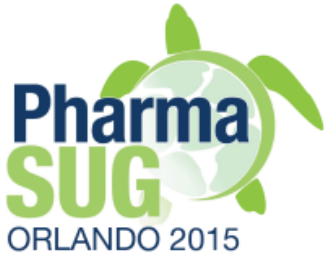


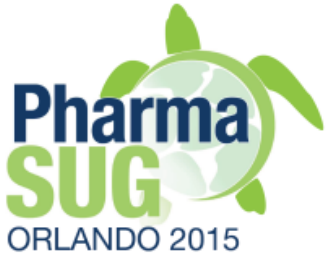
The Most Common Issues in Submission Data

Sergiy Sirichenko, Pinnacle 21
Max Kanevsky, Pinnacle 21

PharmaSUG 2015
Paper #SS06

Introduction

Need for High Quality Standardized Data



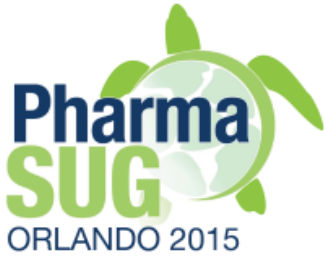
- FDA finalized requirements for submitting data in standardized format
- Automated analytics and data-driven tools allow to perform review more efficiently
- FDA's definition for high quality data
  - *Compliant* means the data confirms to applicable data standards
  - *Useful* means the ability of data to support the intended use

FDA DataFit

- To ensure "High Quality Data", FDA launched the DataFit project
  - OpenCDISC Enterprise software
  - Detailed assessment of submitted data performed very early in the review process
  - Based on intended use requirements
  - Helps to understand if there are any data-quality issues that could prevent reviewers from doing their job
  - Performed as part of JumpStart service that provides FDA review team with additional exploratory data analyses

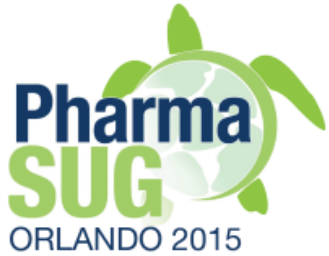
Metadata Issues

Define.xml



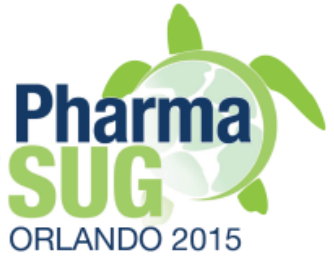
- The most important part of electronic dataset submission for regulatory (FDA, 2015, p.16)
- Most often noted to be deficient

Define.xml v1.0 is outdated



- 10 years old
- Cannot handle Value Level metadata
  - No reference to Variable it applies to
  - Important for ADaM data
- No reference to standard CT (NCI Codes)
- Limited structural consistency
  - Origin example: CRF → Pages, Derived → Method
- Define.xml v2.0 is robust enough to handle review needs

Incorrect or missing codelists

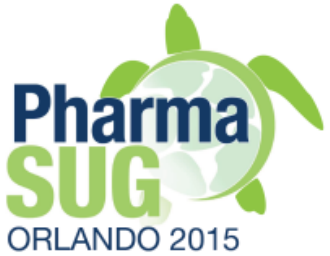


- Missing codelists for study specific elements
  - --CAT, --SCAT, EPOCH, etc.
- Missing codelists for Value Level
  - In SUPPQUAL domains
- Codelists for variables collected as a free text
- Collapsed codelists for multiple variables across domains
  - Single (UNIT) codelists for all --DOSU,--ORRESU, --STRESU variables
  - Codelist should be variable specific
  - Confusion between variable codelist and Control Terminology applied for variable

Missing, unclear or invalid Computational Methods

- All “Derived” variables must have clear and detailed description of computational Method
- Missing Method for study specific variables
  - EPOCH, SESTDTC, RFPENDTC
- Reference to non-available information
  - EDC variables, look-up conversion tables
  - If “Yes” then CMENRF is “ONGOING”’
    - What is “Yes”? What variable or CRF page does it refer?

