CSP-1033 CRF Level Data De-Identification Notes

And

Database Descriptions

July 9, 2024

CSP-1033 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.

- Subject: All subject IDs were transformed to factless key values bearing no relationship to the collected study data. Multiple subject-related ID fields were present in each of the study tables, however, only one of them served as a sufficient linkage across all of the study datasets. Therefore, these additional subject fields were nulled for security and to avoid adding confusion.
- Site ID: All site ID fields were hashed.
- 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. consent date, conmed initiated during screening) 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 values were nulled, as in all cases there existed a complete date field alongside them that could more easily be converted using the method described above.
    - Partial date sets of variables, e.g. just a Month and Year, may be erased if they refer to potentially identifying milestones, especially in a small patient pool.
- 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; narrative descriptions of living situation, illnesses, concerns, drug/alcohol use, feelings/interests.
- Raw Fields: The study datasets contained a number of 'Raw' labeled fields which were duplicates of other fields already present within the dataset. To simplify the data de-identification process, these Raw-labeled fields were nulled.
- Text Field De-identification Exceptions:
    - AE and medical condition terms were *retained* in the database, but 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, refer to the “CSP-1033-NULLED-COLUMNS” Excel workbook file.