Production of Deidentified Datasets at the Conclusion of Human Subjects Trials

Korpak A, Schoenfeld D; MGH Biostatistics Center

ABSTRACT

At the conclusion of a study, researchers often have an obligation to publish their data, but simultaneously are required to preserve confidentiality of human subjects. For instance, NIH research contracts require distribution of data after publication, and call upon the researchers to provide a deidentified dataset, one that is appropriately stripped of all information that might be used to identify subjects or be deemed sensitive information. There are published guidelines for how to produce a deidentified dataset, but these are written in general terms. There is clearly a need for more developed guidelines. The ARDS Network has produced three deidentified datasets for NIH studies. Through these projects, we have produced detailed documentation regarding the creation of such datasets, rooted in the NIH guidelines and refined through application. The final process is generalizable to other studies involving human subjects, and serves as a useful guide for researchers facing the same challenge.

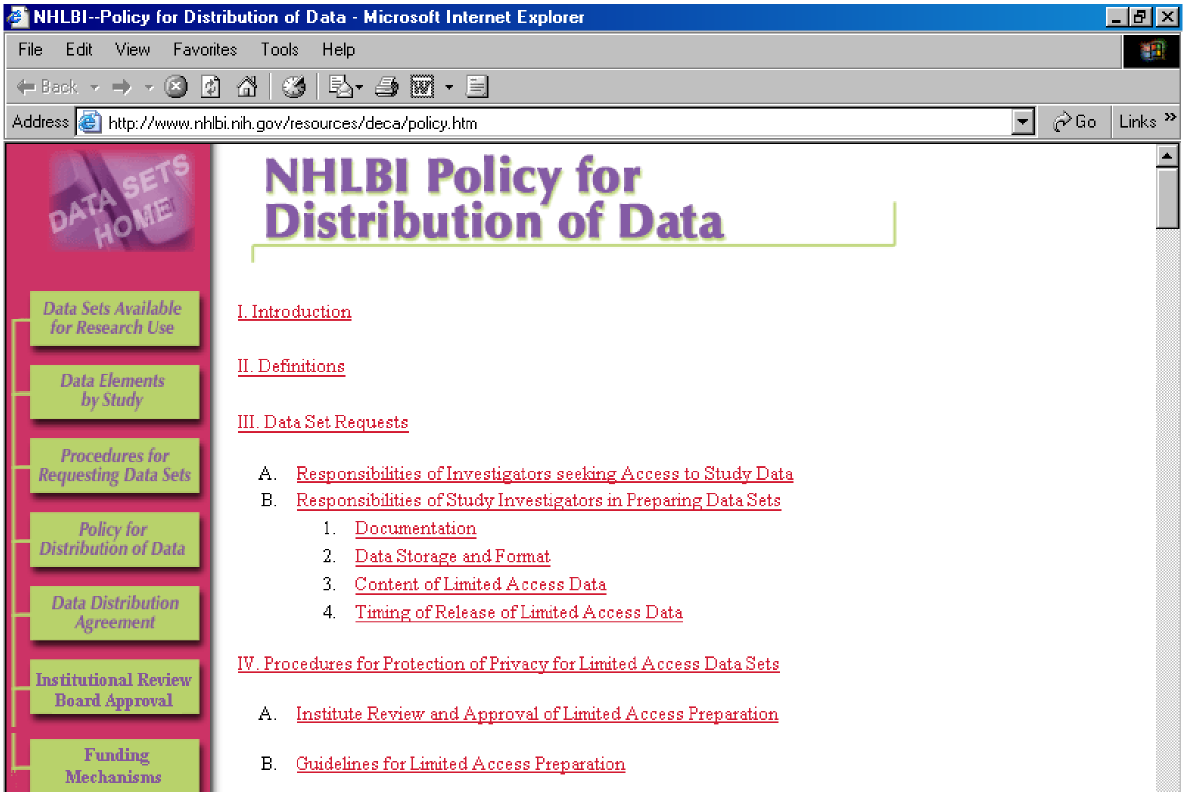
The ARDSNet procedure for producing a deidentified dataset addresses an array of issues, including:
- The audit trail, which is required for datasets produced through electronic data capture systems, produces fields which are prohibited in a deidentified dataset. These fields include user names and date of data entry, which potentially provide geographic and temporal clues to patient identity.
- Data points such as age, weight, and height are not necessarily identifying information, but can be considered so in the case of uncommon values. For instance, reporting an age greater than 89 years old is considered a risk to confidentiality.
- Most dates are potentially identifying data. By converting dates to integer study days, relative to a reference date such as the date of enrollment, this information may be maintained without risk of identifying a subject.
- Patient numbers can be considered identifying if they contain data about a patient's location, date of enrollment, or similar information. For the ARDSNet trials, we created a list of simple sequential patient IDs and then matched it to an unsorted list of the original patient numbers; the resulting table was then used to translate patient IDs in all study tables to the new patient ID.
- Free text fields require special consideration, as they potentially contain any kind of information. Obviously, references to dates, names, locations, etc. should be removed. Less predictably, references to sensitive information that is not necessarily identifying in nature should also be removed; references to drug abuse fall into this category. In cases of free text variables, the difficulty of systematically eliminating all prohibited information must be carefully weighed against the importance of the legitimate information recorded in the data field.

