blood advances

ASH Research Collaborative: a real-world data infrastructure to support real-world evidence development and learning healthcare systems in hematology

William A. Wood,[1] Peter Marks,[2] Robert M. Plovnick,[3] Kathleen Hewitt,[4] Donna S. Neuberg,[5] Sam Walters,[6] Brendan K. Dolan,[7] Emily A. Tucker,[4] Charles S. Abrams,[8] Alexis A. Thompson,[9] Kenneth C. Anderson,[10] Paul Kluetz,[2] Ann Farrell,[2] Donna Rivera,[2] Matthew Gertzog,[3] and Gregory Pappas[2]

[1]Division of Hematology, Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC; [2]U.S. Food and Drug Administration, Silver Spring, MD; [3]The American Society of Hematology, Washington, DC; [4]ASH Research Collaborative, Washington, DC; [5]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA; [6]Breakthrough Healthcare, Baltimore, MD; [7]The University of Wisconsin School of Medicine and Public Health, Madison, WI; [8]Department of Medicine, University of Pennsylvania, Philadelphia, PA; [9]Department of Pediatrics, Northwestern University, Chicago, IL; and [10]Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA

Key Points

- The ASH Research Collaborative includes a patient-level data platform for SCD and MM, expanding to other conditions in the future.
- The ASH Research Collaborative gathers input from patients, clinicians, researchers, industry, and government representatives.

The ASH Research Collaborative is a nonprofit organization established through the American Society of Hematology's commitment to patients with hematologic conditions and the science that informs clinical care and future therapies. The ASH Research Collaborative houses 2 major initiatives: (1) the Data Hub and (2) the Clinical Trials Network (CTN). The Data Hub is a program for hematologic diseases in which networks of clinical care delivery sites are developed in specific disease areas, with individual patient data contributed through electronic health record (EHR) integration, direct data entry through electronic data capture, and external data sources. Disease-specific data models are constructed so that data can be assembled into analytic datasets and used to enhance clinical care through dashboards and other mechanisms. Initial models have been built in multiple myeloma (MM) and sickle cell disease (SCD) using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) and Fast Healthcare Interoperability Resources (FHIR) standards. The Data Hub also provides a framework for development of disease-specific learning communities (LC) and testing of health care delivery strategies. The ASH Research Collaborative SCD CTN is a clinical trials accelerator that creates efficiencies in the execution of multicenter clinical trials and has been initially developed for SCD. Both components are operational, with the Data Hub actively aggregating source data and the SCD CTN reviewing study candidates. This manuscript describes processes involved in developing core features of the ASH Research Collaborative to inform the stakeholder community in preparation for expansion to additional disease areas.

Introduction

The cost and time to generate evidence in hematology represent barriers to progress in research and transformation of clinical practice.[1] However, digitalization of health care data and the availability of clinical, administrative, laboratory, and patient-reported information derived from routine clinical care represent

Submitted 9 August 2021; accepted 2 October 2021; prepublished online on *Blood Advances* First Edition 21 October 2021; final version published online 13 December 2021. DOI 10.1182/bloodadvances.2021005902.

Requests for data sharing may be submitted to William Wood (wawood@med.unc.edu).

opportunities to accelerate evidence generation. These opportunities are supported by recent legislation, including the 21st Century Cures Act and the reauthorization of the Prescription Drug User Fee Act (PDUFA VI), that have encouraged trial modernization and the evaluation of data from sources outside of traditional clinical trials to support regulatory decision making.[2]

This report describes the ASH Research Collaborative and its primary components, the Data Hub and the Sickle Cell Disease Clinical Trials Network (SCD CTN). The development and implementation of the initiative's methods for data collection and tools for users of the data are reviewed, and initial use cases for the generation of real-world evidence (RWE) and the improvement of clinical care are discussed. The report concludes with directions for the future evolution of the ASH Research Collaborative.

Development of the ASH Research Collaborative

The ASH Research Collaborative was founded as a nonprofit organization by the American Society of Hematology in 2018, to improve the lives of people affected by blood diseases by enhancing research and clinical practice. The Data Hub aggregates curated data from multiple sites for a variety of uses by researchers, providers, and other stakeholders, while the SCD CTN engages institutions to introduce efficiencies in multicenter clinical research.

The ASH Research Collaborative is governed by an executive committee comprised of appointed members, with subcommittees, working groups, and task forces that support specific areas of focus. The Data Hub oversight group and the SCD CTN oversight committee provide programmatic guidance. Though the initial focus for the Data Hub and SCD CTN is initially in multiple myeloma (MM) and sickle cell disease , early work in these areas will develop templates that will be used for future program expansion.

Data Hub

The Data Hub is building a shared information resource for the global hematology community. Clinical care sites, such as health care systems, hospitals, and outpatient practices, can participate in the Data Hub and establish EHR data integration that facilitates data submission. Data submission formats for structured EHR data include Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) and Fast Healthcare Interoperability Resource (FHIR) standards. OMOP and FHIR, terms that may be unfamiliar to many hematologists, refer to 2 contemporary and increasingly used clinical informatics models to extract and/or transmit data from the electronic health record. Based on the Data Hub experience to date, most participating institutions have information technology and informatics expertise, and in some cases, have approached these methods to organize and transmit data for other data aggregation efforts outside of hematology. As new transmission methods for data submission become available, such as application programming interfaces (APIs) between third-party applications and the EHR, these opportunities will be accepted by the Data Hub as well. Besides EHR data, other data sources will be incorporated into the Data Hub such as patient-reported outcomes (PROs) from patient-facing portals and apps, patient-generated health data (PGHD) from connected digital sensors, case report form-based electronic data capture for clinical information maintained outside structured EHR fields, genomic and molecular data, and population data from a variety of sources. In some instances, data specifications, including accepted formats, vocabularies, and standards including details related to data acquisition, curation, and linking to external sources, are still evolving.

