PLOS ONE

RESEARCH ARTICLE

CureSCi Metadata Catalog–Making sickle cell studies findable

Huaqin Pan[1]*, Cataia Ives[1], Meisha Mandal[1], Ying Qin[1], Tabitha Hendershot[1], Jen Popovic[1], Donald Brambilla[1], Jeran Stratford[1], Marsha Treadwell[2], Xin Wu[1], Barbara Kroner[1]

1 RTI International, Research Triangle Park, Durham, NC, United States of America, **2** Children's Hospital & Research Center Oakland, Oakland, CA, United States of America

* hpan@rti.org

Citation: Pan H, Ives C, Mandal M, Qin Y, Hendershot T, Popovic J, et al. (2022) CureSCi Metadata Catalog–Making sickle cell studies findable. PLoS ONE 17(12): e0256248. https://doi.org/10.1371/journal.pone.0256248

Editor: Alfredo Vellido, Universitat Politecnica de Catalunya, SPAIN

Received: July 30, 2021

Accepted: October 27, 2022

Published: December 12, 2022

Data Availability Statement: There is no data set involved in the CureSCi MDC features described in this paper. The tool uses metadata to support users finding name of studies and to point users to the location of the study datasets. All of the metadata documentation that was used in the curation of the studies has been uploaded to the CureSCi MDC website (https://curesicklecell.rti.org/) and can be accessed via the documentation tab on the "Study Detail" page for each of the included studies.

Abstract

Objectives

To adopt the FAIR principles (Findable, Accessible, Interoperable, Reusable) to enhance data sharing, the Cure Sickle Cell Initiative (CureSCi) MetaData Catalog (MDC) was developed to make Sickle Cell Disease (SCD) study datasets more Findable by curating study metadata and making them available through an open-access web portal.

Methods

Study metadata, including study protocol, data collection forms, and data dictionaries, describe information about study patient-level data. We curated key metadata of 16 SCD studies in a three-tiered conceptual framework of category, subcategory, and data element using ontologies and controlled vocabularies to organize the study variables. We developed the CureSCi MDC by indexing study metadata to enable effective browse and search capabilities at three levels: study, Patient-Reported Outcome (PRO) Measures, and data element levels.

Results

The CureSCi MDC offers several browse and search tools to discover studies by study level, PRO Measures, and data elements. The "Browse Studies," "Browse Studies by PRO Measures," and "Browse Studies by Data Elements" tools allow users to identify studies through pre-defined conceptual categories. "Search by Keyword" and "Search Data Element by Concept Category" can be used separately or in combination to provide more granularity to refine the search results. This resource helps investigators find information about specific data elements across studies using public browsing/search tools, before going through data request procedures to access controlled datasets. The MDC makes SCD studies more Findable through browsing/searching study information, PRO Measures, and data elements, aiding in the reuse of existing SCD data.

Funding: This project was supported by the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health (OT3 HL147798, https://www.nhlbi.nih.gov/) to B. Kroner. The funder had no role in the conceptualization, analysis, interpretation, or decision to publish this manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Sickle Cell Disease (SCD) is a rare hereditary blood disorder affecting an estimated 100,000 individuals in the United States [1]. The National Heart, Lung, and Blood Institute (NHLBI)-funded Cure Sickle Cell Initiative (CureSCi) is a collaborative, patient-focused research effort designed to accelerate promising gene therapies to cure SCD [2]. CureSCi's Data Consortium component is creating a robust data resource for SCD researchers, with the goal of cataloguing, harmonizing, and standardizing data collection to advance science to allow for the development of safe and effective SCD curative therapies.

Using standard measures and common data elements for SCD will improve data quality and comparability. This will enhance cross-study analyses for more powerful discovery of smaller effect sizes that may influence patient outcomes or selection of effective treatments and interventions. Many initiatives and resources have been developed to promote the FAIR (Findable, Accessible, Interoperable, Reusable) principles to enhance data sharing and address the data silo challenge, such as the consensus measures for Phenotypes and eXposures (PhenX) Toolkit SCD Research Collection [3], the consensus recommendations for clinical trial end points developed by a partnership of the American Society of Hematology (ASH) and the US Food and Drug Administration [4, 5], the series of ASH guidelines for SCD [6–9], and the CureSCi Common Data Elements (CDEs) Catalog [10]. These resources provide valuable tools and resources for SCD research and trials but are not associated with study datasets.

To maximize the value of completed and ongoing studies, efforts to deposit NHLBI-funded study datasets of patient-level data in a centralized Data Commons at BioData Catalyst (BDC) has just started [11]. This effort will enable combining datasets to ask novel scientific questions and overcome constraints emanating from the study design such as those related to statistical power, subgroup identification, generalizability, and sources of variance. For most studies at BDC, data access is controlled through a data request process. However, identifying relevant studies to request often involves a time-consuming manual review of the publications and metadata from each available study. Using metadata to describe study datasets has been recognized and promoted by the FAIR data principles [12, 13]. Study metadata, including study protocols, data collection forms, and data dictionaries, provide information about the study and its collected variables and is accessible without the formal request required to access the dataset. Curating and indexing the metadata into a browsable and searchable resource would greatly enhance researchers' ability to evaluate appropriate datasets to support the research question. Prominent National Institutes of Health (NIH) Data Repositories and Data Commons, such as the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC), NHLBI BDC, and the database of Genotypes and Phenotypes (dbGaP) at the National Library of Medicine's National Center for Biotechnology Information (NCBI), currently have study level metadata well curated and indexed but lacks the curation and indexing at the common measure or data element level. The NIH Helping to End Addiction Long-term® Initiative, developed a HEAL Data Platform that allows users to discover studies using keyword searches or curated tags of study level metadata related to study setting, study design, data type, subject characteristics, interventions/treatments, substance use, data resources [14]. NIH Environmental influences on Child Health Outcomes (ECHO) has developed a metadata browsing tool (currently only available to the ECHO researchers), with the data element categories organized by a pre-defined set of concepts for environmental influences and child health outcomes [15]. The CureSCi Data Consortium has a relatively small number of SCD-focused studies, making it feasible to take advantage of this unique opportunity to curate and index study variables at the data element and measure levels.

In this report, we describe the design and development of the CureSCi MetaData Catalog (MDC) to make SCD study datasets more Findable by curating study- and variable-level

metadata and making them available through an open-access web portal. To our knowledge, at the time of publication this is the first publicly available tool of its kind, having curated data element level metadata, designed to increase study findability for SCD research. This tool does not offer study datasets. Instead, users can find information about specific data elements across SCD studies before going through data request procedures to access controlled datasets.

Materials and methods

SCD studies in the MDC

The studies included in the CureSCi MDC were initially identified as clinical and observational studies of patients with SCD that were funded by the NIH and NHLBI. There was also the expectation that patient-level study data are or could be made available to the research public through a request and approval process. This initial criterion identified 14 completed or ongoing studies. Three additional SCD registries not funded by NIH were added to the MDC because of their large sample size (i.e., the Sickle Cell Disease Association of America Get Connected Registry) or their comprehensive and longitudinal clinical and patient survey data (i.e., St. Jude's Sickle Cell Clinical Research and Intervention Program Registry and Medical University of South Carolina's South Carolina Sickle Cell Disease Access to Care Pilot Program). The study protocol, data dictionary, and data collection forms and basic information regarding number of patients enrolled or target enrollment, funding source, location of the data, and Principal Investigator contact information were either extracted and downloaded from BioLINCC or dbGaP, or they were requested from each study investigator and made available through the web portal.