OFFICIAL GUIDELINES

Researchers are responsible for ensuring that datasets provided for distribution are appropriately modified to remove identifying information.

NHLBI policy requires removal of all obvious identifiers, such as:
- name
- social security number
- family relationships / pedigree
- hospital record numbers

An NHLBI website provides guidelines regarding other data elements.


NHLBI--Policy for Distribution of Data - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Address
Go
Links
DATA SETS HOME
Data Sets Available for Research Use
Data Elements by Study
Procedures for Requesting Data Sets
Policy for Distribution of Data
Data Distribution Agreement
Institutional Review Board Approval
Funding Mechanisms

Address: http://www.nhlbi.nih.gov/resources/deca/policy.htm

NHLBI Policy for Distribution of Data

I. Introduction

II. Definitions

III. Data Set Requests

A. Responsibilities of Investigators seeking Access to Study Data
B. Responsibilities of Study Investigators in Preparing Data Sets
1. Documentation
2. Data Storage and Format
3. Content of Limited Access Data
4. Timing of Release of Limited Access Data

IV. Procedures for Protection of Privacy for Limited Access Data Sets

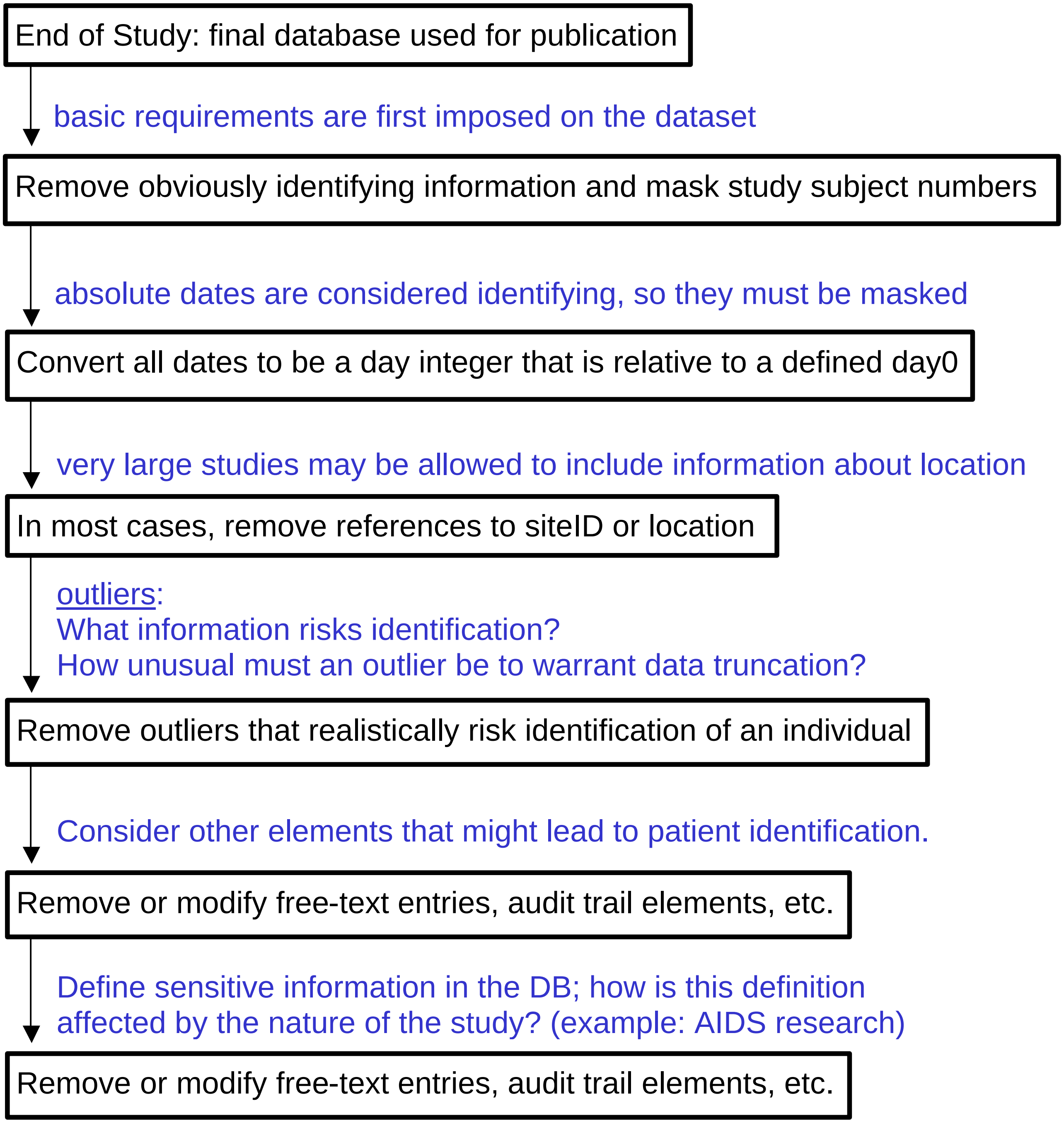
A. Institute Review and Approval of Limited Access Preparation

