, Alabama, USA
[9]Department of Bioinformatics, University of Tennessee Health Science Center, Memphis, Tennessee, USA
[10]Duke University School of Nursing, Durham, North Carolina, USA
[11]Duke Office of Clinical Research, Duke University School of Medicine, Durham, North Carolina, USA
[12]Tracking California, Public Health Institute, Oakland, California, USA
[13]Department of Pediatrics, Medical College of Wisconsin, Milwaukee, Wisconsin, USA

Corresponding Author: Matthew P. Smeltzer, PhD, MStat, Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, 222 Robison Hall, Memphis, TN 38152, USA; msmltzer@memphis.edu

ABSTRACT

Objective: Population-level data on sickle cell disease (SCD) are sparse in the United States. The Centers for Disease Control and Prevention (CDC) is addressing the need for SCD surveillance through state-level Sickle Cell Data Collection Programs (SCDC). The SCDC developed a pilot common informatics infrastructure to standardize processes across states.

Materials and Methods: We describe the process for establishing and maintaining the proposed common informatics infrastructure for a rare disease, starting with a common data model and identify key data elements for public health SCD reporting.

Results: The proposed model is constructed to allow pooling of table shells across states for comparison. Core Surveillance Data reports are compiled based on aggregate data provided by states to CDC annually.

Discussion and Conclusion: We successfully implemented a pilot SCDC common informatics infrastructure to strengthen our distributed data network and provide a blueprint for similar initiatives in other rare diseases.

LAY SUMMARY

Sickle cell disease (SCD) is a complex health disorder that requires frequent scheduled and unscheduled healthcare visits. In the United States, we do not have comprehensive data on the numbers of persons with SCD and where they receive healthcare. In order to understand this better, The Centers for Disease Control and Prevention (CDC) is supporting SCD surveillance through state-level Sickle Cell Data Collection Programs. Because the data are collected within each state, they are not standardized across states. In order to standardize the datasets across states, we have constructed a common data model. This will allow participating states to standardize their data so that it can be presented together for multi-state projects. In our model, each state keeps its own data, but can work with other states to combine information. We hope this common data model will be useful to other groups working with SCD data or data on other rare diseases.

Key words: sickle cell, data model, surveillance

BACKGROUND AND SIGNIFICANCE

Sickle cell disease (SCD) is a complex, chronic health disorder affecting approximately 100 000 persons in the United States,[1] although the exact number is unknown.[2] The burden of SCD lies in severe complications, associated comorbidities, the need to access specialized and coordinated care, and high reliance on acute healthcare services. Affecting marginalized populations in the United States, health inequities and social determinants further contribute to negative health outcomes. With no population-based national registry, the availability of data on the

Received: 31 August 2022. **Revised:** 10 February 2023. **Editorial Decision:** 4 May 2023. **Accepted:** 23 May 2023

epidemiology of SCD and the healthcare needs of the population to inform clinical practice and health policy are sparse.

The Sickle Cell Data Collection Program (SCDC) funded by the Centers for Disease Control and Prevention (CDC) is addressing this need by developing state-level, multi-source surveillance programs. Data required for this effort are maintained in many state-specific systems and formats including, state newborn screening, state Medicaid databases, state all-payer claims, hospital discharge data, death records, and hospital-based electronic medical records (EMR). Given the variability in data sources and structures, a standardized process across SCDC states is essential to ensure reproducible and valid results for national reporting.

Common informatics infrastructures such as common data models allow distributed data networks to standardize, integrate, and analyze data across multiple sources.[3,4] Using the same data definitions, multi-site analyses can be conducted efficiently and with higher quality. Common informatics infrastructures are particularly useful for evaluating rare diseases, outcomes, or therapies as aggregating data across multiple states can provide a more generalizable representation of the population and improve statistical power for analyses.[5] This is especially important for SCD, given that much of the existing health services research for SCD relies on administrative case definitions using ICD codes that could underestimate the SCD population.[6]

Common informatics infrastructures are currently utilized by numerous shared health data networks that have applied common data models for a variety of purposes, including cancer care, vaccine safety, drug safety, and healthcare delivery.[3] The numerous successful applications demonstrate the feasibility and value of common data models and shared data networks.