Study-level metadata curation

We evaluated other prominent NIH data resources providing study-level information (such as the ClinicalTrials.gov, NHLBI Data Repository at BioLINCC, and the database of Genotypes and Phenotypes [dbGaP]) for their organization of summary-level information about a study. Based on the resources' categorization, categories for the CureSCi MDC were chosen to organize basic information about the studies. Key study-level metadata were curated, including 25 data fields organized into six groups: "General," "Research," "Access," "Study Population," "Documentation," and "Publications." General information, such as study design, study period, and number of subjects, was collected from study protocols and manual of operations. The "Research" group summarized the studies' focus areas and primary outcomes based on information from the literature review, the study investigator, and the ClinicalTrials.gov data resource [16]. The "Access" group included information about data, biospecimen, and genomic data availability from BioLINCC, dbGaP, and BioData Catalyst, as well as information about data and biospecimen consent extracted from study consent forms. The "Study Population" group included information about the age of the participants; the number of subjects; and the inclusion and exclusion criteria obtained from consent forms, protocols, and study publications. For the "Documentation" group, data dictionaries, protocols, and individual data collection forms were collected, and permission to share them in the MDC was obtained. Finally, key publications were selected from a literature review of publications related to each study.

Data element curation

The CureSCi MDC employs a three-tiered conceptual framework to organize and curate study variables. This hierarchical classification system starts with the concept category and is

followed by the subcategory and data element. This concept category hierarchy starts with existing standardized vocabularies and ontologies such as the Medical Subject Headings (MeSH) [17] and the Sickle Cell Disease Ontology (SCDO) [18], and input from domain experts within the CureSC Data Consortium. Expansion and creation of addition data elements are informed by the study variables. Curators evaluate the concept of a variable, and the context of the associated collection form, and map them to the data element categories. For example, "have you visited a hospital in the last month?" can be categorized into two separate categories based on context: "Health Care/Acute Care Utilization" if it's from a Form on assessing quality of care, or "Clinical Status/Respiratory/Acute Chest Syndrome" if it's from a Form assessing "Acute Chest Syndrome". Variables in each study are first cross-referenced with the existing data elements in the MDC to see if they can be categorized into existing categories. If none of the existing data elements are appropriate for the variable at hand, a new data element is created. Some study variables were excluded from the MDC, including (1) study-specific variables related to study administration, (2) variables reporting health information protected by the Health Insurance Portability and Accountability Act that could be used to identify study participants, (3) derived variables that lack documentation in the study forms preventing them from being re-derived as needed, and (4) variables not related to scientific or medical content. All variables selected by curators for exclusion were individually reviewed prior to omission from the MDC.

Patient-Reported Outcome (PRO) measure curation

PRO Measures are derived from outcomes reported by patients. Prominent PRO Measurement systems that evaluate and monitor patients' physical, mental, or social well-being that were reported in MDC-curated studies include the Adult Sickle Cell Quality of Life Measurement Information System (ASCQ-Me®) [19], Patient Reported Outcomes Measurement Information System (PROMIS®) [20, 21], Quality of Life in Neurological Disorders (Neuro-QoL™), and the NIH Toolbox®. These PRO Measures play a key role in assessing SCD outcomes such as pain, quality of care, and quality of life. Using the Health Measures resource [22], PRO Measures were identified and accessioned in the MDC with the measure name and source.

Adoption of controlled vocabularies and ontologies

To increase interoperability, controlled vocabularies and ontologies such as Medical Subject Headings (MeSH) [17] and the Sickle Cell Disease Ontology (SCDO) [18] were referenced and adopted for initial category, subcategory, and data element creation. Domain experts were consulted regarding terms listed in the study documentation, and terms were harmonized when appropriate. Widely used study terms not included in a controlled vocabulary were included as synonyms for terms in the vocabulary. For instance, the SCDO term "SCD Related Pain" (SCDO:0001023) has nested terms for "Chronic Sickle Cell Pain" (SCDO:0000233) and "Sickle Cell Painful Event" (SCDO:0001053) that are synonymous with a pain episode.

Developing the web-based MDC

Best practices regarding database design, data constraints, and data normalization were adopted to optimize performance, scalability, efficiency, and integrity. The CureSCi MDC web portal was developed as a fast, scalable, and interactive multi-platform web application using open-source languages such as.Net Core, JavaScript, and Bootstrap. The web portal and associated database were designed to use HTTPS, a secure and standardized communications protocol, to ensure that metadata are retrievable via the unique identifier and to maximize data

accessibility. Javascript was employed to enable instant dynamic display of selected search filters/facets results.

Results

The most recent release of the CureSCi MDC (Version 2.6, June 30, 2021) includes curated metadata from 16 studies. The studies in the MDC are quite diverse and vary with respect to study design, including clinical trials and case-control, cross sectional, registry, and cohort studies (Fig 1A). Study measure distribution is quite variable with some studies being small and specific (e.g., Hematopoietic Cell Transplant for SCD) whereas others are large and comprehensive (e.g., Cooperative Study of Sickle Cell Diseases) (Fig 1B). Sifting through each study individually would be a difficult and time-consuming task; therefore, the MDC has incorporated several tools to make these studies more Findable.

The CureSCi MDC (https://curesicklecell.rti.org/) offers several browse and search tools. Within a study, users can browse the study-level information in 19 categories with study variables organized in the category/subcategory/data element hierarchy. Across studies, the "Browse Studies by Data Elements" tool allows users to see the number of studies that collected a given category/subcategory/data element. Two search modalities were implemented to allow queries of study-level and data element–level features, by keyword and prepopulated conceptual categories backed by the three-tiered conceptual framework. Search modalities may be used independently or in combination to enhance search capability and maximize discovery of relevant content. To achieve effective and interactive browsing, traditional search techniques were supplemented with a faceted navigation panel by grouping study variables in the three-tiered conceptual framework, allowing users to narrow down search results by applying multiple data elements filters. A tutorial of MDC features is available on the study site,https://curesicklecell.rti.org/player/CureSCi_Metadata_Catalog_v3_player.html. To date, 19 studies have been selected for curation, and metadata from 16 studies have already been uploaded to the web portal, as listed in Table 1.

