DATA HARMONIZATION PROCEDURE [DRAFT MARCH 18, 2021]

SECTION		ORIGINATION DATE	March 2021
SUBJECT	Data Harmonization	EFFECTIVE DATE	(draft)
POLICY NUMBER	PCDC-	REVISION DATE	March 18, 2021
POLICY OWNER	PCDC	REVISION NUMBER	1.0

PURPOSE

The University of Chicago, as the service provider of the Pediatric Cancer Data Commons (PCDC), is committed to creating an efficient and effective process for data to be harmonized and ingested into the Commons. The purpose of this document is to describe the process and procedures of data harmonization efforts and outline the roles and responsibilities of the PCDC team and the disease consortia.

POLICY

The PCDC data harmonization process consists of two parts.

First, each disease group consortium needs to have three basic components in place:
1. A data dictionary
2. A tiering of data elements and availability of data by element that is driven by priority research question(s)
3. Data quality rules and governance

These elements are explained in further detail below.

Second, the PCDC will provide guidance on how to submit data and how to ensure that this process will conform to PCDC quality standards.

PROCEDURE

A. BEGINNING ELEMENTS
   Prior to beginning the data harmonization process, the disease consortium from which data will be contributed should have the following elements established.
   1. A stable and complete version of the data dictionary.
   2. A tiered system for prioritizing variables in the dictionary to guide harmonization efforts. Ideally, this prioritization will occur during the data dictionary development process.
   3. Data quality rules and governance.

For large dictionaries, not all variables are of equal importance nor do they all exist in upstream source data. Furthermore, there may not be sufficient resources available for harmonization of all data elements. The establishment of prioritization tiers helps to ensure that the consortium has consistent and usable data ingested into the commons. This system should be established by consortium executive leadership and representatives from the data dictionary creation process. In the absence of guidance to contributors from the consortium executive leadership, the data could become a patchwork of sparsely populated variables for large dictionaries.

4. Data quality rules and governance.
   The PCDC needs metrics by which to evaluate incoming data, and there should be a consortium-level data quality group to which the PCDC technical team can report its findings. Particularly important here are: an enumeration of which fields are required, acceptable ranges for numeric fields, internal consistency checks (e.g. no metastatic sites if staging indicates non-metastatic disease, etc.). High-quality data should be a baseline expectation rather than a downstream aspiration. Having the set of rules on which data quality will be evaluated is immensely helpful for individuals who will be tasked with doing the harmonization work.

B. CONSORTIUM-SPECIFIC DATA SUBMISSION TEMPLATES
   The PCDC technical project manager will create a data submission template for data contributors and those who need to harmonize that data, which will be made available online for data contributors to access.

C. HOW TO POPULATE TEMPLATES WITH DATA
   The PCDC provides guidance to data submitters on how to populate templates with data. This includes information and case descriptions that will be available on the PCDC website. The PCDC will also offer virtual information sessions to provide an overview of the process and answer specific questions that the PCDC community may have.

D. QUALITY CONTROL
   The PCDC will provide quality check guidelines and accompanying software that contributors can execute locally to warn them of data quality issues prior to sending data to the PCDC. In addition, automated tests will run before the data are finalized and uploaded in the common.

FORMS

DEFINITIONS

Data harmonization

RELATED STANDARD OPERATING PROCEDURES

REFERENCES

REVIEW STATEMENT

PCDC will maintain the status of this activity and conduct audits to ensure compliance. This procedure will be reviewed by the **leader of the PCDC Quality Program** every two years or in a timely response to changes in local or federal regulations or in response to a significant change in PCDC operations and technology. Modifications to the procedure will be made, as needed.

APPROVAL

Reviewers
- Members of the PCDC Data Quality Program:
    - PCDC Leadership
    - PCDC Pillars Leads

Approver(s)
- Director of the PCDC