RWD27

Creation of a Data Quality Framework for a United States Electronic Medical Record-based Registry for Individuals with Spinal Muscular Atrophy

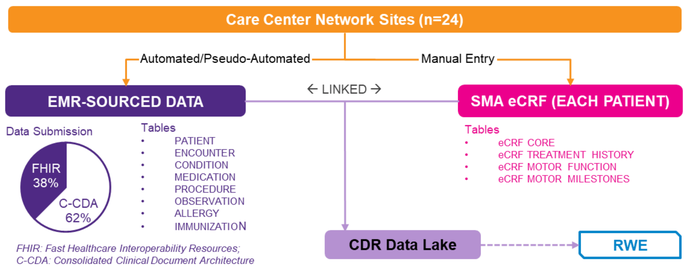
Sarah M Whitmire MS[1], Erin F Welsh MPH[1], Lisa Belter MPH[1], Ashwin Rai MS[2], Jolene Damon, MPH[2], Ariel Berger MPH[2], Mary Curry ND[1], and Mary Schroth MD, FAAP, FCCP[1]

[1] Cure SMA, Elk Grove Village, IL, US; [2] Evidera, Inc., Bethesda, MD, US

Background

- Spinal Muscular Atrophy (SMA) is a rare genetic neurodegenerative disease with a birth prevalence of approximately 1 in 15,000 individuals[1] in the United States (U.S.).
- Real-world data (RWD) can be used to understand the current disease landscape and help evaluate relevant outcomes, resulting in the generation of real-world evidence (RWE) that can improve standards of care (SOC).
- Cure SMA, a patient advocacy organization that provides support for SMA research and care, created the SMA Clinical Data Registry (CDR) in 2018 to inform development of evidence-based SOC.
  - The CDR is comprised of data from >1,150 consented individuals with SMA and is among the largest electronic medical record (EMR)-based SMA registries worldwide.
  - The CDR ingests monthly submissions of structured EMR data from 24 participating U.S. clinical care sites (Cure SMA Care Center Network) into a shared data mart.
  - High-priority SMA data unavailable from discrete EMR fields are collected via linked electronic case report forms (eCRFs) completed annually.

Figure 1: Cure SMA CDR Data Sources



- Ensuring high-quality data across participating institutions is challenging, given heterogeneity in data availability, reliability, infrastructure, submission methods, and extraction/mapping.
- Cure SMA developed a customized CDR quality framework to ensure data quality to inform SOC development and enable additional use cases.

Methods

- The CDR quality framework follows “ALCOA+” (Attributable, Legible, Contemporaneous, Original, and Accurate) principles, from which foundational checks that focus on data conformance, plausibility, and completeness at multiple levels (e.g., record, patient, care center site) were developed.
- Data quality checks from existing published frameworks[2-4] were used to inform those developed for the SMA CDR and were supplemented with customized checks based on past CDR quality issues.
- Table-driven macros were created using RStudio® (Posit, Boston, MA) to perform data quality algorithms through direct connection to the data lake.
- The process runs monthly to assess the quality of current extracts, and performs threshold driven comparisons to existing data for a subset of checks.

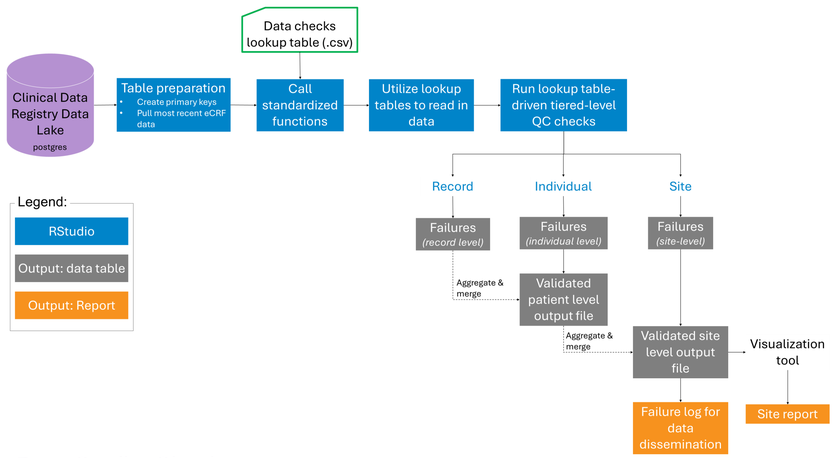
Acknowledgements

- Thank you to the SMA community for sharing their data and supporting the Cure SMA Care Center Network.
- Thank you to the Cure SMA Care Center Network for their commitment to improving care for people with SMA and contributing consented patient data to the SMA Clinical Data Registry. The Cure SMA Care Center Network includes 24 integrated SMA Care Centers across the US who provide multidisciplinary care for people with SMA.
- Funding was provided by the Cure SMA Real World Evidence Collaboration, which includes Novartis Gene Therapies, Biogen, and Genentech/Roche.