Browsing by study-level metadata

The "Browse Studies" tool provides access to study-level metadata for the studies included in the MDC. Detailed study information is organized into topics spanning (1) general study

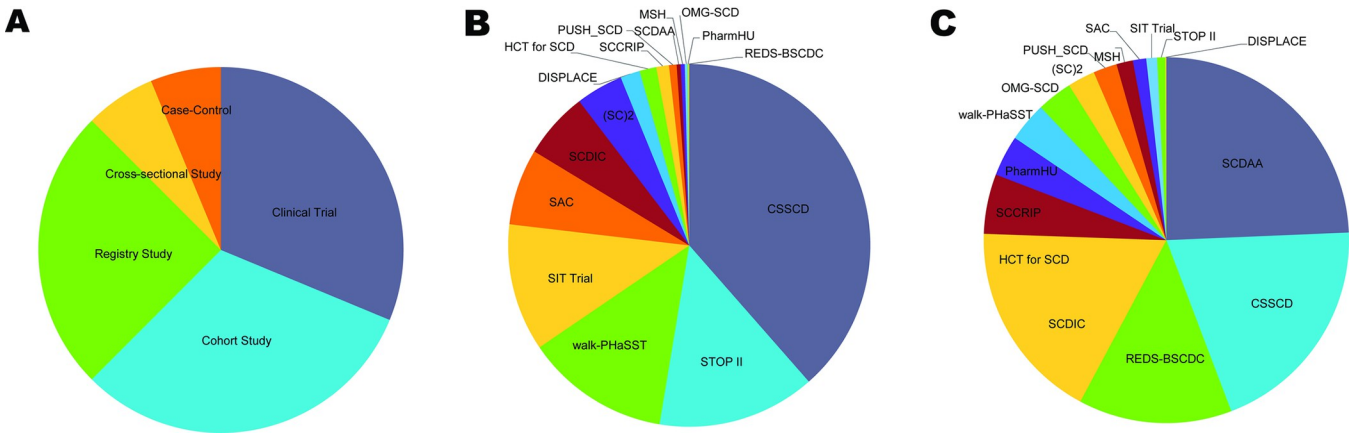

Fig 1. Distribution of 16 curated studies in the Metadata Catalog by experimental design, collected variables, and patient size, available on the CureSCi MDC homepage. A. Proportion of curated studies by study design. **B.** Distribution of study variables (>10,000) across the curated studies. **C.** Distribution of subjects (>20,000) among the curated studies. Republished from https://curesicklecell.rti.org/ under a CC BY license, with permission from Public Library of Science (PLOS), original copyright 2022.

https://doi.org/10.1371/journal.pone.0256248.g001

Table 1. List of SCD studies released in CureSCi MDC.

Study Name	Acronym	Funding
Cooperative Study of Sickle Cell Diseases	CSSCD	NHLBI
Dissemination and Implementation of Stroke Prevention Looking at the Care Environment	DISPLACE	NHLBI
Hematopoietic Cell Transplant for Sickle Cell Disease	HCT for SCD	NHLBI
Howard University Center for Sickle Cell Disease Study	PUSH_SCD	NHLBI
Multicenter Study of Hydroxyurea in Patients With Sickle Cell Anemia	MSH	NHLBI
Optimizing Primary Stroke Prevention in Children with Sickle Cell Anemia	STOP II	NHLBI
Outcome Modifying Genes in Sickle Cell Disease	OMG-SCD	NHLBI
Recipient Epidemiology and Donor Evaluation Study	REDS-BSCDC	NHLBI
Sickle Cell Disease Implementation Consortium	SCDIC	NHLBI
Silent Cerebral Infarct Multi-Center Clinical Trial	SIT Trial	NHLBI
Sleep and Asthma Cohort—II	SAC—II	NHLBI
The Pharmacogenomics of Hydroxyurea in Sickle Cell Disease	PharmHU	NHLBI
Treatment of Pulmonary Hypertension and Sickle Cell Disease With Sildenafil Therapy	walk-PHaSST	NHLBI
South Carolina Sickle Cell Disease Access to Care Pilot Program	$(SC)^2$	Doris Duke
Sickle Cell Disease Association of America—Get Connected Registry	SCDAA	SCDAA
Sickle Cell Clinical Research and Intervention Program	SCCRIP	St. Jude Children's Research Hospital

https://doi.org/10.1371/journal.pone.0256248.t001

information, (2) research overview, focus, and outcomes, (3) access to data or biospecimens, (4) study population description, (5) individual forms and documentation available for viewing, download, or sharing via URL, and (6) data elements associated with this study. The data elements are categorized based on the three-tiered category/subcategory/data element structure. Hovering over each variable reveals a detailed description and values linked to the variable from the study forms.

Browsing by data element

The "Browse Studies by Data Element" tool enables browsing/filtering studies that include a chosen data element selected from a user-friendly expandable menu of the category/subcategory/data element classification framework (Fig 2). This functionality can be extended by clicking "Show Variables" to display study variables mapped to selected data elements, providing information about variable name, description, value set, and the associated study and data collection form (Fig 3).

Browsing by PRO measure

Although data elements capture **what** is collected (e.g., depression severity), PRO Measures capture **how** the data element is collected. To date, 23 different PRO Measures across eight different PRO categories are available spanning outcomes related to functional exercise capacity, mental health, neurodevelopment, neuropsychology, pain, quality of care, and quality of life and sleep. The number of measures will increase, and categories will diversify as more studies are added to the MDC. Although indexed and searchable PRO Measure metadata fields are rarely available in public data resources, the ability to identify study datasets at the measure or instrument level reduces the potential burden of harmonization. Similar to data elements, the "Browse Studies by PRO Measure" tool enables identification of studies that measured study outcomes using selected PRO Measures (Fig 4). Furthermore, data elements and PRO Measures can be combined to refine the study search. For example, it is possible to search for studies that collected the "Pain Attack" data element AND the "ASCQ-Me Pain Episode" PRO