For example, genomic data could include either "ground truth" data files or interpreted data summaries, but details related to data acquisition, curation, and linking to external sources remain to be determined. The Data Hub plans to facilitate the collection of PROs and PGHD through patient-facing portals; because patients are linked to records at the site level, data can be simultaneously collected directly from patients while being shared back to sites for clinical care and site-level population analyses. Exploration of data integration from industry, government datasets, registries, and other U.S. or international sources is also planned, with assessments of feasibility, appropriateness, and fitness for purpose in different contexts.

The multistakeholder approach to building the Data Hub

The ASH Research Collaborative adopted a multistakeholder collaborative philosophy that includes patients and community members, clinical care sites, research sponsors, clinicians, researchers, and federal entities, including the U.S. Food and Drug Administration (FDA), National Institutes of Health (NIH), U.S. Department of Health and Human Services (HHS), and others. The ASH Research Collaborative Data Hub works closely with patients to understand priorities, incentives, and barriers to research participation. The patient community is recognized as a key driver to the Data Hub's ability to enable the broadest possible use of real-world data (RWD) to accelerate evidence development, advance research, and improve care. An electronic consent platform has been developed with patient input, modeled after best practices that have been demonstrated in the National Institutes of Health's All of Us research program.[3]

Patients have also been integrally involved in key ASH Research Collaborative projects, including the SCD Learning Community (LC), which aims to iteratively inform Data Hub design alongside the particular clinical improvement priorities identified. Patients will become increasingly involved as PRO and PGHD capabilities are identified and integrated. At this point, direct patient querying of aggregate Data Hub data is not anticipated but could be considered in the future. Nonetheless, patients will continue to have an important voice in Data Hub development and execution so that their priorities are identified and implemented. Providing sites with data tools that help them directly assess their clinical care and patient outcomes and engaging a network of care providers who exchange best practices and lessons learned are key elements that support care enhancements and program sustainability. Tools provided to participating sites include real-time dashboards that provide unique site care and outcome metric results and comparisons to aggregate data from all sites. Sites also have access to their own data using online queries and cohort analyzer tools and data exports for local research. As the Data Hub grows, researchers can submit research proposals to access deidentified data across all participating sites. To preserve confidentiality for participating sites, clinicians, and patients, research data sets exclude site and clinician identifiers and include a limited data set or deidentified data set for analysis. The level of identification is determined on a case-by-case basis and complies with institutional review board (IRB) oversight and the Data Hub's governance policy.

For research sponsors, the Data Hub can provide access and analysis of RWD for use in contemporaneous control arms, postapproval surveillance, and identification of cohorts for pre and postmarket research.

The ASH Research Collaborative is working to align its interests with those of federal agencies in promoting the development and access to safe, efficacious, and patient-centered therapies by generating new clinical and translational research discoveries and promoting equity in access to evidence-based care and novel therapies throughout the field of hematology.

The following paragraphs describe in more detail the rationale for the inclusion of the initial 2 hematologic conditions (SCD and MM) and will use these 2 areas to illustrate overall Data Hub operations and strategy.

SCD

ASH has a long-standing commitment to addressing disparities associated with SCD and facilitating the development of therapeutic strategies. The SCD drug development pipeline is active,[4] and FDA-regulated research includes small molecules, monoclonal antibodies, gene therapies, and other approaches to ameliorate disease-related symptoms and to achieve a cure. As improvements in SCD have extended survival for affected children and young adults, new questions are being developed about the course of the disease in older individuals and/or those with end-organ dysfunction. The ASH Research Collaborative's patient-centered approach and community engagement are important program elements to answer these questions and to accelerate clinical trials and accumulation of RWD with longitudinal follow-up.

MM

MM was selected as the initial disease area of focus within malignant hematology as data sharing opportunities utilizing RWD in this disease are particularly timely. Improvements in current prognostic models (eg, an improved version of the Revised International Staging System) and the development of new prognostic biomarkers (eg, the role of minimal residual disease to guide evaluation and selection of therapies) require large datasets for development and validation.[5,6] The refinement of genomic markers for risk stratification and treatment selection continues to evolve. In the setting of recently approved treatments for MM, sequencing of therapies and the optimal role of autologous hematopoietic cell transplantation and CAR-T cell therapies remain unclear and will benefit from the accumulation of RWD over long periods of time.[7] Further, racial and ethnic barriers[8,9] to clinical trial enrollment and access to effective therapies outside of trials mandate creative problem-solving approaches.

Data model development

Though full EHR data are captured within the Data Hub for each included participant, each disease-specific program within the Data Hub has a harmonized data model that includes evidence-based data elements with accompanying validation rules. Here, data harmonization is referring to the way in which data with different formatting, naming, and organization frameworks can be brought together and transformed into a cohesive data set to facilitate "apples to apples" comparisons for visualizations and analysis. The harmonized data models do not limit the scope of full EHR data ingestion but support the Data Hub's ability to optimize the use of the data. Core data elements are iteratively updated over time in parallel with new advances in clinical care and research and are derived from a common construction methodology and commonly shared data language(s). Closely tied to the core data elements are metrics that use specific algorithms to align with outcomes of interest.

