Cerebral Palsy Registry Network

Data Coordinating Center

at the

Division of Health System Innovation & Research

Department of Population Health Sciences

University of Utah, School of Medicine

Vikrant G. Deshmukh, Ph.D.

Paul H. Gross

Jacob Kean, Ph.D.

Susan D. Horn, Ph.D.

Rachel Hess, M.D.

March 12, 2016

Version 0.3

Table of Contents

Executive Summary

The Cerebral Palsy Research Network (CPRN) aims to improve treatments and outcomes for persons with cerebral palsy through high quality clinical research and quality initiatives. Central to the CPRN is a CP registry that will be used to collect longitudinal treatment and outcomes data from multiple participating sites in the CPRN. The CPRN registry will be hosted at the University of Utah, Dept. of Population Health Sciences, a.k.a., the Data Coordinating Center (DCC). This document outlines the key elements needed to develop the CPRN registry.

The CP registry will be implemented in phases, and Phase I has started with the development of CPRN's Common Data Model (CPRN CDM), a collaborative effort to standardize the dictionary used for data collection. Phase II will involve lead sites with Epic & Cerner Electronic Health Record (EHR) systems developing a series of structured clinical documentation (SCD) forms in their respective systems to collect identified data at the point of care on CP patients, i.e., Electronic Data Capture (EDC). Phase II will also involve sites that do not use EDC to collect data using the Research Electronic Data Capture (REDCap) software hosted by the DCC. Phase III will involve the aggregation of data across different sites that have implemented EDC in Epic, Cerner, and other EHR systems, as well as the DCC-hosted REDCap sites by relying on the Clinical Data Interchange Standards Consortium (CDISC) standards including the Operational Data Model (ODM), XML-based export formats. Phase IV will involve the distribution of these aggregated data to PIs at various sites under a collaborative governance structure, which is currently being finalized.

Technical Overview

Architectural Considerations

1. <u>The Data Coordinating Center (DCC).</u>

The DCC will serve as the main site for collection and aggregation of data from all the participating sites in the CPRN (Figure 1). The DCC infrastructure will include: (i) a Secure File Transfer Protocol (SFTP) site for receiving extracts from participating sites; (ii) an Extract-Transform-Load (ETL) environment for processing extracts received from the participating sites; (iii) a Database (DB) environment for hosting the aggregated extracts from various sites; and (iv) a frontend web portal to serve the data needs of the CPRN community.

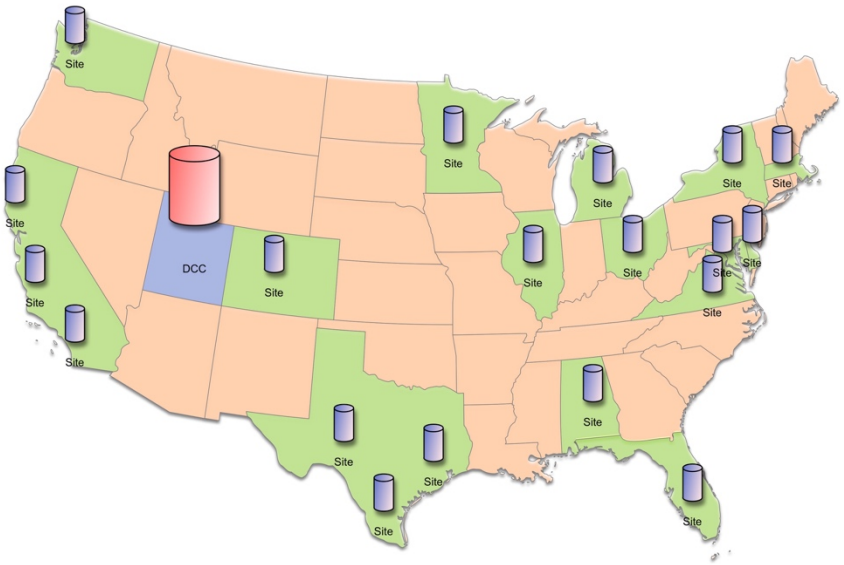

Figure 1: Member CPRN sites (green) and the DCC (blue) in the United States and other places like Canada (not shown) will contribute data to the DCC using automated or manual methods.