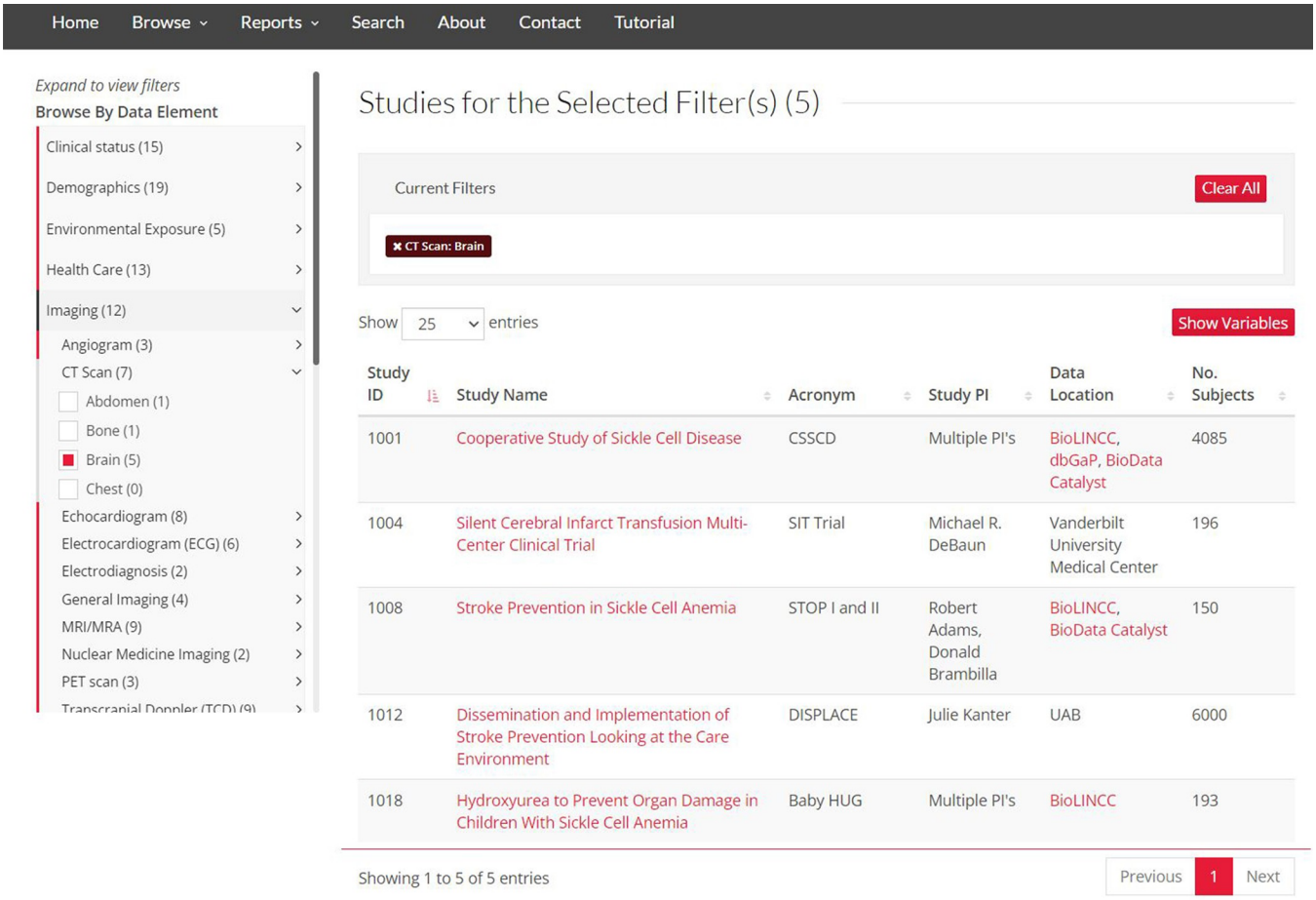

Fig 2. The "Browse Studies by Data Element" tool. The tool dynamically displays studies matching the selections of data elements organized in the three-tiered classification framework.

https://doi.org/10.1371/journal.pone.0256248.g002

Measure. Because many of the SCD PRO Measures were developed in recent years, PRO Measures were found only in recent studies. The power of labeling and indexing PRO Measures will be realized as additional studies adopt their use.

Searching with keywords or concept category filters

The CureSCi MDC includes a keyword and category search tool to help users identify relevant studies. The "Search Study Metadata by Keyword" tool allows users to search for a keyword in study name, study acronym, design, outcome, and data element, among others. The "Search Data Element by Concept Category" leverages the hierarchy of category, subcategory, or data element to select studies. For example, users interested in identifying studies collecting cardiac Magnetic Resonance Imaging (MRI) can select the "Imaging" concept category, "MRI" subcategory, and "Heart" data element. In addition, metadata keyword and data element concept category searches can be combined to refine user-defined searches further (Fig 5). One search example regards the identification of studies and data elements associated with the assessment of pain, a critical attribute in SCD studies. Terms for variables reporting pain vary with different synonyms or classification granularities, as listed in Box 1. Without a structured

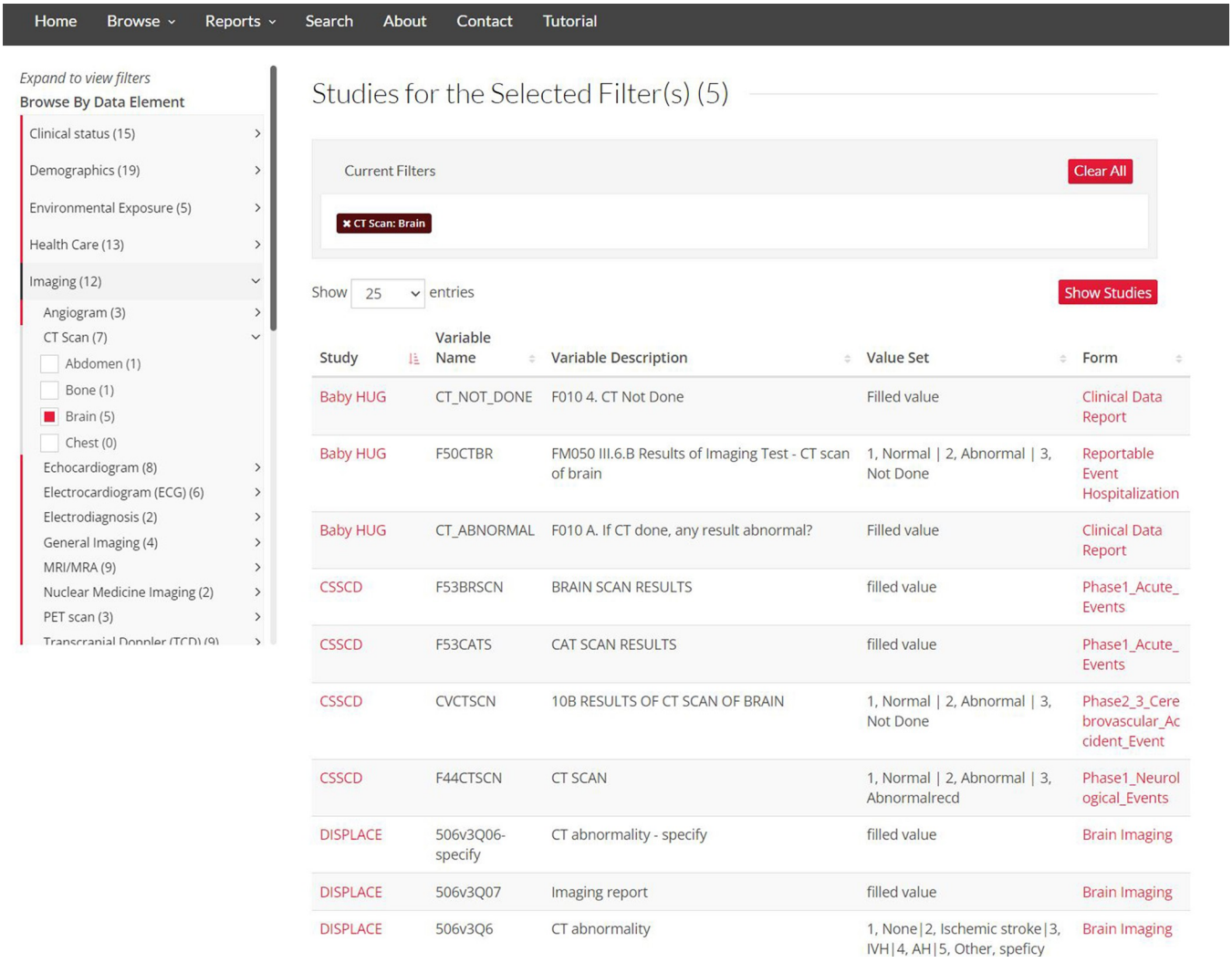

Fig 3. The "Browse Studies by Data Element" tool. The "Show Variable" feature lists study variables information. Republished from https://curesicklecell.rti.org/ under a CC BY license, with permission from Public Library of Science (PLOS), original copyright 2022.

https://doi.org/10.1371/journal.pone.0256248.g003

categorization, a simple keyword search for "pain" could miss data elements collected with a different synonym or granularity. To find studies that collected data for pain within the CureSCi MDC, a user can browse the "Pain" category, revealing six associated subcategories and 49 data elements for 441 variables across six studies. Searching data elements in combination with various study metadata fields can identify variables such as dactylitis that would have been missed by keyword search for "pain".