OBJECTIVE

Most common informatics infrastructures are based on data obtained from clinical trials or directly from EMR. To our knowledge, no common informatics infrastructure has been developed to integrate and harmonize source data from medical claims and newborn genetic screening for public health surveillance. Furthermore, there are no common data models specific to SCD or many other rare conditions. We describe the process for establishing and maintaining a pilot common data model in the SCDC surveillance setting and identify key data elements for public health reporting on SCD. The proposed model may be of value to facilitate the expansion of public health surveillance for SCD and other rare diseases.

MATERIALS AND METHODS

The SCDC was established by the CDC in 2015 and currently includes 11 states. Core data sources are state-level newborn screening and state-level Medicaid claims data. Most programs have additional data sources including EMR, state all-payer datasets, hospital and emergency department discharge data, and clinical cohorts.

Each state uses a state-specific study protocol and creates a comprehensive list (index file) of an SCD case. We developed consensus around key data elements included in the index file for each data instrument, variable definitions with coding examples, and considerations for prioritization among multiple data sources. By design, states maintain individual-level data and combine aggregate data across states for national reporting, facilitating both single-state and multi-state projects. Figure 1 outlines the process of data compilation, construction of analytic datasets, analysis, and reporting. Multi-state projects have a lead group that coordinates data analyses and standardizes specific project methodology. Methods for handling outliers and missing data are determined on a project-by-project basis. States conduct analyses in parallel and return aggregate results to the lead group (Figure 1). Aggregate results yielding small cell counts (less than 5, 10, or

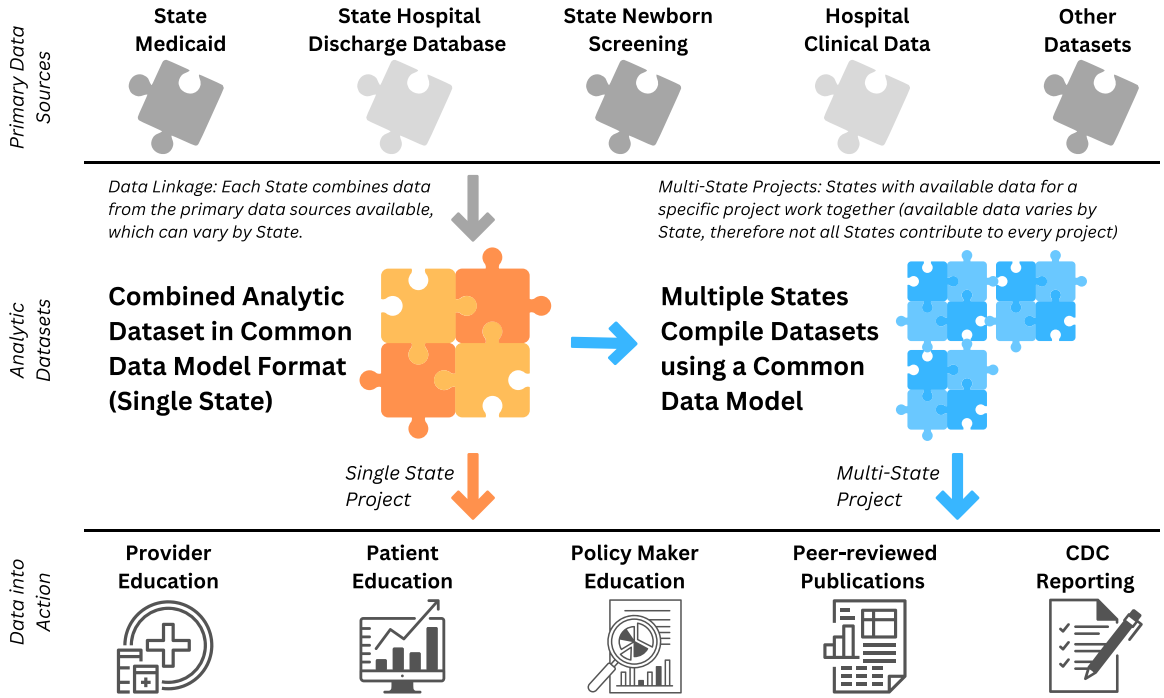

Figure 1. Process of data compilation, construction of analytic datasets, analysis, and reporting (single- and multi-state projects).

25) are suppressed as a part of our privacy-preserving data plan.[7] Internal and external investigators may propose and lead projects with approval of the SCDC Data Use Committee.