The SFTP site will run on high availability (HA) servers to receive datasets from participating sites (Figure 2). Two transfer mechanisms will be supported on the SFTP site: automated & manual. An automated SFTP option (1,2) will allow participating sites with Electronic Data Capture (EDC) to rely on their own ETL mechanisms to upload ODM XML to the DCC. The ETL

option has inherent advantages in streamlining data collection from sites with EHRs, and will reduce manual overhead across the CPRN. A manual option will be provided on the frontend web portal for sites where automated ETL options are unavailable. The manual upload option is necessary but less preferable than the automated option, and will require participating sites to designate a person to upload the extracts to the frontend portal. In the future, if the data needs of the CP registry grow to a point where transfers of large data-sets become necessary, the DCC will consider alternative solutions like Globus File Transfer.[1]

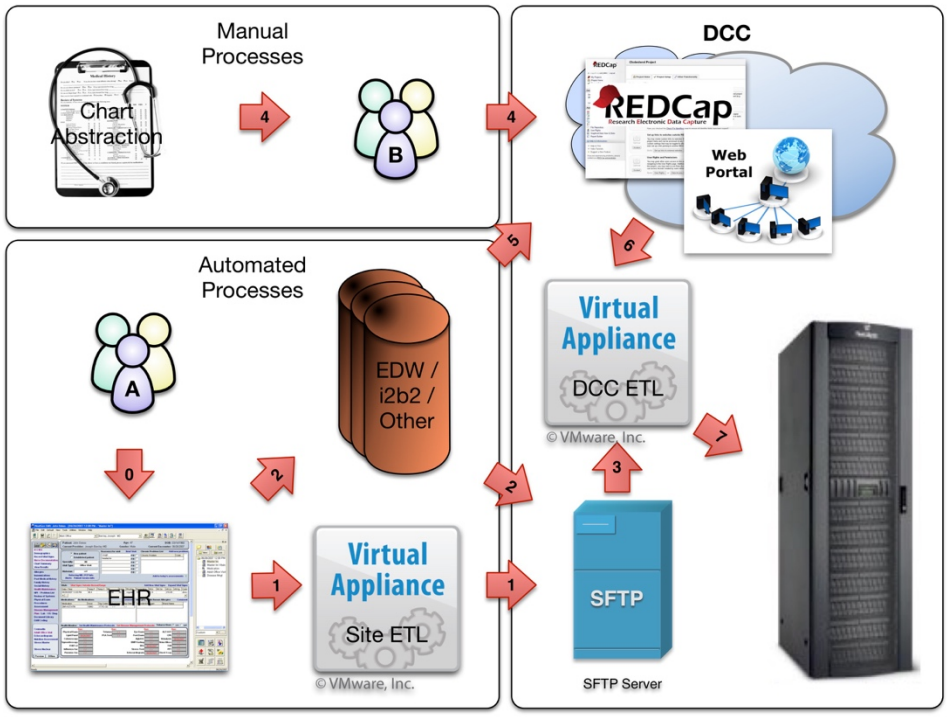

Figure 2: A top level view of data transfer from CPRN sites to the DCC. Users with automated processes (A) will chart at the point of care using EHR (1, 2), while manual chart abstracters (B) will use REDCap (4). Automated sites can either transfer data to the SFTP Server using ETL Virtual Appliance (1) or other processes (2). Some sites may also upload data to the Web Portal hosted at the DCC (5). DCC will load data from SFTP Server (3), REDCap & Web Portal (5) servers by relying on a separate ETL processes running on a Virtual Appliance hosted at the DCC (6).

[1] For an overview of Globus Online data transfer, see https://www.globus.org/how-it-works . This approach has already been tested at the University of Utah's Center for High Performance Computing.

The DCC ETL environment will consist of a Talend Open Studio for Data Integration, a high performance, free and open-source ETL tool.[2] The DCC will develop ETL processes in Talend to process the extracts uploaded by each of the participating sites to the DCC SFTP site via either method described above (Figure 3). Upon successful processing of the extracts, the ETL process will record the extraction statistics, and move the data files over to an archive location. Upon unsuccessful processing, if the cause of failure is identified as either the format or content of the extract files, the ETL process will notify the respective sites as well as the DCC team. If the failure is unrelated to the extracts themselves, the ETL process will only notify the DCC staff. Processing of extracts received from various sites will occur in a first-in-first-out (FIFO) approach, with batches processed during the day. The DCC staff will investigate all failures, collaborate with, and advise participating sites on their findings to formulate appropriate solutions.

