Considering De-Identification?

Legacy Data

Kymberly Lee
16-Jul-2015

AstraZeneca

Introduction

This presentation provides an overview of Clinical data sharing, clinical data privacy, and clinical transparency.

Discuss the nuances and experiences in working with Legacy data to complete the de-identification process in order to preserve data privacy while maintaining scientific importance.

Data Sharing, Privacy and Transparency

- Clinical Data Sharing is the ability to share data
- Clinical data privacy encompasses privacy laws, HIPAA, etc.
- Clinical Data Transparency determines the levels of data de-identification to protect the patient's personal data

Unique Approach to De-identification
Legacy Data versus SDTM/CDISC Data

- Differences/Challenges Legacy Data versus SDTM/CDISC
- Legacy data not required to be converted to be CDISC compliant. However, the de-identified data file should be structurally ready for SAS transport file conversion when completed
- If the raw data source has been converted already to CDISC SDTM complaint and the analysis file are still legacy data
- This presentation will demonstrate how to convert legacy data points only. Please reference public documents on CDISC compliant data.

Ensuring Consistency

- Consistency/Traceability: Shorten Label/Variable names
- Traceability:
  - De-identified raw/source data/analysis data
  - How to ensure traceability of the patient going from de-identified raw/source data to the de-identified analysis data
  - Ensure patient consistency from de-identified double-blinded phase trials and open or extension phase de- identified data

Considerations to Simplify the Approach

If not using an industry tool, a viable approach would be to create a separate de-identified linking file based from unique patient/subject identification number, site/center number, demographic information, geographic location, and any other personal data points.

- Ensure patient/subject identification is completely re-randomized or scrambled.

- Ways to achieve this process:
  1. New Patient identifier
  2. Site/Center Number (where applicable)
  3. Age Grouping
  4. Race category re-randomized or scrambled

Considerations, cont'd

- Keeping consistency within patient level information across raw/source and analysis data files as well as open/extension studies is crucial

- Processes Applied:
    1. Merging Datasets with de-identified patient level data
    2. Dropping Original Patient Information
    3. Renaming variables/label
    4. What to do with the linking file information

Domain Considerations: General Points

- Date variables: Why you should consider calculating a relative day variable within each domain
  (i.e. Visit dates or assessment dates, time to event dates, adverse event dates, etc)

- Timing variables expected within protocol design
  (i.e. Labs, Exposure, Vitals, Pharmacokinetics, etc)

Different data or analysis domains require special attention to ensure patient's identity is protected especially when looking across data collections, medical coding, and types of assessments collected. Let's review a few….

Data Considerations cont'd
Special Attention Data Files

- Adverse events, medical and disease history, concomitant medication/procedures, prior and concomitant/subsequent therapy should be closely scrutinized

  What to do when older versions of medical dictionaries are used:

  Does Requester want to update the coding
  Patient safety/Scientific relevance
  Redaction

- Laboratory and ECG Data:
  What variables to de identify and why

- Vital Signs Data (including weight and height):
  What variables and why? Should therapeutic area or analyses be considered?

Data Considerations cont'd
Exposure data and Other Analyses

- Exposure data:

  Does the data file consist of other important patient level identities?

  What variables should be considered? Should they be de-identified or removed?

- Consider all therapeutic areas and data used for efficacy. (i.e. tumor locations, genomic data, translational medicines, asthma equipment, cardiovascular digital equipment, ambulatory serial numbers,etc.)

- Ensure any data pieces that may provide personal identity of a patient to any knowledgeable medical/technical employee.

Documentation

Upon completion of de-identifying the data, documentation is important for internal purposes as well as safety of patient's personal data

Recommend the following:

1. Complete documentation to explain which variables were de-identified and to what level of de-identification (i.e. explain the process). See additional slides for options of recording.
2. Check data for conformance and traceability of variable/label name from raw/source data to analysis data
3. Complete SAS transport files
4. Provide to requestor as specified per Data Sponsor

De-identification Reviewer's Guide

Contents

Introduction

1.1 Purpose

This document provides context for the sole purpose of de-identification of Individual Patient-Level Data (ILDP) as agreed upon and specified within ***Data Sponsor*** as specified in ***current process*** governances and Clinical Data Sharing Agreement. In addition, this document provides a summary of the data points included as well how IPLD data information was de-identified in a conformance with the original data findings and written agreement.