B. Guidelines for Limited Access Preparation

PROCESS

A chart of *one possible* process for producing a deidentified dataset (also known as a "limited access" dataset) is shown. The text between boxes suggests some items for consideration at each step in the process.

There are few hard-and-fast rules. Study personnel should collaborate with the study sponsor to determine what is appropriate in the given case. The process you develop for your own study should be well-documented, with documentation distributed along with the resulting dataset.



End of Study: final database used for publication

basic requirements are first imposed on the dataset

Remove obviously identifying information and mask study subject numbers

absolute dates are considered identifying, so they must be masked

Convert all dates to be a day integer that is relative to a defined day0

very large studies may be allowed to include information about location

In most cases, remove references to siteID or location

outliers:
What information risks identification?
How unusual must an outlier be to warrant data truncation?

Remove outliers that realistically risk identification of an individual

Consider other elements that might lead to patient identification.

Remove or modify free-text entries, audit trail elements, etc.

Define sensitive information in the DB; how is this definition affected by the nature of the study? (example: AIDS research)

Remove or modify free-text entries, audit trail elements, etc.

ARDSNET EXAMPLE

The NHLBI ARDS Clinical Trials Network

Data Elements were Removed:

- Patient initials
- Site coordinator names (user names)
- Site identifiers
- Some free-text fields (example: M.D. names)
- Screened / non-enrolled patient data
- Fields that were relevant only with reference to the data entry software
- Date/time of data entry (audit info.)

Data Elements were Modified:

- Ages over 89 years converted to "89"
- Race/ethnicity information
- Height and Weight: truncated to remove identifying extreme values
- Subject IDs recoded to prevent site identification
- Calendar dates recoded to be relative to each patient's study day 0

CONSIDERATIONS

Many studies will need to strike a balance between data utility and patient confidentiality concerns. Examples include:

- AIDS researchers who are interested in examining data from other fields of research may find less accessible information for tying into their topic. This is due to the designation of HIV-positive status as sensitive information.
- Location can be relevant to interpretation of data, such as in elevation-adjusted P/F ratio.
- Studies whose data depend upon pedigree/heredity information have to decide how best to share their data with other researchers, given confidentiality concerns.

DOCUMENTATION

Documentation provided with the dataset should address both the details of the study and any potential confusion that could result from changes that were made.

Documentation should include:

- Description of all changes made
- Study protocol documents
- Case Report forms & instructions
- Dataset description

RECOMMENDATIONS

The ARDSNet experience has highlighted some of the elements to be mindful of during any effort to produce a "deidentified" distribution dataset. Many of these are described above. Others include:

- Audit trails are required in electronic datasets, but access is restricted to the original researchers and sponsor.
- Before a study even begins, design of the data collection tools can prevent problematic data at the end.
  Example: well-chosen pick-lists as a good alternative to free-text; benefits to analysis and data deidentification.

REFERENCES

NHLBI Policy for Distribution of Data. NHLBI. 25 October 2002. <http://www.nhlbi.nih.gov/resources/deca/policy.htm>.

ARDS Network SOPs - Data Management SOP on Limited Access Datasets (internal document). 28 July 2003, 7 June 2005.

Please visit the MGH Biostatistics website at:

http://biostatistics.mgh.harvard.edu