SDTM De-identification for CTN0009 11/20/2007

The CTN0009 data was collected using some form of dynamic data management which led to the following data processing rules to make the data more useful and compliant with the SDTM:

1. In DM, study termination date was taken as the first available study termination date from the study termination form.
2. In DM, the last available demographic page visit date for a subject was used to populate DM.
3. In AE, absolute duplicate records were removed (excluding clinician signature date).
4. In SC, the last available demographic page visit date, the last enrollment page visit date, the last randomization page visit date, and the last inclusion/exclusion page visit date for a subject were used to populate SC.
5. In IE, the last available inclusion/exclusion page visit date for a subject was used to populate IE.
6. In RP, there are two redundant records that were removed.

The CTN0009 SDTM database has been de-identified according to the HIPAA Safe Harbor rules. Those rules are:

> The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:
> (A) Names;
> (B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial current publicly available data from the Bureau of the Census:
>
> > (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
> > (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000.
>
> (C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
> (D) Telephone numbers;
> (E) Fax numbers;
> (F) Electronic mail addresses;
> (G) Social security numbers;
> (H) Medical record numbers;
> (I) Health plan beneficiary numbers;
> (J) Account numbers;
> (K) Certificate/license numbers;
> (L) Vehicle identifiers and serial numbers, including license plate numbers;
> (M) Device identifiers and serial numbers;
> (N) Web Universal Resource Locators (URLs);
> (O) Internet Protocol (IP) address numbers;
> (P) Biometric identifiers, including finger and voice prints;
> (Q) Full face photographic images and any comparable images; and
> (R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section; and
> The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

The following rules were applied to the CTN0009 SDTM database to de-identify them:

1. All (**DTC) dates converted to year only.
2. All USUBJIDs converted to a random id number called DEIDNUM.
3. Age verified to be less that 89.
4. Specific file by file de-identification changes follow by domain.

AE:

- Many “Other AE action taken” responses had dates that had the month and day removed.
- An adverse event verbatim term had to have a month and day removed.

SUPPAE:

Numerous dates present in AE description had to have month and day removed.

BR: Clean

CM:

Some indications had dates that had to have month and day removed.

SUPPCM:

Some results were in YYYY-MM-DD format, so the MM-DD was removed.

CO:

All records are potentially identifying info. This dataset should not be given as de-identified data.

DA:

Some drug detail results had dates that had to have month and day removed.

DM:

- Make SITEID blank.
- Recode several (ASIAN, AMERICAN INDIAN OR ALASKAN NATIVE, NATIVE HAWAIIAN OR PACIFIC ISLANDER , OTHER (SPECIFY)) low count RACEs to “OTHER”.

DS: Clean.

SUPPDS:

Clean.

EX:

Some exposure adjustment details had dates that had to have month and day removed.

SUPPEX:

Clean.

IE: Clean.

LB: Clean.

SUPPLB:

Some results were in YYYY-MM-DD format, so the MM-DD was removed.

MH:

Some medical history terms had dates that had to have month and day removed.

SUPPMH:

Some medical history detail had dates that had to have month and day removed.

QS:

- Some QSORRES and QSSTRESC = DD/MM/YY so made to be xx/xx/YY.
- Some QSORRES contained unique race categories which have been recoded to OTHER.
- Some QSORRES contained unique employment data which have been recoded to generic equivalents.

RP: Clean.

SC:

- Dropped NODE and RANDOMIZATION NUMBER.
- Dropped all RACE based SCTESTs and results that could identify someone based on a specific race entry. RACE should be gotten from DM.RACE.

SUPPSC:

All records are potentially identifying info. This dataset should not be given as de-identified data.

SE: Clean.

SU: Clean.

SV: Clean.

TA, TE, TI, TS, and TV are all trial metadata files and have no identifying information

TU: Clean.

SUPPTU:

Clean.

VS: Clean.

SUPPVS:

Some results were in YYYY-MM-DD format, so the MM-DD was removed.