The pilot SCDC Common Data Model was developed to standardize key data elements including identification of data instruments, specific variables, and variable formatting. The model allows pooling of table shells across states for comparison, simplifying reporting and analyses. Because many key variables are available from multiple data sources for the same individuals, we developed a semiautomated data hierarchy based on the most valid data source for that variable. For example, birth year is obtained first from newborn screening records or alternatively birth certificates (Supplementary Table S1). This establishes standardized methodology for determining which sources should be prioritized for specific variables, using both literature and expert consensus. Because not every state data will have the same level of detail, pragmatic choices are sometimes necessary, which allow each state flexibility to individualize the hierarchy.

SCDC process for common data model development

To develop the pilot SCDC Common Data Model, we built consensus around key data element groups in categories of variables, which we term "Common Data Model Instruments". The format allows development of new instruments as needed (Figure 2). Three necessary instruments have been identified thus far: the Core Surveillance Data Instrument, the Health Outcomes Instrument, and the Pharmacy Instrument. We formed a SCDC Common Data Model committee with expertise in sickle cell clinical care, administrative healthcare data, epidemiology, data base construction, and data analysis, including SCD champions. In the future, all documentation for the proposed SCDC Common Data Model will be housed in an accessible online public repository.

RESULTS

We implemented the Core Surveillance Data Instrument of the Common Data Model in Tennessee, North Carolina, and Michigan. The SCDC has identified core results that are feasible for most states to attain, typically measured with reasonable accuracy, and have high public health and policy relevance. These Core Surveillance Data are reported annually by states and compiled by CDC. We implemented the approach outlined above to develop the Core Surveillance Data Instrument, including all variables necessary to create four sets of tables: births, case number estimates, deaths, and healthcare utilization. Linkage variables, though not reported publicly, are also important to implement the preprocessing of data, linkage, and privacy preserving encryption. The data dictionary for the Core Surveillance Data Instrument (Table 1) was developed based on expert experience with SCD data, a survey of the data structures states currently utilize, and a review of other common data models.

Births

We use newborn screening data to report numbers of SCD births (1-year and 5-year increments) by sex, race, county, ethnicity, and SCD type, following the data dictionary with the suggested hierarchy (Supplementary Table S1). Newborn screening records are the primary source, with state laboratory confirmatory testing used to identify SCD type. Each site

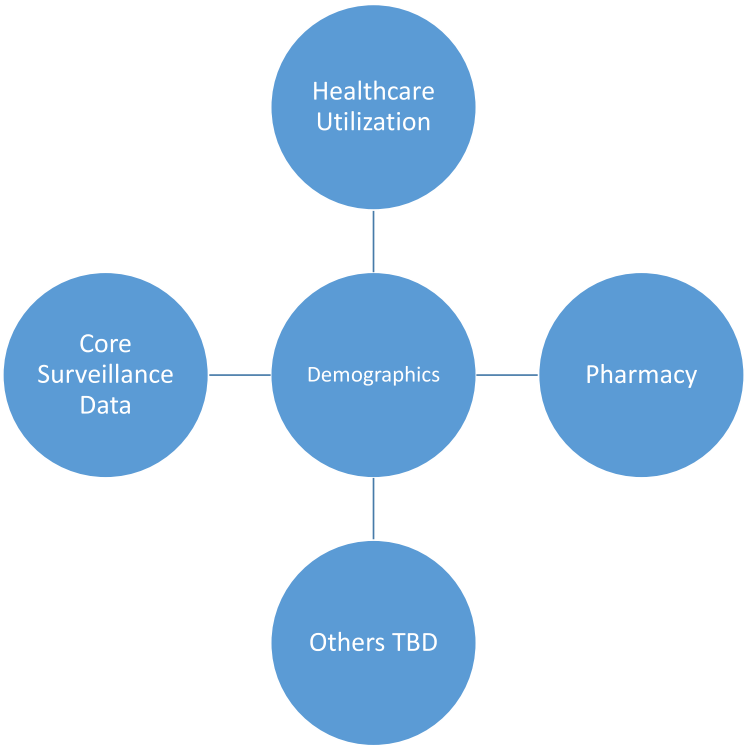

Figure 2. Variable groupings (instruments) for SCDC common data model.