The multiple search features in the MDC leverage curated and indexed study and data element metadata to enable nested searches combining keywords and multiple concepts in the three-tiered concept framework. These search options provide flexibility and specificity with refined granularity. We provide a search use case to demonstrate these search capabilities, as shown in Fig 5. A search for "hydroxyurea" as a data element keyword resulted in 13 studies. Adding another search criterion of having MRI/Magnetic Resonance Angiography for "heart" narrows the list to two studies. Adding "DNA" to the search study metadata criteria of "Biospecimen" narrowed the list to only one study.

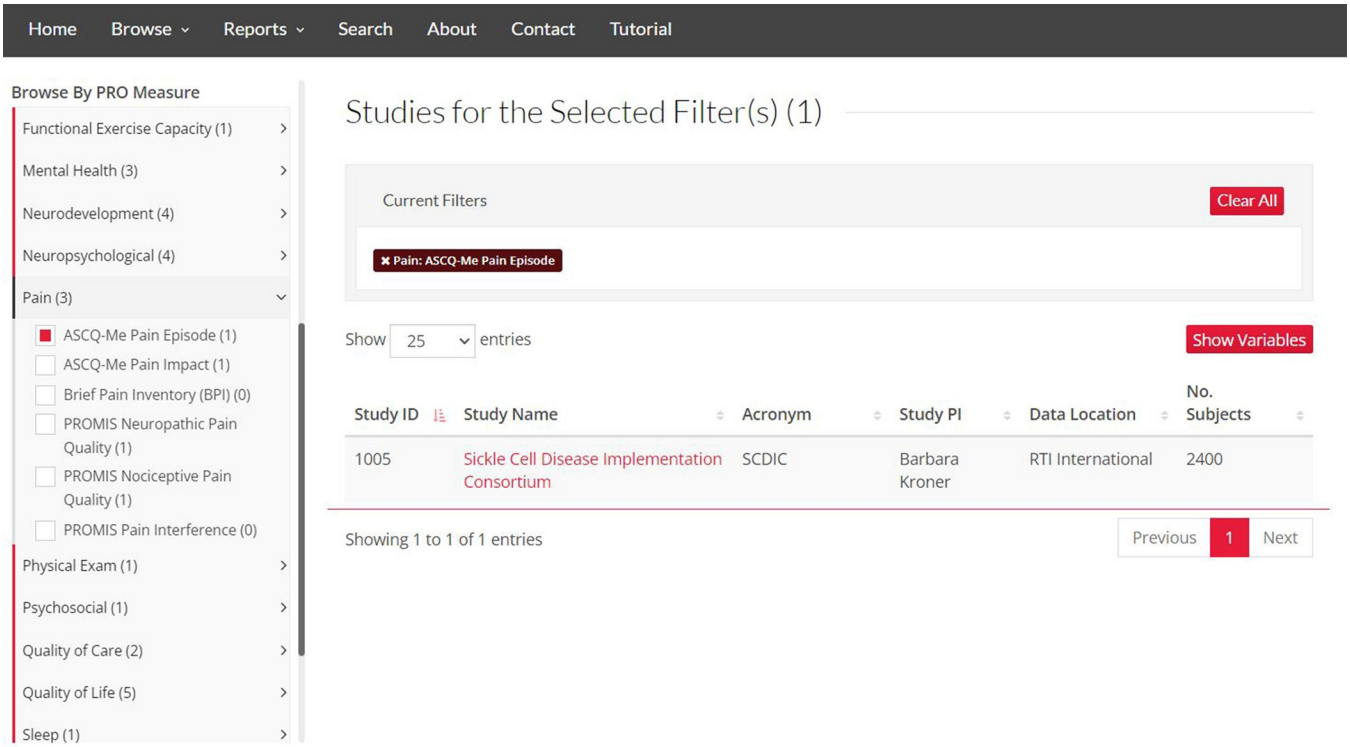

Fig 4. The "Browse Studies by PRO Measures" tool. This tool dynamically displays studies matching the selected PRO Measures. Republished from https://curesicklecell.rti.org/ under a CC BY license, with permission from Public Library of Science (PLOS), original copyright 2022.

https://doi.org/10.1371/journal.pone.0256248.g004

Discussion

Currently, the CureSCi MDC includes curated metadata from 16 different studies, >10,000 study variables, and data collected from >20,000 subjects. This open-access tool allows users to quickly identify collected data from historical studies relevant to current research questions, prior to going through lengthy data request procedures necessary to access the controlled datasets. Study metadata describing collected data can be browsed and searched by keyword, PRO Measures, and ontologies for single or multiple studies. Study metadata were curated to provide standard measures and common data elements through a three-tiered framework consisting of categories, subcategories, and data elements for increased findability. This framework also enables the identification of comparable study variables to support combining similar studies to increase sample size and statistical power and performing cross-study analyses. This way, the MDC provides a method to expedite research studies and data analysis resulting in improvement of the lives of people affected by SCD and supporting paths to cures.

Unique feature of data element level curation

The primary unique feature of the CureSCi MDC is in its variable-level curation. Major NIH data repositories that host various SCD studies, including BioLINCC, BioData Catalyst, and dbGaP, provide information by curating and indexing their datasets based on study-level metadata without the ability to browse and filter at the PRO Measures or variables levels. The CureSCi MDC, a resource solely dedicated to SCD research, fills this gap by curation at the data element (variable) level. Therefore, the MDC offers the flexibility for efficient browsing of SCD research at the study, PRO Measures, and data element levels. Researchers can identify