The Data Hub core data elements and metrics are established by an iterative process and are intended to be informed by a wide variety of drug development stakeholders. A modified Delphi process that engages patients, clinicians, regulators, payers, health technology assessment (HTA) groups, drug developers, and other key stakeholders as used for other initiatives (eg, coreSCD)[10] is being tailored and optimized within the ASH Research Collaborative Data Hub to identify priority metrics and outcomes across decision contexts for regulatory purposes, coverage and reimbursement, and patient care.

Similar approaches that consider multiple stakeholders and sources of data are being used to identify core data elements in SCD and MM and are planned for other malignant and nonmalignant hematologic diseases. The core data elements, metrics, and overall data model development process is coordinated across ASH Research Collaborative multistakeholder subcommittees.

Data elements, e-phenotyping, and metric development

The development of a RWD platform for both SCD and MM presents a set of challenges common to other RWD initiatives and unique to rare diseases. There are limitations with EHR data consistency, accuracy, and completeness that require multidisciplinary approaches for analysis, verification, and augmentation via linkage to secondary data sources. Rare diseases like SCD are subject to chronic miscoding due to the highly specialized knowledge required for accurate diagnosis and treatment. As with other clinical data points, SCD and MM RWD are most valuable when analyzed using complex computable phenotypes that incorporate data across domains, including but not limited to laboratory results, diagnoses, procedures, visits, and imaging studies. The concept of a computable phenotype includes the composite of data elements, and representative codes obtained longitudinally at multiple clinical encounters can be brought together through a value set (a set of codes where "any one counts") for identification and uses rules-based algorithms ("if/then"), queries, and/or machine learning (ML)/artificial intelligence (AI) (agnostic to what underlying relationships might exist) to represent a single health concept. The relationship between the computable phenotype and the known underlying health concept (gold standard) is measurable with traditional test characteristics. The availability and reliability of these data vary across and within health systems, necessitating supplemental data collection and validation approaches to enhance the quality of RWD to ensure fitness for purpose. The SCD and MM Data Hub programs are being developed by hematology experts who examined disease-specific endpoints, clinical guidelines, and other evidence documents to inform the creation of clinical core data elements and metrics of interest. Working with clinical informaticists, hematologists are assessing the reliability and validity of structured EHR data to determine additional manual verification or data entry needs. Baseline computable phenotypes (e-phenotypes) for clinical concepts are constructed using value sets comprised of clinical terms derived from EHR documentation codes such as ICD-10, SNOMED-CT, LOINC, and RxNorm. Where possible, published value sets from the Value Set Authority Center (VSAC)[11] are used. The VSAC is a central repository of clinical concepts and their associated terminology definitions, hosted by the National Library of Medicine. Value sets in the VSAC are developed by a variety of health care entities, and value sets authored by national organizations and frequently updated have been prioritized.

Future Data Hub projects will include the development of a comprehensive phenotype knowledge base[12] with further collaboration to develop and share e-phenotypes for broad use. The Data Hub envisions the development and maturity of e-phenotypes across stages. In the first stage, an inclusive list of codes is generated to achieve acceptable sensitivity for capturing the health concept of interest. In the second stage, rules-based logic is used to develop algorithms of codes to further improve sensitivity, specificity, and positive predictive value of the underlying health concept (eg, one of several diagnosis codes, repeated in at least 2 outpatient visits, accompanied by 1 or more from a set of laboratory and imaging results, and exclusive of other specific codes from the EHR). In the third stage, ML and AI can be applied to agnostically determine other contributors to e-phenotypes where prespecified rules may not exist. Not every stage may be required for every e-phenotype; the amount of development and validation required depends on consensus standards across the stakeholder community as well as the needs of specific use cases. As this area is not yet well defined, the Data Hub will contribute to efforts involving the FDA and others to help define when an e-phenotype is "fit for purpose" for inclusion within a disease program.

The Data Hub is constructing an e-phenotype "innovation lab" where Data Hub participants can review results of proposed e-phenotypes using their own data and provide feedback to guide refinement prior to publication as proposed national standards. An example of a first stage ASH Research Collaborative e-phenotype using an inclusive list of EHR codes for vaso-occlusive episodes is shown in supplemental Table 1. The concept at this stage is to find all potentially applicable codes that might be used to identify an individual with SCD in order to improve the sensitivity (true positive) of the technique for picking up all potential patients. Additional work will be conducted as needed, with new data elements and/or repeated measures such as laboratory values and other diagnoses to further refine this e-phenotype using second-stage and third-stage processes to improve specificity and positive predictive value of the resulting patient identification algorithm.

Supplemental data capture mechanisms

In addition to data directly derived from the EHR, we identified high-value data elements that cannot be reliably extracted from EHRs due to variable documentation across sites. For example, these could include concepts such as genotype in SCD, or disease progression (or "time to treatment failure" [TTF]), or the Revised International Staging System in MM. Though the data elements undergirding these concepts could be present in the EHR (in myeloma, the serum protein electrophoresis, free light chain values, lactate dehydrogenase, β 2 microglobulin, albumin, etc), they may not adequately or reliably be captured for a given patient at the primary institution contributing data. Or, short of application of natural language processing, FISH/cytogenetic data may not be immediately accessible through the EHR data because of the format for original capture of the data. When needed, supplemental data such as the examples provided here have formed the basis for electronic Case Report Forms (eCRFs) for each disease. The Data Hub's eCRFs were designed to meet several program goals: (1) to reliably capture data elements not well-structured or present within EHRs; (2) to allow sites who are unable to connect their EHRs to submit data to the Data Hub to access Data Hub data for benchmarking and research; and (3) to provide a mechanism for amending EHR data that was miscoded, missing, or otherwise inaccurate in their EHR.