Table 1. Dictionary for core surveillance data instrument

Data variable description	Variable name	SAS data type	Annual report value set
Data elements from birth certificates and newborn screening			
Year of Birth	birth_year	Numeric	Integer
Sex on birth certificate	sex_nbs	Numeric	0 = Male 1 = Female 2 = Ambiguous 555 = Other 999 = Unknown
Race on birth certificate	race_nbs	Numeric	1='American Indian' 2='Asian' 3='Black or African American' 4='Native Hawaiian or Pacific Islander' 5='Caucasian' 6='Other Race' 7='Unknown' 8='more than one race'
Ethnicity on birth certificate	ethnicity_nbs	Numeric	0='Non-Hispanic' 1='Hispanic' 999='Unknown'
Newborn screening diagnosis	dx_nbs	Character	SS= 'HbSS'[a] S0= 'HbSBeta0'[a] SC='HbSC' SP='HbSBetaPlus' OT= 'Other'
Mother's county of residence at birth	mothercounty_nbs	Numeric	five digit state and county FIPS code
Data elements from medical claims or records (encounter level)			
Demographics			
Age[b]	ageYY	Numeric	Integer
Age category	age_cat	Numeric	1≤10y, 2 = 10–19y, 3 = 20–29y, 4 = 30–39y, 5 = 40–49y, 6 = 50–59y, 7 = 60+y
Sex	sex	Numeric	0 = Male 1 = Female 2 = Ambiguous 555 = Other 999 = Unknown
Geography			
Name of county	county	Numeric	five digit state and county FIPS code
Metro/nonmetro indicator by county (RUCC)	RUCC_2013	Character	ME= Metro (codes 1–3) NM= Nonmetro (codes 4–9)
Utilization			
Year	Year_utilized	Numeric	Integer
Hospital admission	Type_Hosp	Binary	0=No 1=Yes
Hospital Length of Stay (days)	Hosp_Days	Numeric	Integer
Emergency Department (ED) Treat and Release Visit	Type_ED	Binary	0=No 1=Yes
Primary payer	payer	Numeric	0="Self-pay" 1="Private insurance" 2="Medicare" 3="Medicaid" 4="Other"
30-day readmission	readmit_30	Binary	0=No 1=Yes
ED visit precedes hospital record	Ed_part_Hosp	Binary	0=No 1=Yes
Data elements from vital records			
Death			
Age at death	AgeDeath	Numeric	Integer
Age group at death	AgeDeathgrp	Numeric	0=<20 1 = 20–49 2=≥50 years

[a] SBO/SS combined for the data report.
[b] Age is calculated from date of birth, as age on December 31 of the report year (eg, age on December 31, 2017 for the 2017 report) and coded as ageYY (eg, age17 for 2017 report).

Table 2. Applying the common data model in three states

	Tennessee *N* (%)	Michigan *N* (%)	North Carolina *N* (%)
Year	2016	2017	2016
No. of participants	1999	3331	5118
Sex			
Male	864 (43.22)	1382 (41.5)	2343 (45.8)
Female	1135 (56.78)	1949 (58.5)	2775 (54.2)
Unknown	0 (0)	0 (0)	0 (0)
Race			
Black	1907 (95.4)	2710 (81.3)	4890 (95.6)
Other	61 (3.05)	239 (7.2)	185 (3.6)
Unknown	31 (1.55)	382 (11.5)	43 (0.8)
Ethnicity			
Hispanic	1–25	32 (0.9)	99 (1.9)
Non-Hispanic	451 (22.56)	2281 (68.5)	4908 (95.9)
Unknown	1531 (76.59)	1018 (30.6)	111 (2.2)
Confirmed SCD type	NA[a]		NA[a]
S/S or S/B0 thal		971 (29.1)	
S/C		479 (14.4)	
S/B+ thal		0 (0)	
Other		1–25	
Unknown		1874 (56.3)	
Age group (in years)			
<10	463 (23.16)	711 (21.3)	947 (18.5)
10–19	398 (19.91)	568 (17.1)	882 (17.2)
20–29	424 (21.21)	825 (24.8)	1240 (24.2)
30–39	311 (15.56)	464 (13.9)	895 (17.5)
40–49	204 (10.21)	357 (10.7)	584 (11.4)
50–59	136 (6.8)	230 (6.9)	352 (6.9)
60+	63 (3.15)	176 (5.3)	218 (4.3)
Births with SCD			
Year	2016	2017	2016
No. of births with SCD	48	64	85
Sex			
Male	27 (56.25)	30 (46.9)	50 (58.8)
Female	1–25	34 (53.1)	35 (41.2)
Unknown	NR	0 (0)	0 (0)
Race			
Black	Censored	60 (93.8)	82 (96.5)
Other	1–25	1–25	1–25
Unknown	0 (0)	NR	NR
Ethnicity			
Hispanic	Censored	51 (79.7)	1–25
Non-Hispanic	1–25	1–25	78 (91.8)
Unknown	0 (0)	1–25	1–25
SCD type based on NBS			
S/S or S/B0 thal	30 (62.5)	42 (65.6)	52 (61.2)
S/C	1–25	1–25	27 (31.8)
S/B+ thal	1–25	1–25	1–25
Other	0 (0)	0 (0)	1–25
Unknown	0 (0)	0 (0)	0 (0)

