

## **Data Sharing Plan**

**Primary Goal:** The Clinical Trials Network (CTN) is creating an electronic environment that allows data from their completed clinical trials to be distributed to investigators to promote new research, encourage further analyses, and disseminate information to the community. This data sharing plan is designed to enable access to as much data as possible, while protecting human subject identity and ensuring proper use. Our goal is to remove any patient, site or other sensitive information from each completed CTN trial's final database to make it suitable for public use (see the NIH public data sharing policy at [http://grants.nih.gov/grants/policy/data\\_sharing/](http://grants.nih.gov/grants/policy/data_sharing/)). This will allow researchers interested in drug abuse studies to download data from completed trials in order to conduct analyses that improve the quality of drug abuse treatment.

**Introduction:** Effective data sharing includes communicating to the research community that data are available, providing sufficient descriptive about the data, enforcing compliance to standard semantics and structure, rendering the data in a usable format, and making data accessible. Whether data standards, including standard controlled terminology, are implemented in the data collection design, a master data model supports subsequent consolidation. This content model should be designed based on the known data collected for the therapeutic area so that it is comprehensive enough to include future trials while maintaining backward compatibility.

**Protecting Human Subjects:** The primary concern when sharing data from any study is the protection of human subjects. Thus, we need to understand our responsibility for confidentiality, protected health information (PHI), and subject identities during the planning for sharing data derived from the research.

Individual identifying information is defined as any information that is a subset of health information, including demographic information that is either created or received by a health care provider, health plan, employer, or health care clearinghouse. This demographic information could relate to the past, present, or future physical or mental health or condition of an individual, the provision of health care s/he receives, or the past, present, or future payment for the provision of health care to an individual. Thus, this type of data either identifies the individual or there is a reasonable basis to believe the information could be used to identify the individual. (see *45 CFR Part 160 subpart A*)

Sensitive data such as data regarding alcohol abuse, illegal substances, mental health, sexual activity, sexually transmitted diseases, HIV, violence, financial status, or homelessness raise special concerns about confidentiality and the protection of privacy. These data have an increased likelihood of causing harmful social, economic, or legal consequences if they become available. The National Institute on Drug Abuse (NIDA) Clinical Trials Network (CTN) has recognized this issue and taken steps to prevent an individual's clinical trial data from being brought into a court of law under subpoena, such as putting a process in place for obtaining Certificates of Confidentiality.

The collection of sensitive data does not preclude public sharing data. The NIH expects and supports the timely release and sharing of final research data from NIH-sponsored studies. The sharing of data promotes new research and makes possible analyses that were not explored by the original investigators. In a policy issued on February 26, 2003 (<http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>), the National Institutes of Health (NIH) reaffirmed its support for the concept of "data sharing," that is, making data from a research study available to other interested researchers. This policy states: "We believe that data sharing is essential for expedited

translation of research results into knowledge, products, and procedures to improve human health. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.”

The secondary analyses produced from data sharing multiply the scientific contribution of the original research. Use of universal standards on data shares has the potential to further magnify this impact by providing uniformity across therapeutic areas and translational boundaries. An example of the NIH Roadmap in action, the interoperability obtained from implementation of data standards facilitates combining research results obtained in vastly separate areas to explore unique questions using data common to all datasets. Using data obtained from multiple clinical trials would enable more meaningful meta-analyses. Thus, to further the research on drug abuse treatments, the CTN has created a Data Sharing Policy for any researcher who would like to use data from completed CTN trials for research, analyses or publication.

**Data Protection:** Sensitive data do call for a higher level of security during collection, analysis, and storage and special consideration when preparing datasets for broader use. The rights and privacy of people who participate in NIH-sponsored research must be protected at all times. Therefore, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects.

Assuming that no restrictions are imposed by institutional policies, the local IRBs, as well as local, state and Federal laws and regulations, including the Health Insurance Portability & Accountability Act (HIPAA) privacy rules, we are prepared to make available data generated from all CTN completed trials via a public web site. Our proposed data sharing plan includes complete de-identification of the proposed datasets according to HIPAA definitions. A de-identified data set is not subject to HIPAA's minimum necessary standards, so **ALL** of the data in the de-identified dataset can be included and shared. Under recent guidance from OHRP, a de-identified dataset can be declared to be "not human subjects" and thus not subject to the recipient's IRB review.

**Preparing Shared Data (De-Identification):** To de-identify the data all Personal Health Information (PHI) must be removed in compliance with the HIPAA privacy rule. This will make the data free of identifiers that would permit linkage to the research participants and free of content that would create unacceptably high risks of subject identification. Also, indirect identifiers that are not listed as PHI but could lead to “deductive disclosure” of subject’s identity will be removed. This is more likely in small, geographically limited or specialized populations. Examples of data that we consider for de-identifying on a variable or field level are the following: comment fields, optional fields, other specify fields, site number, node number, investigator, and site name. De-identification of a dataset means removing the following variables:

- Names
- Geographic information (including city, state, and zip code)
- Dates such as birth, hospital admission & discharge, and death dates
- Telephone and fax numbers
- Electronic mail addresses
- Social Security Number
- Medical record and prescription numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate or license numbers

- Any vehicle identifier or serial number, including license plate number
- Any device identifier or serial number
- Web Universal Resource Locator (URL)
- Internet Protocol (IP) address number
- Any biometric identifiers, including finger or voice prints
- Full face photographic images or any comparable images
- Any other unique identifying number, characteristic, or code consisting of any segments of the previously listed identifiers.