Supplemental eCRF data will be critical in the calculation of metrics such as TTF for MM, where the clinical data necessary to determine treatment initiation and criteria for treatment failure were not consistently coded or readily available from most EHRs. While the Data Hub will address current gaps in clinically relevant EHR data capture through eCRFs, the Data Hub will also work to facilitate the development and implementation of EHR clinical documentation standards in order to reduce redundancies and inefficiencies associated with duplicate data entry.

In addition to data captured via EHR and eCRF, the Data Hub can also receive data extracted from sites' local research programs and submitted via a simple flat-file format. Sites that submit local research data will harmonize data element definitions and map fields to the Data Hub's standardized data dictionary to facilitate reliable meta-analyses. Figure 1 is a representative schematic for pathways to Data Hub data submission and validation.

Data quality can be challenging for EHR RWD because data are documented by a variety of providers across care settings and systems. An important goal for the ASH Research Collaborative is to provide sites with a data quality report with each data submission. Site data quality reports focus on areas of high variability in EHR data to help sites identify potential concerns surrounding data completeness and accuracy. If data quality issues are identified, it may not be possible to reconcile EHR data with Data Hubstandards. In this case, sites can use prepopulated eCRFs to amend a patient's record in the Data Hub. Subsequent EHR transmissions will not overwrite data that has been amended through the site's eCRF (eg, the eCRF is treated as the primary source of truth when data are in conflict). Future data quality assessments will include crossreferencing longitudinal chart data to identify inconsistencies, duplication, and potential miscoding that impact the accuracy of metrics and research analyses.

Site dashboards

The Data Hub's site dashboards provide real-time access to metric results for the site submitting data with deidentified aggregate comparison results across all submitting sites. Because measurement needs vary across expert groups with regard to SCD and MM care, the Data Hub has developed different data visualization and analysis approaches for the SCD and MM dashboards. For example, the majority of the SCD dashboard data are best understood in the context of trends over time for specific and repeatable metrics largely derived from structured EHR data, with a need to identify potential outliers for further investigation. See Figure 2 for a list of metrics included in the SCD site dashboard. The SCD site dashboard uses p-charts to illustrate each site's trend for a metric, with an associated

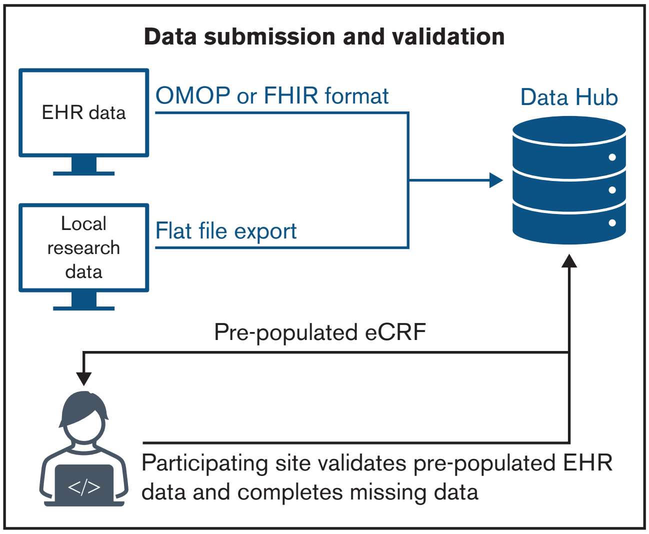

Data submission and validation
EHR data
OMOP or FHIR format
Data Hub
Local research data
Flat file export
Pre-populated eCRF
</>
Participating site validates pre-populated EHR data and completes missing data

Figure 1. The ASH RC Data Hub model for data submission and validation.

mean value across the analysis period and 1, 2, and 3 standard deviations to show potential outliers in need of further analysis. In contrast, the MM site dashboard (Figure 3) uses a blend of descriptive statistics such as frequencies and means, as well as Kaplan-Meier survival curves to analyze metrics of interest, initially overall survival (OS) and TTF. The MM site dashboard also includes multiple filters to create subgroups based on complex logical relationships between data elements such as treatments, disease type and subtype, disease history, and the presence or absence of specific genetic aberrations. Site data can be visualized alongside aggregate data from all other sites participating in the Data Hub, in entirety or by filtered subgroups. For both site dashboards, when a cohort of interest has been specified for a set of metrics, sites can drill down to participant-level detailed data to facilitate advanced statistical analysis and quality improvement activities. Site dashboards do not include the complete set of metrics that are tracked by the Data Hub. For example, the development of new or secondary malignancies is not included in the current dashboard versions but is an important long-term clinical outcome that the Data Hub will track.

Data flow, interoperability, and quality

The Data Hub database is primarily populated with patient EHR data submitted by sites. Centers using any type of EHR (Epic, Cerner, or others) are able to participate. In the future, EHR-agnostic API-based data transmission solutions from third-party apps may also be considered. The ASH Research Collaborative has partnered with Prometheus Research (PR), an IQVIA company, to aggregate and curate data from multiple sources such as EHRs, and, in the future, patient-reported outcomes, PGHD, genomics, and other data sources. The Data Hub data curation method ingests the totality of EHR data from sites using the OMOP CDM or FHIR data exchange format. This approach to data acquisition is intended to reduce the burden for participating Data Hubsites and may better support long-term sustainability by reducing the resources required for redundant data capture. As previously noted, a web-based data entry tool is available to facilitate additional or alternative data entry as needed. This tool also allows sites to validate e-phenotypes and override inaccurate data. Generally, initial data submission involves support from the institutional information technology team, though this support is not expected to be significant over the long run. Some sites can expedite data submission when IT personnel have participated in similar projects, and others have benefited from modest ASH Research Collaborative grant support to help prioritize data integration work needed to submit data.

