PRODUCTION OF DEIDENTIFIED DATASETS AT THE CONCLUSION OF HUMAN SUBJECTS TRIALS

Korpak A, Schoenfeld D
MGH Biostatistics Center

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.