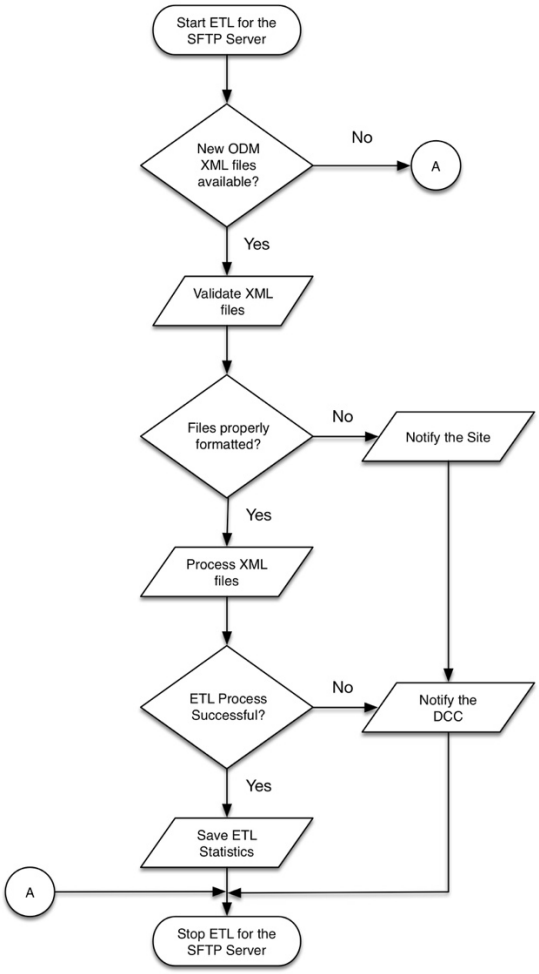

Figure 3: ETL Process at the DCC for ODM XML files received via SFTP, notification mechanisms, etc.

[2] Talend Open Studio for Data Integration is a high performance, free, open-source ETL tool. More information here: https://www.talend.com/download/talend-open-studio#t4

The DB will consist of an Oracle Real Application Cluster (RAC) environment with multiple redundant nodes on Oracle Virtual Machines (OVM) for performance and scalability, which is the standard at the University of Utah Health Care environment.[3] The DB will be organized into a Staging Area and Data-Marts. Primary ETL processes will load the raw extracts from each site into the Staging Area, and secondary ETL processes will then transform and load these data into Data Marts (DM), which are logical, subject-oriented areas like procedures, medications, etc. We will adopt the Observational Medical Outcomes Partnership (OMOP) Common Data Model (OMOPCDM) format for the primary DMs to facilitate representation of longitudinal data and downstream analysis.[4] Additionally, OMOP may be developing self-service query tools which use the OMOPCDM to provide end-users the ability to easily query these data. We will evaluate the OMOP self-service tools when they become available, as well as the Informatics for Integrating Biology and the Bedside (i2b2) self-service query tool[5] and pick the most suitable tool to provide participating sites the ability to run self-service queries from within the frontend web portal, also hosted at the DCC (Figure 2). Data loaded in the DM will be a limited data set, whereas data loaded into the i2b2 instance Data loaded into the DM should be identifiable. Data loaded into i2b2 will be will be de-identified to minimize risk of unwanted disclosure of Protected Health Information (PHI), and access to these data will be granted in accordance with

[3] Oracle RAC provides clustering and HA for performance, scalability and resilience, by distributing the workload among several nodes that tap into a common database in a share-everything configuration. See http://www.oracle.com/us/products/database/options/real-application-clusters/overview/index.html for more information.

