Title: A Policy Framework for Inclusion of Uploaded Publicly Available External Data Sets in the N3C Data Enclave	
Version No: 1.0	**Effective Date: 2021-02-21**
Authors:	**Dave Eichmann, Anita Walden**

Version history

Version	Description of changes
All versions	**10.5281/zenodo.4562982** This DOI represents all versions and will always resolve to the latest one.
1.0-2021-02-21	Approved for implementation

Purpose

This policy establishes expectations and a process (available here) for requesting and approving inclusion of a publicly available dataset (PAD) in the National COVID Cohort Collaborative (N3C) Data Enclave, as well as the request for use of an external data set available in the Enclave in a given Data Use Request (DUR). This policy and process is designed to help manage the privacy and security of the Enclave data and ensure that the data contributed to the LDS are protected and in compliance with the terms of the Data Transfer Agreement (DTA) and Data Use Agreement (DUA), and any statutes or regulations while enhancing the potential value of research conducted in N3C by adding other publicly available datasets.

Scope and Applicability

The scope of this policy includes the request to link to additional publicly available datasets (PADs) by the N3C community and others as identified in the process document, their importation into the N3C Data Enclave, and their use within the Enclave. All individuals requesting addition of a PAD and/or requesting use of a PAD must be users within the Enclave who have an active Data Use Agreement (DUA) with NCATS.

This policy does not apply to the addition or use of external datasets that are not publicly available, such as institution or investigator(s) owned dataset(s) or others. Inclusion of those datasets will be considered in the future and covered under a separate policy on addition of non-public datasets or through N3C activities related to the Privacy Preserved Record Linkage (PPRL, or sometimes referred to as hashing) process. Once ready, information on those policies and processes will be available here.

Definition

Publicly Available Dataset (PAD): A publicly available dataset is one that is either readily available through such means as a public website or is readily provided by the holder of the dataset to anyone who requests access to it. A PAD is an external dataset requested by the community to enhance research after going through a multi-step review process described below and before being made available in the Enclave. These datasets are generally freely available, though there could be some nominal costs or licensing considerations. PADs may include identifiable information, but because the information is publicly available, it is not considered exempt human subjects research (46.104) under the Common Rule. Because of the nature of some datasets, they may be considered for use within N3C through the PPRL/hashing process rather than through this policy and process.

Reference Documents

External Dataset Process Document

Key Aspects of the Policy:

1. ***Addition of a Publicly Available Dataset in the N3C Data Enclave***: Publicly available datasets (PADs) will be considered for inclusion in the N3C Enclave to expand and enhance research conducted using the N3C Enclave data. Any researcher who has been granted access to N3C and who has an active DUA may submit a request to include a PAD. Prior to inclusion of a PAD, a multi-step process will be undertaken to evaluate the benefits of including the dataset into the N3C Data Enclave. NCATS will determine which PADs will be available within the N3C Data Enclave, based on several factors when NCATS and the N3C Community assess the "value" of including a given PAD, including but not limited to:
    - **Scientific utility/value,** particularly whether the dataset will contribute to and/or enhance the understanding of COVID-19 and adds value to the use of the N3C Data Enclave;
    - **Level of effort** (LOE) required to include the PAD;
    - **Copyright, licensing, cost implications** for inclusion and determination of who will pay (if there are costs associated);
    - **Public Availability**: Confirming that the dataset is publicly available; and
    - **Security Assessment**: The prohibition of re-identification is part of the framework of using the N3C Data Enclave through the terms and conditions of the Data Transfer Agreement, Data Use Agreement, and the User Code of Conduct; however, NCATS will assess whether there are data security or privacy risks that would reduce enthusiasm for including a specific PAD.

    Review and assessment of these factors will be conducted by NCATS Staff. The N3C Community Tools and Resources Review Committee will provide input on the scientific value/utility. NCATS or the N3C Community Tools and Resources Review Committee may also consult N3C Community domain experts, as needed. The Standard Operating Process (SOP) for requesting addition of a PAD outlines this.

2. ***Use of an External Dataset in the N3C Data Enclave***: Users (or individuals requesting access) to the N3C Data Enclave may use PADs that have been imported into the N3C Data Enclave. Users who wish to add one or more PADs will have access in the Enclave to any that have been approved by NCATS and made available. Although NCATS and the N3C have determined that there is value to inclusion of a given PAD in the N3C Data Enclave, it is the responsibility of the user(s) involved in a given DUR(s) to determine whether use of the PAD(s) makes N3C patient-level information readily identifiable and to ensure that their institution and/or Institutional Review Board (IRB) (or other such group) have assessed the risks/protections of human subjects in the context of their specific research question and any federal, sate, local, or other regulations or policies that may apply to them.