[a] Currently, confirmed SCD type only available for individuals in newborn screening data.

uses semiautomated approaches to creating the linkage files and standardized approaches to encryption of data.

Case number estimates

We estimated the prevalent cases across each state by county, sex, and age group. Primary sources for these tables are linked healthcare claims datasets, newborn screening, and clinical datasets. Sources of claims include state Medicaid data, state all-payer data, and EMR or hospital discharge data (recommended hierarchy in Supplementary Table S1). A multi-tier case definition identifies individuals as possible, probable, or confirmed based on a standardized, validated case definition.[8] Reports are structured to be concordant with standard epidemiologic reporting for population-level characteristics in the United States.

Deaths

Death information is reported by age at death, stratified by sex (Table 1). Total numbers are reported across a range of dates, based on the available data. State vital records derived from death certificates are the source of death information.

Healthcare utilization

Core Surveillance Data for healthcare utilization report acute care utilization including number of hospitalizations, hospital length-of-stay, and number of emergency department visits (without admission). Results are reported by age group and payer type (Table 1). Availability and hierarchy of these results depend on the data sources each state is able to obtain, link, and deduplicate (Supplementary Table S1).

Data compilation

Core Surveillance Data reports are compiled for all states, based on aggregate data provided by states to CDC annually. CDC uses aggregate data for public reporting and informing public health initiatives. Each year, as data are refreshed, the possibility of prior year estimates being updated is possible, when data quality improves or additional data sources become available. Changes are therefore expected and will be noted and tracked. Large changes will be flagged for audit and reasons for differences noted.

Implementation

Tennessee, Michigan, and North Carolina have implemented the proposed SCDC Common Data Model for core surveillance data elements. Supplementary Table S1 shows the data dictionary adopted by each state's program before implementation of the proposed SCDC Common Data Model and the suggested hierarchy. Although the primary information is available in each state, the structure of the data varied substantially before implementation.

We used the Common Data Model to report select elements of the Core Surveillance Data Instrument as a proof of principle (Table 2). Tennessee reported a total of 1999 individuals and 48 births with SCD in 2016. Michigan had a total of 3331 individuals and 64 births with SCD in 2017. North Carolina reported 5118 total individuals with SCD and 85 births in 2016. These data were compiled individually by the state SCDC teams.

DISSCUSSION

To standardize SCD surveillance, we implemented a pilot common data model that will allow SCDC to strengthen their distributed data network. We describe the process of developing and maintaining the proposed common data model in this unique setting and identify key data elements for public health reporting on SCD. Our approach builds one data instrument at a time in a model that is expandable and modifiable. These results will inform the design of the SCDC program in the future, and we hope it will be adopted by data holders beyond SCDC.

The benefits of the common data model framework are numerous.[9–12] With standardization, we can leverage cross-state knowledge and experience to strengthen the group. Adoption of the pilot SCDC Common Data Model improves data query speed, potentially leading to more research and faster dissemination of findings from SCDC studies. By

formalizing processes, we also aim to reduce the learning curve as new states join SCDC.

This approach is similar to PCORNET in several key ways. Both use a distributed data model to streamline results reporting.[13] In both models individual data are not compiled centrally and do not leave the states (SCDC) or institutions (PCORNET). We have multiple data instruments like PCORNET, although the SCDC scale is currently much smaller. Additionally, our goals are similar to the CDC Clinical and Community Data Initiative (CODI).[14] However, CODI utilizes a central Data Coordinating Center that links datasets and joins data from different sites in a Distributed Health Data Network while SCDC houses individual level data within each state. The CODI model places less burden on data providers, but requires a funded and maintained Data Coordinating Center to perform these activities.[14] We have thus far focused on administrative and newborn screening data sources, but other common data models have sought to standardize concepts and vocabulary of EMR across disparate data sources, such as Observational Medical Outcomes Partnership (OMOP), i2b2, and PCORNET.