[4] For more information on the OMOP CDM, please see http://omop.org/CDM . The purpose of the CDM is to standardize the format and content of observational data to enable standard analysis and frontend tools, applications, etc. to be applied to them

[5] The i2b2 platform enables integration of data from disparate sources into an i2b2 data model, and enables self-service queries via a web client. For more information, see: https://www.i2b2.org/software/index.html

the governance policies that are currently under development by the corresponding workgroup. The de-identification strategy will be reused for providing data to PIs. Self-service query tools through the web portal will be a vital end-user functionality provided by the DCC.

2. Participating Sites.

Each of the participating sites would have varying degrees of data management, integration and ETL capabilities. Sites with EHRs would be expected to generate extracts in the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) to ensure consistency and quality of submitted data.[6] The ODM is an XML file, and the specifications for generating this file are available from the CDISC website. Sites that choose the automated transfer of data to the DCC (Figure 2) will need to generate the ODM files using available ETL tools, and upload it to the SFTP site. Other sites might want to generate the ODM XML files using custom programming, and upload them manually through the frontend portal. The DCC will also provide basic ETL mapping specifications for sites that choose to generate the ODM XML files using ETL tools. However, due to differences in implementations, each site will have to perform an initial mapping of the CPRN CDM to the corresponding items in their respective EHR, and maintain an up-to-date version of their local mappings during their participation in the CPRN.

[6] CDISC ODM is designed to enable the acquisition, archive and interchange of metadata and data for clinical research studies. See http://www.cdisc.org/odm for additional information.

Data Collection

1. Point of care data collection via Electronic Health Records (EHR).

The majority of sites in the CPRN will collect data at the point of care using EHR systems like Epic, Cerner, etc. At least one site with the Epic EHR system and one with the Cerner EHR system will take the lead in developing the initial Structured Clinical Documentation (SCD) needed for Electronic Data Capture (EDC). For sites with the Epic EHR, the EDC strategy will require developing SmartForms and other types of SmartData or Flowsheet based SCD. Interviewer-based PROs and use MyChart or other Epic-integrated methods to collect patient reported outcomes. Additionally, REDCap at sites; and (3) REDCap at DCC. Similarly, for sites using the Cerner EHR, the EDC strategy will require developing PowerNotes, PowerForms, and IView based SCD. Sites using other EHR systems may need to rely on their respective forms of SCD and develop their own strategy. Epic sites will share technical knowhow regarding SCD design via Epic Userweb,[7] Cerner sites will share it via uCern,[8] and sites using other EHRs will have to develop similar strategies for sharing technical knowhow on SCD development. Since EHR vendors can vary significantly in the amount of customized content they allow sharing across their customer-base as well as exercise some amount of control over the method for sharing such knowhow, the sites will need to work within the constraints imposed by their respective EHR contracts. Consequently, the DCC may not be able to share this knowhow

[7] Epic Userweb is Epic's online user community, documentation, and collaboration portal: https://userweb.epic.com

[8] Cerner uCern is Cerner's online user community, documentation, support & collaboration portal: https://www.ucern.com/

directly, but will help play a role in facilitating such collaboration through the sharing of contact information, etc.

2. Chart Abstraction and Secondary EDC.

Certain sites may not possess the infrastructure or the resources to develop their own mechanisms for point of care SCD, and would have to rely on alternative approaches. The DCC will offer EDC using Research Electronic Data Capture (REDCap) software to such sites. This approach will require engaging Research Coordinators at the respective sites to perform chart abstraction on consented patients in the CPRN, and enter the various data points using REDCap forms hosted by the DCC. Of note, REDCap is 21 CFR 11[9] capable when configured and used correctly. The has been certified by the University of Utah Information Security Office has certified the DCC instance of REDCap as 21 CFR 11 compliant and Health Insurance Portability and Accountability Act (HIPAA) compliant.