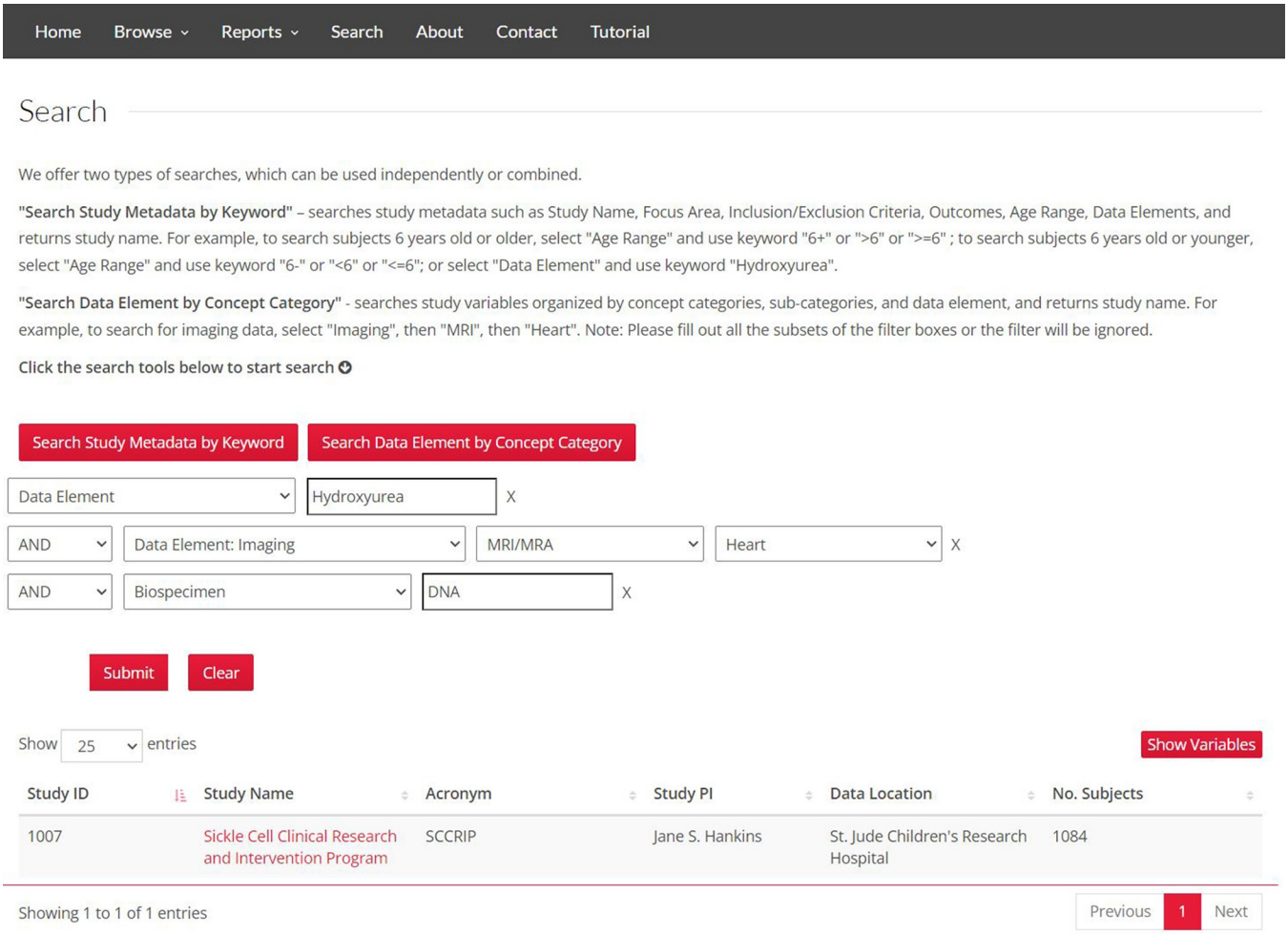

Fig 5. Advanced search combines options to search by keyword and prepopulated conceptual categories. Study identification can be flexibly refined by using keyword searches within a concept category/subcategory/data element or metadata field. The two search modalities may be used independently or in combination with AND/OR/NOT relations. Republished from https://curesicklecell.rti.org/ under a CC BY license, with permission from Public Library of Science (PLOS), original copyright 2022.

https://doi.org/10.1371/journal.pone.0256248.g005

common data elements shared across multiple SCD projects and identify datasets of interest by using the nested search features offered by the MDC. It builds a bridge connecting the CDE resources and the data repositories among the studies for the SCD research community.

Defining, classifying, and refining data elements

A significant curation challenge involves defining a set of common data elements to which study variables can be mapped and organizing the data elements in the hierarchy of categories and subcategories. Selecting a useful level of granularity for each data element is a balancing act. When granularity is too high, variables that might otherwise be meaningfully combined across studies are grouped into separate data elements, limiting the ability to pool variables from multiple studies. Therefore, we have implemented a routine systematic review to combine/ expand data elements, when appropriate, to evaluate balance to allow users to find and evaluate how relevant closely related variables are for testing new hypotheses. We started with a very strict approach with fine granularity that resulted in a long list of concepts collected by

Box 1. A structured categorization of the list of terms for variables reporting pain vary with diverse names, synonyms, or classification granularities to enable browsing and search

- Abdominal Crises
- Abdominal Pain
- Acute Anemic Episodes
- Acute Chest Crisis
- Acute Chest Syndrome
- Acute Pain
- Acute Painful Crisis
- Acute Painful Episodes
- Acute Painful Events
- Allodynia
- Arthritic Pain
- Back Pain
- Breakthrough Pain
- Central Sensitization Syndromes
- Chest Pain
- Chronic Pain
- Chronic Refractory Pain
- Dactylitis
- Headache
- Hyperalgesia
- Leg Ulceration
- Limb Pain and Swelling
- Neuropathic Pain
- Pain Episode
- Painful Crisis
- Pelvis Pain
- Proximal Limb Bone Pain
- Respiratory Crises
- Vaso-Occlusive Crisis
- Very Severe Pain

https://doi.org/10.1371/journal.pone.0256248.t002

only a single study; this concept list has limited utility to identify multiple studies. We recently revised the approach to group multiple related data elements into a single "parent" data element to increase the ability to identify multiple studies, for example, we grouped four data elements (Gallstone Disease, Gastroenteritis, Gastroesophageal Reflux, Gastrointestinal Problem) into "Gastrointestinal Problem". This inclusive revision offers users the flexibility of browsing a list of closely related variables and selecting the subset of variables suitable for their specific topic of interest. Following consultation with experts in the SCD research community, a set of 11 concept categories and 76 subcategories were defined to which data elements were assigned, which was recently revised to 9 concept categories and 57 subcategories. The curators perform periodic reviews for proposed changes to category and subcategory or data element placement within the hierarchy. An important action item at the conclusion of each periodic review is to update the data element assignments for previously curated studies based on the review outcomes. This ensures that search outputs are accurate because variables across all studies are mapped to the current data element definitions.

Standardized vocabulary and ontologies

Existing standardized vocabularies and ontologies were referenced and reused during development of the three-tiered conceptual framework, primarily MeSH and SCDO. MeSH is a comprehensive controlled vocabulary developed by the National Library of Medicine for indexing and searching articles and books in the life sciences [23, 24]. It is widely used for disease and condition classification by many applications and tools such as the MEDLINE/PubMed article database and ClinicalTrials.gov. The hierarchical structure and terminology were adopted and reused in our development when applicable and becomes the backbone of our conceptual framework [23]. The SCDO is a community-driven knowledge representation system for terminology and concepts about SCD [25]. Initial efforts to adopt SCDO structure and naming were met with some challenges. Clinicians and epidemiologists reported that some terminologies, like "Abnormal Phenotype/Abnormality of Cardiovascular System, SCDO:0002245," were awkward and inconsistent with currently used terms. Therefore, we adopted more

commonly used terminologies like "Clinical Status/Cardiovascular," and SCDO terms were included as a cross reference property for interoperability. Adaptions were also made based on the depth and significance of a concept to SCD research. For example, "Pain" is one of the most common symptoms and significantly impacts quality of life and patient ratings for quality of care. Classifying "Pain" as a subcategory under "Clinical Status" would limit the classification of data elements related to the type of pain in the established category/subcategory/data element framework. We therefore elevated "Pain" from a subcategory to the primary category level.

Application to support data harmonization

Retrospective harmonization of existing data can be particularly challenging because it requires accessing comprehensive study documentation and data, designing a mapping approach for heterogeneous data that balances scientific precision and content equivalence with the ability to maximize data inclusion, and implementing analytic methods on harmonized data that carefully consider the potential effects of disparate study designs or populations [26]. The MDC was utilized for a pilot meta-analysis on pain assessment. MDC use enabled discovery and access to detailed study documentation and data collection forms from extant studies essential for identification of suitable studies that collected the data elements required for our analysis. The conceptual framework helped in evaluate which data elements to include based on the number of available data sets and the relationships among the data elements. This exercise demonstrated the MDC's utility in promoting reuse of extant datasets for meta-analysis and prospective study design because it provides investigators of new studies the ability to understand how data concepts and elements were collected in past studies, thus providing tools needed to reduce and support future harmonization.