Through our unique approach to disease surveillance, SCDC strives to know the uses and users. We employ an inclusive, multi-partner governance framework. To this end, we aim to start simply and iterate towards complexity. Understanding data limitations and maintaining a high level of data quality are key tenants.

Some challenges arise with implementing a common data model across multiple data sources. During variable definition, unanticipated discrepancies may be identified. Finding the right balance of standardization and flexibility is important, especially due to uniqueness of some very valuable datasets.

Limitations notwithstanding, active surveillance programs using administrative data are of great value in understanding rare diseases. It is paramount to remember the purpose of surveillance, to answer questions about location of individuals with disease, access to healthcare, necessary resource allocation, and policy-impact questions; not to dictate individual care. SCDC will continue to expand the SCDC Common Data Model and disseminate the structure as we move from simplicity to complexity.

FUNDING

This work was supported by the Centers of Disease Control and Prevention, Sickle Cell Data Collection Program, grant number NU58DD000019.

AUTHOR CONTRIBUTIONS

MPS, MK, SR, and WOC contributed writing—original draft. AP, KL, and CR contributed formal analysis. All authors contributed writing—review and editing.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

CONFLICT OF INTEREST STATEMENT

Dr Matthew Smeltzer has worked as a paid research consultant for the Association of Community Cancer Centers. The findings and conclusions in this publication are those of the authors and do not necessarily represent the views of the North Carolina Department of Health and Human Services, Division of Public Health. No other authors have competing interests to declare.

DATA AVAILABILITY

Data used in this paper are housed within individual state SCDC programs under state specific Data Use Agreements. Data are not publically available.

REFERENCES

1. Hassell KL. Population estimates of sickle cell disease in the U.S. *Am J Prev Med* 2010; 38 (4 Suppl): S512–21.
2. Data & Statistics on Sickle Cell Disease | CDC. https://www.cdc.gov/ncbddd/sicklecell/data.html. Accessed May 20, 2022.
3. Brown JS, Mendelsohn AB, Nam YH, *et al.* The US Food and Drug Administration Sentinel System: a national resource for a learning health system. *J Am Med Inform Assoc* 2022; 29 (12): 2191–200.
4. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care* 2010; 48 (6 Suppl): S45–51.
5. Kent S, Dawoud D, Jonsson P, *et al.* Common problems, common data model solutions: evidence generation for health technology assessment. *Pharmacoeconomics* 2021; 39 (3): 275–85.
6. Snyder AB, Lane PA, Zhou M, Paulukonis ST, Hulihan MM. The accuracy of hospital ICD-9-CM codes for determining sickle cell disease genotype. *J Rare Dis Res Treat* 2017; 2 (4): 39–45.
7. Waitman LR, Song X, Walpitage DL, *et al.* Enhancing PCORnet Clinical Research Network data completeness by integrating multistate insurance claims with electronic health records in a cloud environment aligned with CMS security and privacy requirements. *J Am Med Inform Assoc* 2022; 29 (4): 660–70.
8. Snyder AB, Zhou M, Theodore R, Quarmyne MO, Eckman J, Lane PA. Improving an administrative case definition for longitudinal surveillance of sickle cell disease. *Public Health Rep* 2019; 134 (3): 274–81.
9. Tabano DC, Cole E, Holve E, Davidson AJ. Distributed data networks that support public health information needs. *J Public Health Manag Pract* 2017; 23 (6): 674–83.
10. Weeks J, Pardee R. Learning to share health care data: a brief timeline of influential common data models and distributed health data networks in U.S. health care research. *eGEMs* 2019; 7 (1): 4.
11. DeStefano F; Vaccine Safety Datalink Research Group. The Vaccine Safety Datalink project. *Pharmacoepidemiol Drug Saf* 2001; 10 (5): 403–6.
12. Chen RT, Glasser JW, Rhodes PH, *et al.* Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. *Pediatrics* 1997; 99 (6): 765–73.
13. PCORNet System. PCORnet Common Data Model. https://pcornet.org/pcornet-common-data-model/. Accessed June 26, 2018.
14. Clinical & Community Data Initiative. Brochure. CDC. https://www.cdc.gov/obesity/initiatives/codi/CODI-overview-fact-sheet-2021-508.pdf. Accessed May 10, 2022.