3. Patient Reported Outcomes, Surveys and Long-Term Follow-up.

Patient surveys will be needed to obtain long-term follow-up information on patients in the CP registry. Although many healthcare organizations have developed methods for obtaining patient reported outcomes (PRO), and have the ability to send patient surveys through patient portals and other methods, some sites might not possess the infrastructure or be able to

[9] Part 11, as it is commonly called, defines the criteria under which electronic records and electronic signatures are considered trustworthy, reliable, and equivalent to paper records (Title 21 CFR Part 11 Section 11.1 (a)). Part 11 applies to drug makers, medical device manufacturers, biotech companies, biologics developers, CROs, and other FDA-regulated industries, with some specific exceptions. It requires that they implement controls, including audits, system validations, audit trails, electronic signatures, and documentation for software and systems involved in processing the electronic data that FDA predicate rules require them to maintain. A predicate rule is any requirement set forth in the Federal Food, Drug and Cosmetic Act, the Public Health Service Act, or any FDA regulation other than Part 11. See https://en.wikipedia.org/wiki/Title_21_CFR_Part_11#Coverage (accessed 02/07/2016).

dedicate resources to develop their own. The DCC will develop REDCap based surveys that will be used to collect PROs and other types of long-term follow-up information for sites that either do not possess such capabilities, or would otherwise like the DCC to handle this part. The DCC will also share the specifications for PROs and other REDCap surveys with sites that choose to deploy their own infrastructure to collect these data. Sites that choose to collect their own PROs and other long-term survey data will need to integrate these data into the XML extracts they submit to the DCC.

Data Extraction

1. Generation of Extracts from the EHR Database and Other Vendor Solutions.

For sites that are planning to collect data at the point of care using their EHR systems, there are several options for how these data will be extracted and packaged as a CDISC ODM XML file. The key question is whether the sites rely on vendor-provided solutions for extracting data from these EHR systems, whether they have a mature Enterprise Data Warehouse (EDW) or other operational data stores and/or infrastructure like i2b2. The Epic EHR system includes a solution called Clarity for extracting data from the Intersystems Caché[10] backend into a relational database, and some sites may have also implemented Epic's Cogito EDW solution.[11] Similarly, Cerner EHR includes solutions like PowerInsight EDW, etc.,[12] which help extract data

[10] Intersystems Cache is a non-relational database management system that serves as the backend for Epic. More information can be found here: http://www.intersystems.com/our-products/cache/cache-overview/ Since the vast majority of ETL and Analytics tools rely on an underlying relational database, extracting data from the Epic EHR involves first moving it from Cache to a relational DB like Oracle or Microsoft SQL Server.
[11] See https://www.epic.com/software-intelligence.php for more information on Epic Cogito integrated analytics platform, which includes a unified EDW as well as analytics tools like SAP BusinessObjects.
[12] See https://www.cerner.com/Solutions/Analytics_Reporting/Enterprise_Data_Warehouse/ for more information on Cerner's EDW and analytic solutions which also include SAP BusinessObjects.

from the EHR's operational data store. One of the key advantages for sites that have implemented vendor-provided solutions for base data extraction is the ability to share the ETL approaches with other sites that have implemented compatible technologies, and to reduce duplication of effort. The DCC will help with the development of initial ETL specifications for sites that are prepared to allow that level of access to the underlying systems. The DCC will also share this expertise with other sites that have similar deployments, via vendor-approved channels like Epic Userweb and Cerner uCern.

2. Generation of Extracts from Other Sources.

Some sites might have their own custom EDW or i2b2 implementations, instead of relying on EHR vendor provided solutions for data extraction and those sites would need to develop their own custom processes for preparing the CDISC ODM XML files for the registry. The DCC will provide a sample ETL specification for this purpose, and provide basic support for answering questions, but these sites will have to rely on their own developers to develop these processes further. Since i2b2 is a common platform used by many research data warehouses and other research networks, the DCC will develop ETL processes or adapt other i2b2-ODM solutions like the Integrated Data Repository Toolkit (IDRT).[13]

3. ETL Virtual Appliance.

For sites that are planning to collect data at the point of care using their EHR systems, the DCC will develop an ETL Virtual Appliance (EVA) solution (Figure 2) for common EHRs like Epic,

[13] The Integrated Data Repository Toolkit (IDRT) approach relies on Talend Open Studio workflows to import data form an ODM format into an i2b2 repository. We will investigate repurposing some of these workflows to generate ODM mxl files. More information on IDRT is available here: http://idrt.imise.uni-leipzig.de/IDRT-II/ , while source code is available here: https://github.com/tmfev/IDRT-Import-and-Mapping-Tool

Cerner, and for platforms like i2b2, available under a limited distribution. EVA will feature a minimal Linux OS footprint with pre-configured Talend Open Studio ETL workflows to extract data from common EHRs, and to transfer them automatically using SFTP to the DCC. The DCC is working on prototype EVA ETL workflows using Talend, and expects to deliver a deployable virtual appliance toward Phase III / IV of the project. EVA will lower the entry barrier for newer sites wanting to participate in the CPRN. The key steps, phases, and a proposed timeline (in weeks) is shown in Figure 4 below.