The Cure SMA Real World Evidence Collaboration was established in 2021 to leverage the experience, expertise and resources of pharmaceutical and biotechnology companies and nonprofit organizations involved in development of SMA therapeutics to guide the future direction of real world data collection and use in SMA.

Process

Figure 2: Cure SMA CDR Data Quality Check Process



Data Quality Checks

Table 1: Quality Checks Included in the First Phase of the Cure SMA CDR Framework

Conformance Checks	Level
• Refresh dates must be documented	Record
• All dates must be in YYYY-MM-DD format	Record
• Orphan encounter IDs are not present in **CONDITION**, **MEDICATION**, **PROCEDURE**, **OBSERVATION** tables	Individual
• **eCRF MOTOR FUNCTION** assessment score records are within acceptable assessment score ranges	Record

Completeness Checks	Level
• Core tables are provided	Site
• Core variables are present	Site
• Percent of null values in core fields is below a pre-specified threshold for a site	Site
• The percentages of records in **ENCOUNTERS**, **CONDITIONS**, **MEDICATIONS**, **PROCEDURES**, or **OBSERVATIONS** that are missing critical data are below a pre-specified threshold	Site
• Quantitative laboratory **OBSERVATION** records or vital signs specify unit of measure	Record
• Site has had an **EMR** data submission in the last month	Site
• Percent/absolute difference between the number of records with missing data in each table between two consecutive monthly data cycles is ≥ 0 (all records, not incremental)	Variable
• The number of records in any table between two consecutive monthly data cycles is ≥ 0 (all records, not incremental)	Variable

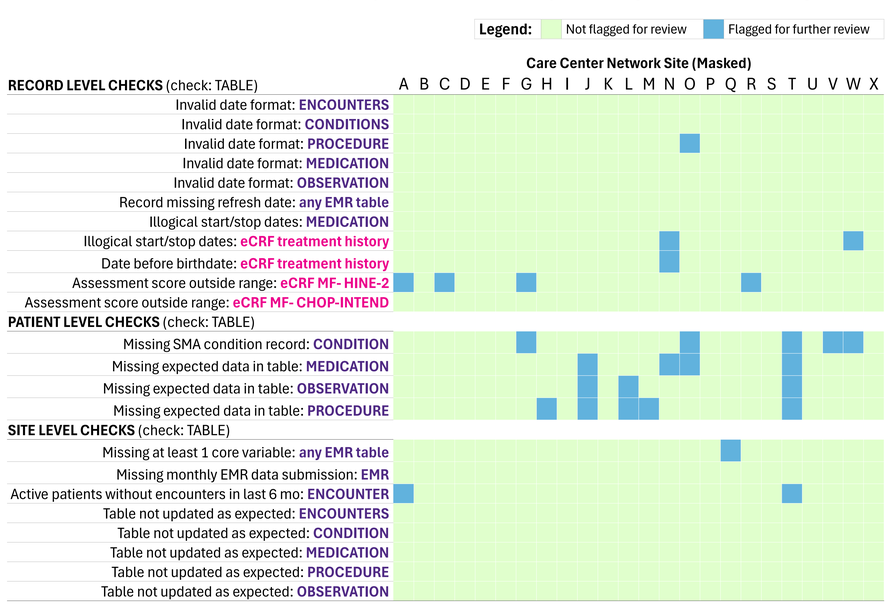
Plausibility Checks	Level
• All individuals have a diagnosis of SMA in the **CONDITION** table	Individual
• The percentage of individuals with at least one **ENCOUNTER** record that do not have any records in other tables is below a pre-specific threshold	Site
• The percentage of active individuals (based on status recorded in **eCRF CORE**) that do not have any **ENCOUNTER** records in the last 180 days at a site is below a pre-specific threshold	Site
• The values recorded for height, weight, diastolic blood pressure, or systolic blood pressure are positive in **OBSERVATIONS**	Record
• SMA treatments recorded in **eCRF TREATMENT HISTORY** are found in the EMR **MEDICATION** table	Individual
• Country of residence in **eCRF CORE** lines up with country of residence recorded in the **PATIENT** table	Individual
• Stop date is not before start date (**CONDITION**, **MEDICATION**, **ALLERGY**, **eCRF TREATMENT HISTORY**)	Record
• Records cannot have a date in the future	Record
• Records for dates that should occur after birth cannot have a date prior to birthdate	Record
• Dates recorded in **eCRF TREATMENT HISTORY** and **eCRF MOTOR FUNCTION** cannot occur after death date	Record
• Dates reported on **EMR** records cannot occur after death date + 7 days (based on data distribution)	Record
• There is not a lag in **EMR** data submitted for any table (lag defined as no new data submitted for any patient at a site in the last 60 days)	Site
• The number of duplicates records in **PATIENT**, **ENCOUNTER**, **CONDITION**, **PROCEDURE**, **OBSERVATION**, **ALLERGY**, and **IMMUNIZATION** must be below a pre-specified threshold	Site
• Percent difference between the number of pure duplicate records between two consecutive monthly data cycles must be less than a pre-defined threshold (all records, not incremental)	Variable

