PRODUCTION OF DEIDENTIFIED DATASETS AT THE CONCLUSION OF HUMAN SUBJECTS TRIALS

Korpak A, Schoenfeld D
MGH Biostatistics Center

Abstract

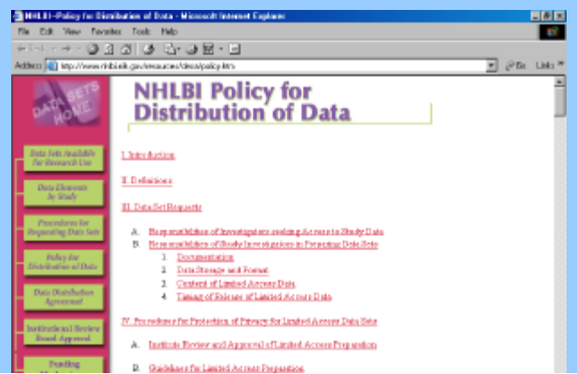
At the conclusion of a study, researchers often have an obligation to publish their data, but simultaneously are required to preserve confidentiality of human subjects. For instance, NIH research contracts require distribution of data after publication, and call upon the researchers to provide a deidentified 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.

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.



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.

End of Study: final database used for publication

basic requirements

Remove obviously identifying information and mask study subject numbers

mask dates

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

very large studies may be allowed to include information on location

In most cases, remove references to siteID or location

outliers:
What information risks identification?
How unusual is an outlier?

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 affected by the nature of the study? (example: AIDS research)

Remove sensitive information: AIDS, drug abuse, history of violence...

Provide documentation of process

ARDSNet example

Data Elements were Removed:
- Patient initials
- Site coordinator names (user names)
- Site identifiers
- Some free-text fields (ex: 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:
- AIDS researchers may be interested in examining data from other fields, and find less accessible information for tying into their topic.
- Location can be relevant to interpretation of data, such as in elevation-adjusted P/F ratio.
- Studies whose data depend upon pedigree/heredity information.

http://biostatistics.mgh.harvard.edu

The NHLBI ARDS Clinical Trials Network

Documentation

Documentation provided with dataset should address both the details of the study and any potential confusion that could result from changes 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:
- Balance between data utility and the need to protect patient confidentiality (examples given in Considerations section at left).
- 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. 28 July 2003, 7 June 2005.