Timeline

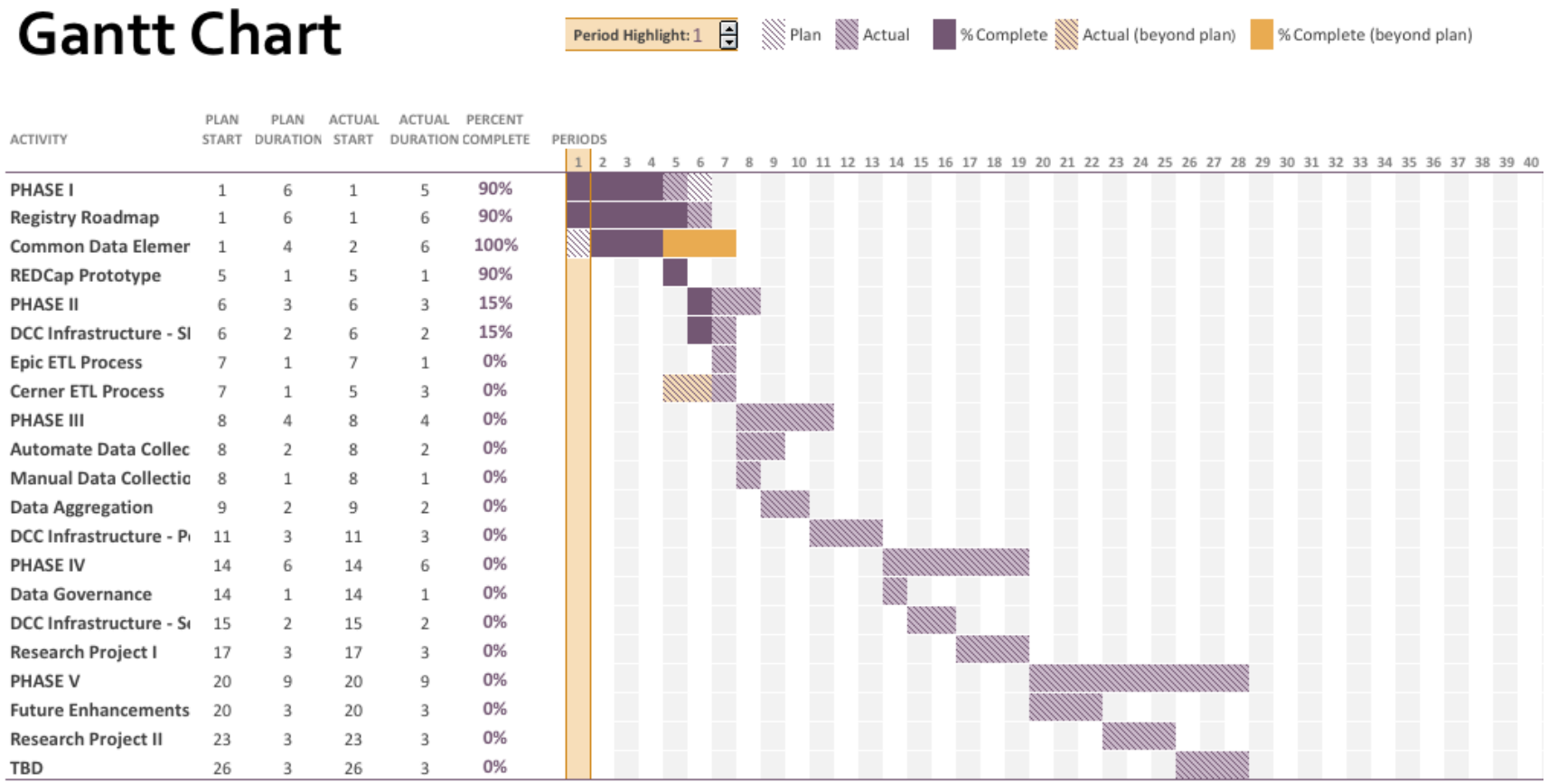

Figure 4: A Gantt Chart chowing various phases, tasks and a proposed timeline (in weeks) for developing the CP Registry.