Subsequent manual data curation at the site level is facilitated by clinical teams who are incentivized to participate based on the particular disease and project for which data are being submitted.

Regardless of the data ingestion method, data captured are longitudinal and comprehensive with detailed information on patient demographics, comorbidities, medications, treatments,

Outcomes			Prevention	Disease-modifying therapies	Kidney disease	
Inpatient stay	ED treat & release	Acute care clinic	Penicillin V	Hydroxyurea	Albuminuria	Transfusion
Ischemic stroke	Pulmonary embolus	Acute chest syndrome	Screening	Crizanlizumab	Chronic kidney disease	Iron overload
Deep venous thrombosis	Hemorrhagic stroke	Silent cerebral infarcts	Transcranial doppler	L-glutamine	Hypertension	Chelation
Priapism	Vaso-occlusive episode	Acute kidney injury	Urine albumin test	Voxelotor	ACE inhibitors & ARBs	Prescribed opioids

Figure 2. SCD site dashboard metrics.

Figure 3. **MM site dashboard metrics.**

Birth sex	Participant accrual	Therapy classes & agents
Age at first active multiple myeloma diagnosis	Chromosomal abnormalities	Active multiple myeloma type & isotype
Stem cell transplant types	Time to treatment failure	Overall survival

utilization, processes, and outcomes, as covered through the United States Core Data for Interoperability (USCDI).[13] See Figure 4 for a list of USCDI v2 data categories and data descriptions. As new versions of the USCDI are created and released, the Data Hubwill correspondingly update its data capture methods to improve the comprehensiveness of the data ingested. The Data Hub will also ingest clinical notes, pathology and imaging reports, and other types of documentation that may contain "unstructured" data. Future approaches to handling unstructured data will include natural language processing (NLP) software to "read" and glean information from the unstructured notes, as well as efforts to bring additional structure into these data sources through advocacy and consensus-building efforts. Regardless of the source, in totality, the data inform research analyses and clinical dashboards and can be queried and exported. Data Hub sites submit data at least quarterly (most select monthly), and data are curated and collated within the corresponding disease-specific data model. When sites submit data, existing records are refreshed. Refreshes do not overwrite information in the eCRFs. Because all potential future data needs and analyses are not known, the entirety of the EHR data for each patient continues to be stored within the Data Hub. The Data Hub has developed a data quality maturity model to ensure data are adaptable, reusable, and scalable for performance improvement initiatives. Once a site submits data, a data quality report is generated.

Interoperability

Through harmonizing approaches to achieve comparable data at the site level, the Data Hub is positioning its data to be used for ML and AI.[14] Advances in digital medicine have not translated easily into implementation mostly due to the lack of standardized data across health systems. The ASH Research Collaborative is working to create widely adopted e-phenotypes that translate EHR codes, such as RxNorm, SNOMED, LOINC, and ICD-10, into clinically relevant data variables, attempting to reduce manual data curation where possible. As data are submitted to the Data Hub, data quality is addressed through data standardization methods, data quality reports, and local data validation procedures. Truly interoperable data can be used for other purposes, including the development of predictive models to inform clinical decision support at point of care for participating sites. The ASH Research Collaborative will align where possible with other efforts such as mCODE (minimal Common Oncology Data Elements) that have overlapping goals.[15]

Longitudinal follow-up

Longitudinal follow-up of patients is a priority, and traditional registries have used follow-up procedures that require direct patient interaction and manual data entry. New approaches to longitudinal follow-up using EHR and claims data will benefit the Data Hub. Linked medical claims data have been used as outcomes data in some studies.[16] Currently, 17 states have all-payer claims

Allergies and intolerances	Assessment and plan of treatment	Care team member(s)	Clinical notes	Clinical tests	Diagnostic imaging	Encounter information
Goals	Health concerns	Immunizations	Laboratory	Medications	Patient demographics	Problems
	Procedures	Provenance	Smoking status	Unique device identifier(s) for a patient's implantable device(s)	Vital signs	

Figure 4. **USCDI v2 data categories for data submitted to the Data Hub.**

databases (including New York and California), and national legislation is pending to create a national all-payer claims database.[17,18] The prospect of a national all claims database may enhance the ASH Research Collaborative Data Hub and the creation of a hematology coordinated registry network similar to those developed by other specialty societies.[19] Expansion of the Data Hub to accommodate a broader network of sites, community participation, and patient-generated data will also help to address the longitudinal loss to follow-up issue.

Patient-generated data and hybrid studies

To facilitate patient-centered research and care, the information provided directly by or obtained from patients is important and difficult to standardize in routine practice. Mobile apps, sensors, and the addition of data collected in usual care can be linked to rosters of patients followed in RWD/RWE studies. An electronic informed patient consent module is available to allow the collection of these types of data and allow patients to be recontacted to inform longitudinal outcome evaluations. Further, RWD/RWE platforms will be leveraged to facilitate more efficient prospective randomized study designs through so-called "hybrid" studies. Prospective hybrid studies can address several important challenges, such as difficulty accounting for known and unknown prognostic factors and differences in endpoint definitions between trial and RWD data. In hybrid trials, patients can be randomized to balance prognostic factors, and endpoint definitions (eg, rwPFS) will be the same for both arms. These trials are especially suited for approved drugs and postmarketing research investigating comparative effectiveness, sequencing of agents, comparative tolerability, and other important objectives. These trials can also be used as part of pragmatic trials and other studies of health care delivery interventions within learning networks.[20] With these approaches, careful analytics and data quality assessment are needed to ensure fitness for purpose.