Legend: **EMR table**; **eCRF table**

CDR Preliminary Results

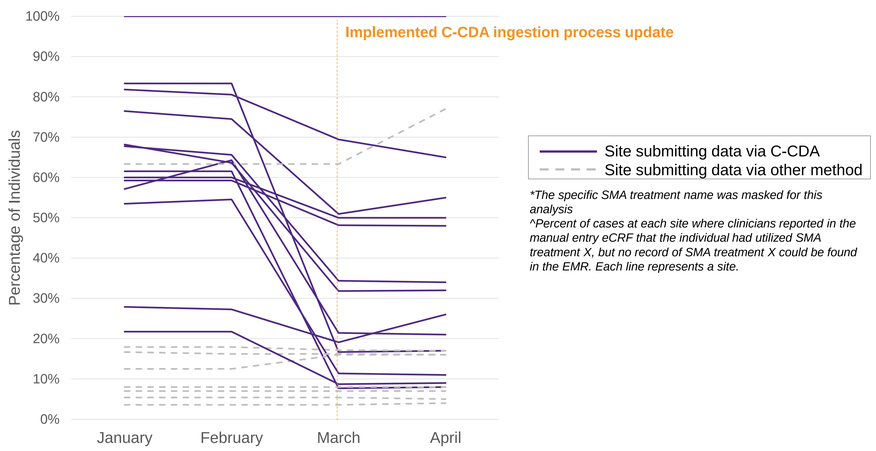
- The CDR was assessed at 4 timepoints: January 2024, February 2024, March 2024, and April 2024
  - A subset of the most recent results from April 2024 are presented in **Table 2**

Table 2: Site-level Results of Data Quality Process Run in April 2024 (Subset Of Checks)



- While results evaluating the current state are valuable, visualizing change over time is necessary for tracking progress or quickly identifying inconsistencies.
  - The importance of site-level temporal review is illustrated in **Figure 3**, which sets forth an example of a C-CDA ingestion process that was updated in March 2024.

Figure 3. Change Over Time: Percentage of Individuals at Each Site where SMA Treatment X* was Reported in the eCRF and Records were Missing In EMR^



**The specific SMA treatment name was masked for this analysis*
^Percent of cases at each site where clinicians reported in the manual entry eCRF that the individual had utilized SMA treatment X, but no record of SMA treatment X could be found in the EMR. Each line represents a site.

Conclusions and Next Steps

- Processes that ensure high confidence in RWD are critical, given increases in reliance on RWE to inform decisions related to medical care and outcomes.
- The CDR quality framework will ensure that relevant and extensive RWD of high quality and completeness is available to create RWE to inform future SMA best care guidelines and other use cases.
- Once the data quality process is finalized, the next step is to create a sustainable process to share relevant feedback/reports with our data platform vendor and Clinical Care Network sites.

References
1. Data on file
2. Qualls LG et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *EGEMS (Wash DC)*. 2018; 6(1): 3.
3. Khare R et al. A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc*. 2017 Nov 1;24(6):1072-1079
4. Kahn MG et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)*. 2016; 4(1): 1244.