Acronyms

Acronym	Translation
IPLD	Individual Patient-Level Data
HIPAA	Health Insurance Portability and Accountability Act

Current process Name Version and Compliances

Standard or Dictionary	Versions Used
Current Process name Version	Final 0.0/ Month/Year

Protocol Description

2.1 Protocol Number and Title

Protocol Number:
Protocol Title:
Protocol Versions/Date:

2.2 Data files included in Individual Patient-Level Data Delivery

Raw/Oracle Datasets?	Yes
SDTM Datasets?	No
Analysis/ADAM Datasets?	Yes

Subject Data Description

3.1 Overview

Date of the Clinical Data Sharing Agreement:

Date: DD-MM-YYYY

Were data de-identified as requested based on Data Sharing Agreement? Yes

Specific the date of the original data files being used for this request:

Date: DD-MM-YYYY

Were the Raw/Oracle/SDTM datasets used as sources for the analysis datasets? Yes

In what data format were the original data stored?

In what data format will the de-identification datasets be delivered?

How were the de-identified data sets transferred?

3.2 De-Identified Source Data Domains

** This can also be the LEGACY Raw dataset names if per request ** For illusions, these are SDTM Domain names.

Dataset – Dataset Label	Efficacy	Safety	Other	SUPP-	Observation Class
AE – Adverse Events		X			Events
CE - Clinical Events		X			Events
CM – Concomitant Medications		X			Interventions
CO – Comments			X		Special Purpose
DA – Drug Accountability			X		Findings
DM – Demographics			X		Special Purpose
DS – Disposition			X		Events
EG – ECG Test Results		X			Findings

SC - Subject Characteristics			X		Findings
SV – Subject Visits			X		Special Purpose
VS – Vital Signs		X			Findings

3.2.1 Data set name – Adverse Events

Variable Name	Description of Process
AETXT	Explain what procedures were applied. Verbatim AE Text was dropped.
AEDY	Added; calculated as (give algorithm applied)

3.2.2 Data set Name – Demographics

Variable Name	Description of Process
RACE	Subjects identified as "Native Hawaiian/Pacific American" = 8; "Jamacian"= 5. - Subjects were re-mapped to Other.
SBJNBR	Unique sequential numeric count applied. Then de-identified using randnum sas generator.
RANDUM	Dropped per Data Sponsor guidance (or list current sponsor process)

3.2.3 Data set Name – Concomitant Medications

Data Conformance Summary

4.1 Data Issues

Use this section to add areas of concerns with the data that were not fixable (i.e. truncation of filenames, etc).

- Variables name ??? and description. (i.e. Lab values missing for all lab test = ALT in original data)
-

Dataset	Description of Issue	Severe	Explanation
AE	AE is Serious but no qualifiers set to 'Y'	Y/N	
PC	Mismatch of VISIT and VISITNUM compare to TV domain		

4.2 Data Issues Summary

- How were the de-identified data sets transferred?

- Final Data Locations:

Dataset	Path Location
Raw	
SDTM	
ADAM or	

Conclusion

In conclusion, what was covered when de-identifying Legacy data:

1. Review the request and Therapeutic area requested.
2. Determine if raw/source and/or analysis files will be needed
3. Decide the domains requested and types of data requested
4. Review and ensure consistency and traceability of variable names and lengths within the data request.
5. Complete the link file with intentions of ensuring patient level de-identified data points have the capability to carry from raw/ source data to analysis and/or open-label extension studies.
6. Review whether protocol or analysis time variables
7. Check all treatment assignments and randomizations
8. Can the data be re-identified by any medical employee/officers?
9. Check all available data points for personal data : de-identifying or redact all data (including comments).
10. Document all process applied in Data Sponsor provided format.
11. Validate the process thoroughly before delivery.

Contact Information and Presentation Disclosure

All contents of this presentation are the sole expressions and experiences of the presenter.

Contact information:
Kymberly Lee
AstraZeneca, Gaithersburg Campus
(301) 398-0715