Using the Data Hub to generate RWE

The ASH Research Collaborative is expanding its stakeholder community to facilitate RWE generation, with an initial focus on SCD genomic therapy research that will require longitudinal evidence generation. The ASH Research Collaborative and the Innovative Genomics Institute (IGI), in collaboration with the FDA, have engaged people living with SCD, clinicians, researchers, industry, and regulators to explore methods to support SCD RWE generation using the Data Hub program.

The initiative's stakeholder participants are working to recommend data to collect and methods to coordinate clinically relevant and reliable RWD. Stakeholders have convened through roundtable meetings and working groups. The first stakeholder roundtable was held in March 2021 to discuss the role of RWE for FDA regulated studies, examples of how coordinated registry networks (CRN) have facilitated the use of RWD for improved safety, efficacy, and label expansion, and the urgency of harmonizing data collection in new genomic therapies. A CRN working group is exploring how the Data Hub could serve as a CRN to provide RWD for a variety of regulatory purposes and linkage to other data sources (claims, EHR, and data collected via apps or remote monitoring) that potentially increase utility, reduce costs, and better reflect patients' experiences compared with traditional methods. The Genomic Therapies Work Group is seeking consensus on data points that should be collected and procedures and assays to be used to generate actionable, regulatory-grade RWE for genomic therapies for these blood disorders. A final report will address recommendations for the collection, curation, storage, and sharing of data collected in clinical settings (therapeutic and research) that can also provide reliable, fit-for-purpose RWE to regulators, health care providers and payers, investigators, and patients about the safety and effectiveness of genome editing and other novel therapies for SCD and other hematologic conditions.

Using the Data Hub to improve clinical care

The Data Hub facilitates the exchange of information through real-time dashboards, queries, and research. Data are also used to highlight gaps in clinical care and patient outcomes to facilitate quality improvement. HHS and the Office of Minority Health (OMH) awarded ASH, the ASH Research Collaborative, and the Learning Networks Program at the James M. Anderson Center for Health Systems Excellence at Cincinnati Children's (Anderson Center) a grant to build an SCD clinical data platform and an SCD LC. Anderson Center has extensive experience developing learning networks using a learning network model, which aligns with the National Academy of Medicine framework of a learning healthcare system.[21] The SCD clinical data platform will leverage the Data Hub's longitudinal patient data to track practice patterns, and the LC will focus on actionable and measurable areas to improve, implementation of evidence-based strategies to support selected areas for improvement, and measurement of change using Data Hub data.

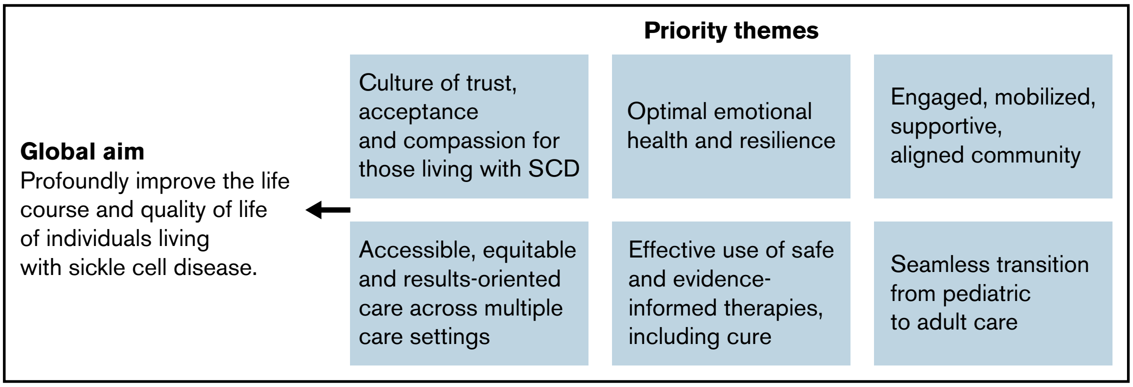
LC stakeholders include clinical teams, patient and family members, implementation scientists, quality improvement experts, and psychologists. Stakeholders are designing a pilot implementation that will include 15 to 20 sites participating in the Data Hub's SCD program. The global aim and priority themes are listed in Figure 5.

Sites will engage with each other to share lessons learned and strategies for success. The pilot LC will inform the launch of a nationwide SCD LC to improve outcomes at a national level. The nationwide LC will be available to all sites enrolled in the SCD Data Hub program. The ASH Research Collaborative is also exploring ways in which lessons from the SCD LC can be applied to develop LC activities in other disease programs supported by the Data Hub.

SCD CTN

The ASH Research Collaborative SCD CTN was launched to improve outcomes for individuals with SCD by expediting the development of treatments and facilitating innovation in clinical trial research. Since the Data Hub provides a valuable and essential resource for the operation of the SCD CTN, all sites in the network will contribute data to the Data Hub. The Data Hub will be used to identify cohorts for trials and provide RWE control arms, among other uses relevant to the SCD CTN.

The SCD CTN provides 3 resources that will be of benefit to investigators, sponsors, and ultimately patients. First, it provides a collection of advisory boards that will help overcome barriers to clinical trial participation, prioritize research areas of interest to the SCD community, improve enrollment, design, and execution of clinical trials, and local community advisory boards at each of the sites to supplement the opinions that are provided by national patient advocates. Second, it is connected to the Data Hub, which provides a centralized data repository, will identify cohorts for research



Global aim
Profoundly improve the life course and quality of life of individuals living with sickle cell disease.

