CENIC-P1S1 Data De-Identification Notes

And

Database Descriptions

January 15, 2020

CENIC-P1S1 Data De-Identification Notes

The data presented here in de-identified form are analysis datasets, rather than as-collected case report form data. They contain data only for those subjects who were enrolled in the study. Datasets included are:

- Primary_Baseline
- Secondary_Baseline
- Primary_PostRand
- Secondary_PostRand
- FollowUp_30Day
- Adverse_Event

The two Baseline datasets and the Followup contain one row per subject, while the two PostRand datasets contain a row for each visit within subject. (Values for the visit variable are described in the Data Scoring Manual under “Identifiers”.) The Adverse_Event dataset contains one row per event.

Note on missing data: Some variables include coding for missing data, using numeric values in the 999x range, e.g. 9998. Explanations of these values are presented by variable or variable group in the Data Scoring Manual.

These de-identified data sets were prepared by reviewing the original analysis datasets along with the data manual to identify variables that contained, or could potentially contain, identifying information, such as personal identifiers, dates and free text narratives. These variables were programmatically removed or modified to produce de-identified datasets.

- Subject ID: Each subject ID was transformed to a factless key value bearing no relationship to the collected study data.
- Site ID: Each site ID, originally coded within the Subject ID values, was transformed to a factless key values bearing no relationship to the collected study data, and placed into a newly created and distinct Site ID variable.
- Dates: All dates 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 or screening dates), while a positive value refers to an event that occurred after randomization (e.g. treatment-phase visit and event dates). Coded-missing values in date variables were changed to true missing rather than a potentially misleading translation into days on study.
- Initials / Names: Any field designed to record a person’s initials or name was erased.
- Other Identifiers: Sample or product barcodes and identifiers were erased.
- Text Fields: In general, text fields long enough to contain dates or narrative were emptied, including
    - Responses derived from “Specify”, “Describe”, “Explain”, or “Other”.
    - Narratives regarding interaction with patient, health conditions, treatment progress or follow-up.
    - Medical history and physical exam narrative text.
    - Job descriptions and narrative descriptions of living situation, drug/alcohol use, or feelings/interests.
- Text Field De-identification Exceptions:
    - Medical event and condition terms, if non-narrative, were *retained* in the database, but detailed descriptions providing additional information about adverse events were erased, such as relevant labs, medical history, drug/alcohol use description, probable cause of death.

  - Medication names and indications, if non-narrative, were *retained* in the database.
- For a complete listing of the erased text fields, refer to the “CENIC-P1S1 Nulled Values” csv file.

See the “Data Scoring Manual” file for descriptions of the data, organized by collection instrument or form.