Clinical study data lacking time variables, demographic information, and information regarding risk factors is often not very useful. Therefore, de-identification must come with tools to rebuild and or retain the aforementioned information. The time information is retained by picking a reference date, the same for all subjects (i.e., enrollment), and calculating all dates as the number of days since the reference date ("Day 0"). Variables are added back to the dataset that express all dates as number of days since the reference date, as well as a variable storing just the year.

Information regarding geographical location and site that would jeopardize identity by having too small of numbers within a geographical vicinity or a small number of sites will not be available in the de-identified data. For comment and specify fields, we first recommend not collecting them unless absolutely necessary. Then, only to report the coded data, or the other, and not the specify field. Data collection situations will always occur for which no predefined solution exists. In these cases, we have always worked with the statistician and trial or network leadership to define an acceptable reporting solution. These study-specific solutions will be documented with each completed dataset.

**Link Field:** The de-identified dataset can include a 'link' field to connect back to the original, identified, data. This field cannot be derived from any data in the original dataset. This linking field is computed using a random number generator and cannot be traced back to the original study identifier without a "key". This "key" will not be maintained once the data are de-identified.

**Data Conversion:** Prior to de-identifying data, data will be converted from its native format to a modified Study Data Tabulation Model (SDTM) standard format. The SDTM is a content standard sponsored by the Clinical Data Interchange Standards Consortium (CDISC). This universal data format will be applied to each completed CTN trial's data. This will facilitate the pooling of shared data across completed studies. SDTM is a well-documented, platform-independent data standard that researchers will have access to training and implementation guides.

**Data Format:** Data will be available for download in two formats, SAS and ASCII. SAS files will be SDTM CDISC transport files and the ASCII data will be CSV files. The data will follow standard SDTM mapping rules including the fact that supplemental qualifiers will be available in separate files and not attached to the main files. Documentation regarding study-specific SDTM methods will be given for each study protocol.

**Documentation:** Adequate study and data documentation describing the data source will be available for each posted CTN trial data in Adobe format, so that researchers can accurately download and use the data. The study documentation for each completed trial will include the final protocol and links to primary publications. The data documentation will include the following:

- Annotated CRF (maps data variables to source case report form (CRF))
- Define.xml File (outlines the structure, key variables, and contents of the dataset and any derived data)
- SDTM and de-identification mapping rules

**Data Availability:** Research activities for a completed study should be shared with the scientific community in a timely fashion. A period of 18 months from completion of the study or after the acceptance of publication of the primary manuscript is a reasonable criterion for when data should be made available to the public. We will post the trial description information, data collection forms, and data structure on a website after the primary manuscript is accepted for publication or 18 months after database lock, whichever occurs first. Until that time, a message will appear on the web site stating that the “Trial data will be available after the primary manuscript is accepted for publication or 18 months after study completion, whichever occurs first.”

**Data Location:** De-identified study datasets will be available for download on a public web site (<http://www.ctndatashare.org>) using CDISC standards. The data will be provided in SAS (CDISC SDTM) and ASCII file formats.

Any data downloaded from the web site will be subject to the CTN Data Sharing Agreement. Requesters will have to register a name, position, affiliation, and valid email in order to download data to state that they are responsible for using the data in accordance with the CTN Data Share Agreement. The data will only be made available to the requester under a registration agreement containing the following stipulations:

1. Not to use the data, either alone or in conjunction with any other information, in any effort whatsoever to establish the individual identities of any of the subjects from whom the data were obtained.
2. To retain control over the received data, and not to transfer any portion of the received data, with or without charge, to any other entity or individual.
3. To notify the Recipient's Institutional Review Board (IRB) operating under an Assurance approved by the Office of Human Research Protections (OHRP), and in accordance with Department of Health and Human Services regulations at 45 CFR Part 46 of any research projects resulting from CTN data.
4. To acknowledge the CTN database and the specific trials accessed in all oral and written presentations and publications resulting from analyses of the received data. CTN data should be cited as follows: “The information reported here results from secondary analyses of data from clinical trials conducted as part of the National Drug Abuse Treatment Clinical Trials Network (CTN) sponsored by NIDA. Specifically, data from trial “CTN-XXXX” were included (actual protocol number). CTN trial databases and documentation are available at [www.ctndatashare.org](http://www.ctndatashare.org)”.
5. To maintain the security and privacy of the received data for as long as necessary per local and federal requirements.
6. To inform the CCTN Data Manager when research paper(s), which include(s) the use of the CTN clinical trials data, is (are) accepted for publication.
7. Complete the requested fields of the registration agreement.

**Other:** There are a few situations where a fully de-identified dataset might not be adequate, for example in doing some cyclic trend activities where either the cycle is less than a year or the actual order in which the cases present to the clinic is important, or an analysis that includes site in the statistical model is required. If any of these analysis situations present themselves, researchers can contact the CCTN Data Manager for assistance. Additional data may be possible as either a (1) Limited Data Set with Data Use Agreement, or (2) Fully Identified Data with subject authorization and IRB approval.