Priority themes

- Culture of trust, acceptance and compassion for those living with SCD
- Optimal emotional health and resilience
- Engaged, mobilized, supportive, aligned community
- Accessible, equitable and results-oriented care across multiple care settings
- Effective use of safe and evidence-informed therapies, including cure
- Seamless transition from pediatric to adult care

Figure 5. SCD LC global aims and priority themes.

(ie, well-characterized patients for inclusion and exclusion criteria), natural history studies, and a contemporaneous control group. Third, the SCD CTN provides well-vetted and engaged clinical trial sites with a culture of collaboration and research, an efficient, coordinated approach to clinical trials research, and centralized IRB & contracting.

The SCD CTN and its partners share a commitment to: (1) forge new relationships with the SCD community to increase their understanding of clinical trials and trust in SCD researchers; (2) eliminate inefficiencies through the use of a centralized data repository (the ASH Research Collaborative Data Hub), a single institutional review board approval, and centralized contracting; and (3) focus on the research opportunities that hold the most promise for individuals living with SCD.

While there are currently only 4 FDA-approved drugs to treat SCD, there is now a robust SCD drug development pipeline poised to drive demand for SCD clinical trials, providing a prime opportunity to advance treatment and care of those affected by SCD.

The SCD CTN has been designed to address the many issues limiting evidence generation and, importantly, to incorporate the voice of patients.[22] It is important to note that the SCD CTN functions primarily as a clinical trials accelerator. The SCD CTN is not a contract research organization (CRO) and thus does not actually conduct the trials itself; instead, it brings sites, investigators, and patients together through shared purpose with a common infrastructure. When fully enrolled, the SCD CTN and Data Hub will encompass approximately half of all patients with SCD in the United States, thereby representing an unparalleled opportunity to benefit this population through accelerated research and improved clinical practice.

The SCD CTN is designed to substantially shorten time-to-trial launch and completion, in part by engaging sites quickly and reaching patients expeditiously, providing value to industry, researchers, clinicians, patients, and other stakeholders. At the same time, the SCD CTN serves as a steward of a culture that is patient-informed, patient-centered, and determined to address and ameliorate disparities in access to potential life-extending therapies.

Responding to the pandemic: the ASH Research Collaborative COVID-19 registry for hematology

The ASH Research Collaborative was well-positioned to accommodate unanticipated and urgent data needs relevant to hematology, including the global COVID-19 pandemic. Early in the pandemic, clinicians were concerned that patients with underlying hematologic conditions could be at risk for adverse outcomes from COVID-19.[23] As the pandemic evolved, there was also increasing recognition that hematologic complications of COVID-19 infection, such as thrombosis, were also prevalent. The ASH Research Collaborative launched the COVID-19 registry for hematology in April of 2020 as a global public reference tool. The registry captures RWD on individuals who test positive for COVID-19 and have a hematologic condition (past or present) and/or have experienced a post-COVID-19 hematologic complication. Initial results from the registry have been recently presented and published.[24]

As data are received and analyzed, real-time observational summaries are made available via a publicly displayed dashboard intended to support clinical decision-making. The registry has been designed as a provider-entered, case report form-based, voluntary submission program. There are efforts underway in which the ASH Research Collaborative is participating to increase COVID-19 testing and promote laboratory standardization, and the ASH Research Collaborative anticipates collecting COVID-19 data from EHR-integrated data capture in the future as part of the SCD and MM programs to gain a better understanding of any potential long-term sequelae or interactions with diseases or treatments.

Accessing the Data Hub: the value of the Data Hub to the hematology community

The examples provided above demonstrate many ways the Data Hub will be used to accelerate research and improve practice and patient outcomes in hematology. Ultimately, the value of the resource depends on the degree to which it is used by providers who input data as well as trained users who use the data for research, analysis, and improvement of clinical practice. Many of the data tools being developed by the ASH Research Collaborative will be available to all participating sites, including site dashboards that incorporate metrics from resources such as ASH evidence-based clinical guidelines, the ability to analyze site-level data and visualize comparisons with aggregate data, and in the future, point of care clinical decision support. Individual researchers also access Data Hub data to facilitate specific analyses, grant submissions, and other objectives. The Data Hub oversight group has developed data access and use procedures which will be made widely available to the hematology community, along with training and ongoing support to ensure high-quality analyses when the Data Hub has accrued sufficient data and is ready to be made available for scientific analyses. In the future, insights and developments in the field based on Data Hub activities will also be communicated broadly with the hematology community through newsletters and other means. The Data Hub will also work to be as useful as possible to the patient community. Future features of a patient-facing portal, in addition to direct data entry from patients, could include the ability for patients to access treatment guidelines, opportunities to participate in clinical studies, and the ability to access information related to centers specializing in the care of individuals like them. Ongoing input from patients and external stakeholders will be actively sought to ensure that the Data Hub provides maximum value to all those who might wish to interact with it.

Future of the ASH Research Collaborative

Over the next several years, the Data Hub will capture longitudinal data on many individuals in the United States living with SCD or MM. Efforts will soon expand to encompass additional hematologic conditions. The ASH Research Collaborative SCD CTN will continue to expand and engage clinicians and patients across the country as new sites are onboarded. The Data Hub will be used for hypothesis-generating research and RWD analytics to inform drug development. Data quality is a prominent priority, and it is anticipated that Data Hub data will be used to create new point-of-care clinical decision support tools. Throughout these activities, the ASH Research Collaborative will continue its process of multistakeholder engagement to ensure that the resulting trials and data are fit for purpose for various entities throughout the regulatory, clinical, research, and health care policy ecosystem.