Process enhancement

Timely curation of new studies into the CureSCi MDC will hasten and encourage data reuse such that the full potential of a research study may be realized. However, mapping study variables to common data elements is currently a manual, multi-step process requiring an intimate knowledge of both the data standards and the study variables. This process can be lengthy depending on the scope and breadth of the study. Automation of key steps involving data dictionary and case report form ingest have already shortened harmonization timelines, and continued efforts will likely result in similar efficiencies.

Integration with other resources

Development of the CureSCi MDC is based on accessible SCD studies. This foundation provides a stable platform to expand and integrate the MDC data elements catalog and PRO Measures with well-established CDE resources, including the PhenX SCD Research Collection, the consensus recommendations for clinical trial end points developed by the ASH and US Food and Drug Administration partnership, the series of ASH guidelines for SCD, and the CureSCi CDE Catalog. This integration will provide researchers with a pivotal link between common SCD data elements with research study datasets. To facilitate dataset exchange, the MDC is building relationships and exploring methods for database interoperability with resources such as BioLINCC and BioData Catalyst, as recently highlighted in "Recent News" at BioLINCC (July 14, 2021) [27]. The MDC provides high-quality manually curated variables, along with the ontologies, which can serve as a robust training set for developing and refining Natural Language Processing algorithms to scale up metadata curation in large data repositories.

Implementing FAIR principles

Since the launch of the CureSCi MDC in April 2020, there have been more than 14,000 views of the MDC. Over the past year, an average of 60 users have accessed the MDC each month. 17% of ever-users are returning users, classified as having multiple visits to the MDC to browse additional studies and access supporting documentation. New datasets are continually being added to the MDC as data generators initiate collaborations with the MDC to incorporate their datasets to the catalog. Beyond making studies more findable, new studies can design data collection with previously used standardized instruments, identified through the MDC, so that the new data can be seamlessly integrated with that of existing studies, promoting data Interoperability and Reuse. Currently we are coupling the features of MDC to inform the BDC tool PIC-Sure to build cohorts and extract participant level data, making data more Accessible. While the MDC is designed to specifically address the Findable) component of FAIR, these examples demonstrate how the MDC touches on each of the various FAIR components.

Promotion of data reuse

The majority of studies included in the MDC have already concluded. Studies currently underway often delay releasing study documentation until after study results are published, which may be several years after data collection has been completed. This creates a time lag where metadata from the most recent and ongoing studies are unavailable for inclusion in the MDC, preventing development and refinement of current CDEs and promotion of more standardized data collection across newly designed studies. With the recent call for using CDEs in multiple NIH Funding Opportunity Announcements [28–33], it is now more important than ever to have an up-to-date resource to find CDE information that can streamline study design, promote dataset interoperability, and ultimately increase integration with other relevant resources. This resource depends on the collaboration and cooperation of study investigators to release their study metadata in a timely manner. Use of the CureSCi MDC is intended to facilitate the SCD research and therapeutic community to identify studies by data element for collaborative research and promote data reuse for future SCD studies. The benefits of this tool to the community will be continually evaluated and feedback incorporated to ensure the highest benefit to advance SCD research.

Sustainability of the CureSCi MDC

Through the support of the NHLBI and the CureSCi, new research studies are underway and will be included in the MDC. Continual outreach to the SCD research community, including outside the NIH, through publications and conference presentations has already and will continue to establish collaborations, resulting in a continual stream of studies to incorporate into the MDC, including CDC funded studies (Public Health Research, Epidemiology, and Surveillance for Hemoglobinopathies [PHRESH], Registry and Surveillance System for Hemoglobinopathies [RuSH], Sickle Cell Data Collection [SCDC] Program). As new study results become available, it is critical for the metadata from these studies to be incorporated into the CureSCi MDC. An ingestion engine will promote the efficient upload of documents from study investigators and ensure that the necessary format for curation is present. To keep up with the influx of studies it will be essential to be able to scale study curation. One approach to addressing this issue is to use the existing curation data to train a machine learning model to perform the initial curation, which can then be reviewed by human experts for accuracy and clarity. Enhancement and refinement of features will also continue in response to the needs of the SCD research community and user feedback.

Conclusion

The CureSCi MDC was designed to be a resource for curation of data elements and study documentation for past and current studies of SCD populations. The web-based portal presented in this manuscript provides a platform to make SCD research more Findable and promote FAIR data principles, allow investigators to browse and search metadata from existing studies to determine what data elements were collected and are common in SCD research, and finally, to identify compatible studies for cross-study analyses. The CureSCi MDC is a living entity and not a static resource. Providing timely access to new study documentation from the SCD community will help sustain the utility of this resource. Ultimately, the CureSCi MDC will aid in the appropriate reuse of existing data to answer pertinent research questions and the expedited design of new studies with measures and outcomes that are compatible with cross-study analysis efforts.

Acknowledgments

The authors acknowledge the contributions of the CureSCi Data Consortium members as well as Michael DeBaun at Vanderbilt University Medical Center and Josh Richardson and Jorgen Waldermo at RTI international. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

Conceptualization: Huaqin Pan, Cataia Ives, Meisha Mandal, Ying Qin, Donald Brambilla, Jeran Stratford, Marsha Treadwell, Xin Wu.

Data curation: Huaqin Pan, Cataia Ives, Meisha Mandal, Jeran Stratford, Xin Wu.

Formal analysis: Huaqin Pan, Meisha Mandal, Ying Qin, Jeran Stratford.

Funding acquisition: Barbara Kroner.

Investigation: Huaqin Pan, Cataia Ives, Meisha Mandal, Xin Wu, Barbara Kroner.

Methodology: Huaqin Pan, Cataia Ives, Meisha Mandal, Ying Qin, Jen Popovic, Xin Wu.

Project administration: Huaqin Pan, Cataia Ives, Tabitha Hendershot.

Resources: Cataia Ives.

Software: Cataia Ives, Ying Qin.

Supervision: Huaqin Pan.

Validation: Huaqin Pan, Cataia Ives, Meisha Mandal.

Visualization: Huaqin Pan, Cataia Ives, Jeran Stratford, Xin Wu.

Writing – original draft: Huaqin Pan, Cataia Ives, Ying Qin, Tabitha Hendershot, Jen Popovic, Jeran Stratford, Marsha Treadwell, Xin Wu.

Writing – review & editing: Huaqin Pan, Cataia Ives, Ying Qin, Tabitha Hendershot, Jen Popovic, Donald Brambilla, Jeran Stratford, Marsha Treadwell, Xin Wu, Barbara Kroner.

