N3C Results Download Policy

Title: N3C Results Download Policy	
Version No: 1.0	**Effective Date: 2021-05-07**
Authors:	**Chris Chute and Anita Walden**

Version history

Version	Description of changes
All versions	10.5281/zenodo.4743235 This DOI represents all versions, and will always resolve to the latest one.
1.0-2021-5-07	• First draft

Goals

Investigators using the rich data compiled in the N3C Data Enclave are expected to generate quantitative results in the forms of tables, figures, parameter estimates, and aggregated statistics. These results may have broad impact and will be shared, typically in the form of manuscripts, reports, and visualizations for websites and seminars. Additionally, many artifacts such as workflows, value sets, and phenotyping algorithms may be developed in the Enclave and made useful to others by being exported for presentation on public websites such as GitHub, or incorporated as supplementary materials to publications. The purpose of this policy is to establish a process to permit the download of quantitative artifacts used in analyses that are necessary for sharing results, while at the same time ensuring protection and security of the data in the Enclave and compliance with the NCATS Data Transfer and Data Use Agreements.

Documents

Results Download Process Document

Policy

Quantitative Data

Investigators attest to the N3C Data User Code of Conduct, which prohibits downloading data beyond findings of analyses specified in the approved Data Use Request (DUR). DURs outline the goals of the research to address COVID-19 related research questions. Summary statistics of data findings approved by NCATS for sharing and made available on operational dashboards of the N3C website may be downloaded by authorized personnel listed in the operational DURs that correspond to those dashboards. Research findings in the form of quantitative or tabular results can be exported from the N3C Data Enclave environment with the following provisions:

- No observations or patient data at an individual level can be downloaded.
- Aggregate data can be downloaded if the aggregation pertains to 20 or more persons in the N3C dataset. No table or figure can contain cell results deriving from fewer than 20 persons, unless the value is zero. Cells with counts less than 20 will be identified with a symbol (often seen as <20 or *). On a case-by-case basis, users may request whether, on the basis of scientific interpretation, the number could be reduced to 10. The basis of scientific interpretation to reduce this to 10 on a case-by-case basis.
- The obfuscated site IDs used in the Enclave must be re-mapped to a protocol/paper specific remapping, e.g. sites 1-30. Actual Enclave site IDs may not be used in any public documents.
- Ages can be included in the data, unless the patient is age 90 and over. Patients over age 89 can be included in their own category as patients age 90 and older.All quantitative data to be downloaded will be copied to a staging folder on the N3C Data Enclave. It will also be archived on an NCATS hosted permanent location for monitoring purposes.
- Metadata about the downloading user, including their ORCiD, will be attached to the staging and archive area data.
- PHI and date detector software will be run over newly staged materials.
    - Dates may suggest row-level data, the download of which is prohibited
    - Five-digit zip codes as fields may also suggest row level data
- A manual audit of all proposed data will be conducted to ensure it does not represent row-level data.
- The download requestor will be informed by email as to how to retrieve tabular results that are manually determined to be appropriate for download.
- Datasets that are determined to violate these policies after manual review will be denied for downloading. Violations will be addressed as described by potential sanctions outlined in the Code of Conduct document.

The Results Download Review Subcommittee will be sought from N3C community membership as a subset of the N3C Publications Committee. The Results Download Review Subcommittee , a subcommittee of the N3C Publications Committee, will be structured to include members with

clinical, data science, and privacy backgrounds. It will be solicited by N3C operations with membership vetting by the N3C Governance Committee. NCATS and the Results Download Review Subcommittee will review requests, and then refer them to the committee for additional input. Additional details on how the operations of this committee will be implemented will be available in the Results Download Review process document.

Knowledge objects (logic, concept sets, or program code only, no data or protected health information)

Investigators and research teams will generate derived variables, harmonized data elements, concept sets/codesets, machine learning models, code workbooks, logical workflows, and other knowledge artifacts. None of these resources will include row level data or PHI from clinical sources to enable voluntary and fair sharing of such resources. Most of these artifacts will be beneficial to other users within the N3C community. The goal is to enable the fair sharing of such resources with appropriate attribution (particularly through Zenodo and/or GitHub).

Codesets and harmonized variables will be placed into a shared folder in the N3C Data Enclave and indexed in the Data Resource Library (in development), which is an expanded data dictionary. Provenance and attribution metadata will be incorporated into the library. Code repositories and workflows can be stored in N3C or community Github repositories and cloned or downloaded to public Github repositories or other sites. If these workflows were derived from limited dataset (LDS) sources, they can be reused on any Enclave artifact. Users who wish to utilize workflows derived from LDS sources must submit a DUR requesting access to LDS level data. No N3C or NCATS review or permission is required for knowledge object sharing. Please see the N3C Attribution and Publication Policy for updated details regarding attribution of knowledge objects. Requestors may license their code using a non-viral open source license.

Images/visualizations

Analytic workflows can create graphical images or figures. These images can be downloaded by users in svg/png/jpg format without special permission. The download of analytics figures and images is not presently restrictable but is logged. It is technically possible that some images could contain data about a single individual. Release of data specific to the description of an individual or characteristic of a single data contributing site is not allowed per the User Code of Conduct and the NCATS N3C Data Use Agreement. From Section F of the NCATS N3C Data Use Agreement, users agree not to photograph, create screenshots, or download data viewed on the NCATS N3C Data Enclave. Furthermore, users' access to the NCATS N3C Data Enclave will be terminated after one year if an updated DUR is not submitted for renewal. Users' access to their analyses and findings on the NCATS N3C Data Enclave will continue if Data Access Committee (DAC) approval of the DUR renewal is granted. The Results Download Review Subcommittee will periodically inspect a random sample of logged image downloads to determine whether any data on a patient instance has been exported. Violations will follow the process outlined in the User Code of Conduct and the NCATS N3C Data Use Agreement.

Release for archival purposes

Once any artifact is ready for public release - e.g. it is the final code or knowledge artifact used in a published manuscript or is otherwise ready for fully provenanced community reuse, an archival record with a DOI for the artifact should be created in Zenodo. It is straightforward to create a versioned release archive in Zenodo directly from Github (see https://guides.github.com/activities/citable-code/). Please request inclusion of your artifact in the N3C Zenodo collection so that the N3C community can more readily find them.