Authorship

Contribution: W.A.W. and G.P. developed the initial draft of the manuscript. All authors provided substantive input, edited, reviewed, and approved the manuscript.

Conflict-of-interest disclosure: W.A.W. and S.W. have received honoraria or fees from the ASH Research Collaborative. R.M.P. and M.G. are employees of the American Society of Hematology. K.H. and E.A.T. are employees of the ASH Research Collaborative. B.K.D. was an employee of the ASH Research Collaborative. W.A.W., D.S.N., C.S.A., A.A.T., and K.C.A. have leadership positions on oversight groups or committees within the ASH Research Collaborative.

Disclaimer: This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

ORCID profiles: A.A.T., 0000-0003-4961-8103; D.R., 0000-0002-4565-4556.

Correspondence: William Wood, Division of Hematology, Department of Medicine, University of North Carolina at Chapel Hill Houpt Physician Office Building, Chapel Hill; email: wawood@med.unc.edu

References

1. Califf RM, Robb MA, Bindman AB, et al. Transforming evidence generation to support health and health care decisions. *N Engl J Med.* 2016;375(24):2395-2400.
2. United States Food and Drug Administration. Real World Evidence. Available at: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed 18 November 2021.
3. National Institutes of Health. All of Us Research Program. Available at: https://allofus.nih.gov. Accessed 18 November 2021.
4. Telen MJ. Beyond hydroxyurea: new and old drugs in the pipeline for sickle cell disease. *Blood.* 2016;127(7):810-819.
5. Palumbo A, Avet-Loiseau H, Oliva S, et al. Revised international staging system for multiple myeloma: a report from International Myeloma Working Group. *J Clin Oncol.* 2015;33(26):2863-2869.
6. Kumar S, Paiva B, Anderson KC, et al. International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma. *Lancet Oncol.* 2016;17(8):e328-e346.
7. Joseph NS, Kaufman JL, Dhodapkar MV, et al. Long-term follow-up results of lenalidomide, bortezomib, and dexamethasone induction therapy and risk-adapted maintenance approach in newly diagnosed multiple myeloma. *J Clin Oncol.* 2020;38(17):1928-1937.
8. Marinac CR, Ghobrial IM, Birmann BM, Soiffer J, Rebbeck TR. Dissecting racial disparities in multiple myeloma. *Blood Cancer J.* 2020;10(2):19.
9. Fiala MA, Wildes TM. Racial disparities in treatment use for multiple myeloma. *Cancer.* 2017;123(9):1590-1596.
10. Center for Medical Technology Policy (CMTP), Green Park Collaborative. Core Outcomes in Sickle Cell Disease. Available at: http://www.cmtpnet.org/green-park-collaborative/core-outcome-set-initiatives/corescd/. Accessed 18 November 2021.
11. Bodenreider O, Nguyen D, Chiang P, et al. The NLM value set authority center. *Stud Health Technol Inform.* 2013;192(1):1224.
12. Phenotype Knowledge Base (PheKB). A knowledge base for discovering phenotypes from electronic medical records. Available at: https://phekb.org/. Accessed 18 November 2021.
13. HealthIT.gov. United States Core Data for Interoperability (USCDI). Available at: https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi. Accessed 18 November 2021.
14. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. *NPJ Digit Med.* 2019;2(1):79.
15. Osterman TJ, Terry M, Miller RS. Improving cancer data interoperability: the promise of the Minimal Common Oncology Data Elements (mCODE) initiative. *JCO Clin Cancer Inform.* 2020;4(4):993-1001.
16. Krucoff MW, Sedrakyan A, Normand S-LT. Bridging unmet medical device ecosystem needs with strategically coordinated registries networks. *JAMA.* 2015;314(16):1691-1692.

17. Agency for Healthcare Research and Quality. Inventory and prioritization of measures to support the growing effort in transparency using all-payer claims databases. Available at: https://www.ahrq.gov/data/apcd/backgroundrpt/intro.html#uses. Accessed 18 November 2021.

18. The Commonwealth Fund. What can be done to improve all-payer claims databases? Available at: https://www.commonwealthfund.org/blog/2020/what-can-be-done-improve-all-payer-claims-databases. Accessed 18 November 2021.

19. Peters A, Sachs J, Porter J, Love D, Costello A. The value of all-payer claims databases to states. *N C Med J.* 2014;75(3):211-213.

20. Mack C, Christian J, Brinkley E, Warren EJ, Hall M, Dreyer N. When context is hard to come by: external comparators and how to use them. *Ther Innov Regul Sci.* 2020;54(4):932-938.

21. Olsen LA, Aisner D, McGinnis JM, eds. Institute of Medicine Roundtable on Evidence-Based Medicine. *The Learning Healthcare System: Workshop Summary.* Washington, DC: National Academies Press; 2007.

22. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA.* 2018; 320(9):867-868.

23. Cook G, John Ashcroft A, Pratt G, et al; United Kingdom Myeloma Forum. Real-world assessment of the clinical impact of symptomatic infection with severe acute respiratory syndrome coronavirus (COVID-19 disease) in patients with multiple myeloma receiving systemic anti-cancer therapy. *Br J Haematol.* 2020;190(2):e83-e86.

24. Wood WA, Neuberg DS, Thompson JC, et al. Outcomes of patients with hematologic malignancies and COVID-19: a report from the ASH Research Collaborative Data Hub. *Blood Adv.* 2020;4(23):5966-5975.