CSP-1026 CRF Level Data De-Identification Notes

And

Database Descriptions

Aug 30, 2015

CSP-1026 Data De-Identification Notes

The Case Report Form (CRF)-level de-identified data sets were prepared by reviewing the CRFs and the data dictionary to identify variables that contained, or could potentially contain, identifying information, such as identifiers, dates and free text fields. These variables were programmatically removed or modified to produce de-identified datasets.

- Patient ID: All patient IDs were transformed to factless key values bearing no relationship to the collected study data.
- Site ID: All site IDs were transformed to factless key values bearing no relationship to the collected study data.
- Dates: All dates except date of birth were converted to "days on study", where the date of randomization = day 0. A negative value for days on study refers to an event that occurred prior to randomization (e.g., date of birth, consent date) and a positive value for days on study refers to an event that occurred after randomization (e.g. treatment-phase visit and event dates).
    - Date of Birth: Dates of birth were converted to age values, computed as of randomization date.
    - Note that patients who failed screening will have no computed studyday values in date fields.
    - In cases where the date was stored as 3 distinct columns (month, day, year), these component value were converted to a date value before computing days on study; the resulting variable was named following the naming pattern of the 3 component variable names, but substituting the suffix DT.
- Initials / Names: Any field designed to record a person's initials or name was erased.
- Other Identifiers: Sample identifiers and Drug lot numbers were erased.
- Text Fields: Most free text fields long enough to contain dates or narrative were emptied, including
    - Comment fields.
    - Responses derived from "Specify", "Describe", "Reason", or "Other". For example, other race, other termination reason, reason study drug not given, abnormalities on ECG and lab forms.
    - Narratives regarding interaction with patient, health conditions, treatment progress or follow-up.
    - Medical history and physical exam narrative text.
    - Type of work or job description; school program name and major; narrative descriptions of living situation, illnesses, concerns, drug/alcohol use, feelings/interests.
- Text Field De-identification Exceptions:
    - AE and medical condition terms were *retained* in the database, but any detailed descriptions in a separate field providing additional information about adverse events were erased, such as relevant labs, medical history, drug/alcohol use description, probable cause of death.
    - Concomitant medication names and indications were *retained* in the database.
- For a complete listing of the erased text fields by CRF, refer to the "CSP1026 Nulled Values" csv file.

See the "CSP1026 Dictionary" Excel workbook file for descriptions of the data, organized into tabbed worksheets for each respective CRF/dataset. The columns that have been nulled for de-identification purposes are indicated by a 'Y' in the Nulled column.

Known data issues: Two date variables in the final SAS database contain unexpected values.

- CBT dataset – all the date values in variable DATESES3 show a year of 1920.
- ECG dataset – variable READDATE includes values of 23 Dec 1959* that translate into extremely large negative study-day values.

* Note: It is suspected that this value occurs where a -9 was present as this value would translate to the observed date and is used in other columns of this table to indicate ‘Not Available’