Abbreviations & Definitions

CDISC: Clinical Data Interchange Standards Consortium. An open, vendor-neutral, multi-disciplinary, non-profit standards development organization that develops global standards for streamlining the collection of research data and integration with healthcare.

CPRN CDM: Cerebral Palsy Research Network Common Data Model. A standard dictionary for research data collection in the CPRN.

CPRN: Cerebral Palsy Research Network. A collaborative data-sharing partnership of medical centers that aims to improve treatments and outcomes for persons with cerebral palsy through high quality clinical research and quality initiatives.

DCC: Data Coordinating Center. Centralized data management and coordinating center at the Division of Health System Innovation & Research, Department of Population Health Sciences, University of Utah, School of Medicine.

DM: Data Mart. A data mart is a subset or logical area of a data warehouse environment that includes a specific subject area, business line, or a group. In this project, CP data will be organized into a longitudinal DM containing all the data elements needed to answer research questions for the CP registry.

EDC: Electronic Data Capture. The method of electronically collecting a bulk of the data within the EHR at the point of care.

EHR: Electronic Health Record. Clinical information systems like Cerner, Epic, etc., which are used to capture patient records electronically.

ETL: Extract-Transform-Load. ETL processes are used to extract data from SCD fields in an EDC source system or files, process and transform them into a different form and load them into the

target database system. For example, in this project, one of the ETL processes will rely on extracting data from Epic systems SmartData elements, transform them into an XML file, and then upload it to the DCC's SFTP server, where another ETL process will process the filem transform the XML fields into suitably mapped fields for target tabled in the DCC database and load the data into the database.

FIFO: First-In-First-Out. In this context, an approach for processing ODM XML files received by the DCC in the order of upload.

HA: High Availability. HA refers to a type of server environment where multiple, redundant servers are "clustered" together to perform the same type of task, so that the failure or loss of a single server does not lead to the entire application becoming unavailable, because other servers are available to perform those same tasks.

ODM: CDISC Operational Data Model. The CDISC ODM is a vendor neutral, platform independent format for interchange and archive of clinical study data. The model includes clinical data along with its associated metadata, administrative data, reference data and audit information, and is implemented using XML.

OMOP: Observational Medical Outcomes Partnership. OMOP is a public-private partnership established to inform the appropriate use of observational healthcare databases for studying the effects of medical products. The OMOP consortium promulgates standards for data representation modeling and querying for observational studies.

OMOP-CDM: OMOP Common Data Model. The OMOP CDM aims to standardize the format and content of observational data to enable standard analysis and frontend tools, applications, etc. to be applied to them.

OVM: Oracle Virtual Machine. OVM is a server virtualization product from Oracle that incorporates the free and open-source Xen Hypervisor to deliver near native performance and provides horizontal scalability combined with Oracle RAC.

RAC: Oracle Real Application Cluster. The RAC option in Oracle provides clustering and high availability to provide performance, scalability and resilience, and minimize downtime by distributing the workload among several nodes that tap into a common database in a share-everything configuration.

REDCap: Research Electronic Data Capture. REDCap is a browser-based metadata-driven EDC software that can be used to collect forms, patient surveys, etc. The University of Utah instance of REDCap has been certified by the Information Security Office as being HIPAA compliant as well as 21 CFR 11 compliant.

SCD: Structured Clinical Documentation. SCD is a form of EDC where each of the data elements defined in the CDE is captured in a structured form, making it computable and easily accessible to ETL methods. Contrast SCD with narrative notes, which are inherently difficult to extract individual observations from, and which are not as amenable to ETL methods described in this document.

SFTP: Secure File Transfer Protocol. SFTP is a network protocol that provides secure file access, file transfer, and file management

XML: eXtensible Markup Language. XML is machine-readable and human-readable markup language that is defined by the World Wide Web (WWW) Consortium's XML 1.0 specification.