References

1. Hassell KL. Population estimates of sickle cell disease in the U.S. Am J Prev Med. 2010; 38(4 Suppl): S512–21. https://doi.org/10.1016/j.amepre.2009.12.022 PMID: 20331952
2. Curesickle.org [Internet]. It's time to rewrite the story of sickle cell. Available from https://curesickle.org/

3. Eckman JR, Hassell KL, Huggins W, Werner EM, Klings ES, Adams RJ, et al. Standard measures for sickle cell disease research: the PhenX Toolkit sickle cell disease collections. Blood Adv. 2017 Dec 15; 1(27):2703–11. https://doi.org/10.1182/bloodadvances.2017010702 PMID: 29296922
4. Farrell AT, Panepinto J, Carroll CP, Darbari DS, Desai AA, King AA, et al. End points for sickle cell disease clinical trials: patient-reported outcomes, pain, and the brain. Blood Adv. 2019 Dec 10; 3 (23):3982–4001. https://doi.org/10.1182/bloodadvances.2019000882 PMID: 31809538
5. Farrell AT, Panepinto J, Desai AA, Kassim AA, Lebensburger J, Walters MC, et al. End points for sickle cell disease clinical trials: renal and cardiopulmonary, cure, and low-resource settings. Blood Adv. 2019 Dec 10; 3(23):4002–20. https://doi.org/10.1182/bloodadvances.2019000883 PMID: 31809537; PMCID: PMC6963248.
6. Monagle P, Cuello CA, Augustine C, Bonduel M, Brandão LR, Capman T, et al. American Society of Hematology 2018 Guidelines for management of venous thromboembolism: treatment of pediatric venous thromboembolism. Blood Adv. 2018 Nov 27; 2(22):3292–3316. https://doi.org/10.1182/bloodadvances.2018024786 PMID: 30482766
7. Liem RI, Lanzkron S, D Coates T, DeCastro L, Desai AA, Ataga KI, et al. American Society of Hematology 2019 guidelines for sickle cell disease: cardiopulmonary and kidney disease. Blood Adv. 2019 Dec 10; 3(23):3867–97. https://doi.org/10.1182/bloodadvances.2019000916 PMID: 31794601
8. Brandow AM, Carroll CP, Creary S, Edwards-Elliott R, Glassberg J, Hurley RW, et al. American Society of Hematology 2020 guidelines for sickle cell disease: management of acute and chronic pain. Blood Adv. 2020 Jun 23; 4(12):2656–2701. https://doi.org/10.1182/bloodadvances.2020001851 PMID: 32559294
9. Chou ST, Alsawas M, Fasano RM, Field JJ, Hendrickson JE, Howard J, et al. American Society of Hematology 2020 guidelines for sickle cell disease: transfusion support. Blood Adv. 2020 Jan 28; 4 (2):327–55. https://doi.org/10.1182/bloodadvances.2019001143 PMID: 31985807
10. Curesickle.org [Internet]. CDE Catalog. Available from: https://curesickle.org/cde-catalog
11. NHLBI [Internet]. BioData Catalyst: Access biomedical data when you need it and how you need it. Available from: https://biodatacatalyst.nhlbi.nih.gov/
12. Wilkinson MD, Dumontier M, Jan Aalbersberg I, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15; 3:160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244
13. Wilkinson MD, Dumontier M, Jan Aalbersberg I, Appleton G, Axton M, Baak A, et al. Addendum: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2019; 6(6). https://doi.org/10.1038/s41597-019-0009-6 PMID: 30890711
14. NIH HEAL Initiative® HEAL Data Platform [Internet]. Available from: https://healdata.org/landing
15. The Environmental influences on Child Health Outcomes (ECHO) Portal, [Internet]. Available from: https://echoportal.org
16. ClinicalTrials.gov [Internet]. Available from: https://clinicaltrials.gov/ct2/home
17. National Institutes of Health [Internet]. National Library of Medicine: Medical Subject Headings. Available from: https://www.nlm.nih.gov/mesh/meshhome.html
18. BioPortal [Internet]. Sickle Cell Disease Ontology; c2021. Available from: https://bioportal.bioontology.org/ontologies/SCDO
19. Evensen CT, Treadwell MJ, Keller S, Levine R, Hassell KL, Werner EM, et al. Quality of care in sickle cell disease: Cross-sectional study and development of a measure for adults reporting on ambulatory and emergency department care. Medicine. 2016; 95(35):e4528. https://doi.org/10.1097/MD.0000000000004528 PMID: 27583862
20. National Institutes of Health [Internet]. Patient-reported outcome measurement information system; c2019. Available from: https://commonfund.nih.gov/promis/index
21. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): depression, anxiety, and anger. Assessment. 2011; 18(3):263–83.
22. Health Measures [Internet]. Home page; c2021. Available from: https://www.healthmeasures.net/index.php
23. Mao Y, Lu Z. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. J Biomed Semantics. 2017 Apr 17; 8(1):15. https://doi.org/10.1186/s13326-017-0123-3 PMID: 28412964
24. Huang M, Névéol A, Lu Z.Recommending MeSH terms for annotating biomedical articles. J Am Med Inform Assoc. 2011; 18(5):660–7. https://doi.org/10.1136/amiajnl-2010-000055 PMID: 21613640

25. Sickle Cell Disease Ontology Working Group. The Sickle Cell Disease Ontology: enabling universal sickle cell-based knowledge representation. Database. 2019; 2019: baz118. https://doi.org/10.1093/database/baz118 PMID: 31769834

26. Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2017 Feb 1; 46(1):103–105. https://doi.org/10.1093/ije/dyw075 PMID: 27272186

27. National Heart, Lung, and Blood Institute [Internet]. News: NHLBI Cure Sickle Cell Initiative; c2021. Available from: https://biolincc.nhlbi.nih.gov/news/#post-320

28. National Institutes of Health. New Cohorts for Environmental Exposures and Cancer Risk (CEECR) Coordinating Center (U24 Clinical Trial Not Allowed). Funding Opportunity Announcement (FOA) Number RFA-CA-20-050. https://grants.nih.gov/grants/guide/rfa-files/RFA-CA-20-050.html (2020).

29. National Institutes of Health. Rare Disease Cohorts in Heart, Lung, Blood and Sleep Disorders (UG3/UH3 Clinical Trial Not Allowed). Funding Opportunity Announcement (FOA) Number RFA-HL-20-014. https://grants.nih.gov/grants/guide/rfa-files/rfa-hl-20-014.html (2019).

30. National Institutes of Health (NIH). Secondary Analyses and Archiving of Social and Behavioral Datasets in Aging (R03). Funding Opportunity Announcement (FOA) Number RFA-AG-12-005. https://grants.nih.gov/grants/guide/rfa-files/rfa-ag-12-005.html (2011).

31. National Institutes of Health (NIH), U.S. Food and Drug Administration (FDA). Secondary Analyses of Existing Datasets of Tobacco Use and Health (R21 Clinical Trial Not Allowed). Funding Opportunity Announcement (FOA) Number RFA-OD-21-003. https://grants.nih.gov/grants/guide/rfa-files/RFA-OD-21-003.html (2021).

32. National Institutes of Health (NIH). Secondary Analysis and Integration of Existing Data to Elucidate the Genetic Architecture of Cancer Risk and Related Outcomes (R01 Clinical Trial Not Allowed). Funding Opportunity Announcement (FOA) Number PAR-20-276. https://grants.nih.gov/grants/guide/pa-files/PAR-20-276.html (2020).

33. National Institutes of Health (NIH). Secondary Analysis of Existing Datasets in Heart, Lung, and Blood Diseases and Sleep Disorders (R21 Clinical Trial Not Allowed). Funding Opportunity Announcement (FOA) Number PAR-20-078. https://grants.nih.gov/grants/guide/pa-files/PAR-20-078.html (2019).