Missing descriptions for study specific variables



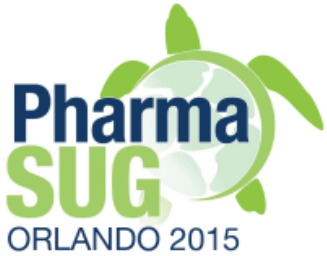
- --SPID, --GRPID, …
- These variable are often Key Variables in domains and responsible for “duplicate” records
- No description, no understanding of study data

- The biggest value of define.xml is to provide a description of study specific data elements!

Incorrect Origin

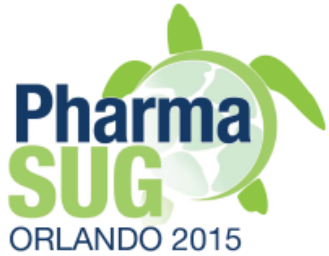
- Due to lack of understanding of define.xml
- Inconsistency in attributes
  - E.g., Origin=CRF with detailed Computational Method
- Confusion between Protocol, Assigned and Derived
- Education is needed

Incorrect Value Level metadata



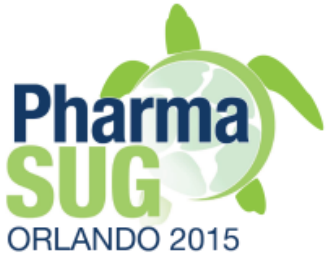
- Value Level metadata populated as a copy of attributes from variable
  - SUPP--.QVAL example
    - SUPPAE.QVAL.AETRTEM has Length 1 char, not 200 chars
- Value Level should be considered as a new variable with independent attributes from host Variable

Issues in Annotated CRF



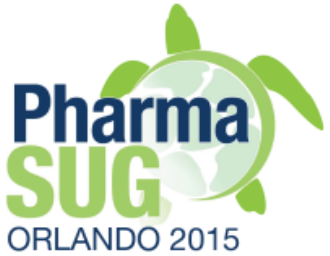
- Missing or incorrect annotations
- Annotations to EDC database instead of SDTM variables
- Annotations as highlighted text instead of PDF Annotations

Reviewer's Guide



- Provides high-level summary and additional context for the submission data package
- Rapid adoption by the industry
- High popularity with FDA reviewers

Reviewer's Guide issues

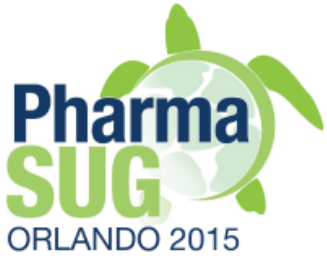


- Not following recommended structure
- Missing or meaningless explanations for data conformance issues
  - *"Expected result", "This is our common practice", "As received from our vendor", "Sponsor decided not to fix", "We did not collect nor derive this data element", …*
- Issues explanations that show incorrect interpretation of CDISC standards and FDA requirements
  - Issue: "*Date is after RFPENDTC*", 10-60% records
  - Explanation: "… *set to the latest DSSTDTC in DS domain where DSCAT='DISPOSITION EVENT'*"

PharmaSUG ORLANDO 2015

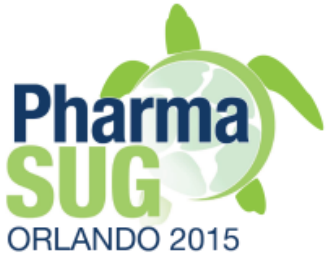
Noncompliance with FDA Business Rules

FDA specific requirements



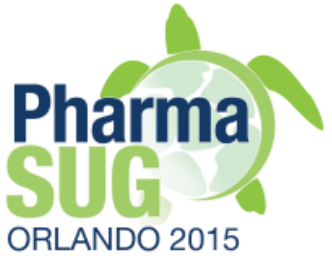
- May 2011
  - "CDER Common Data Standards Issue Document"
- November 2014
  - "FDA Business Rules for SDTM data"
  - "FDA Business Rules for SEND data"
  - December 2014
    - OpenCDISC executable version
- December 2014
  - "Study Data Technical Conformance Guide"

Issues with FDA specific data elements



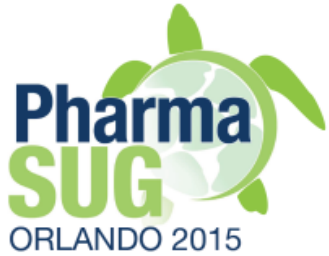
- Missing EPOCH variable
- Missing AE Seriousness Criteria
- Missing AE Treatment Emergent Flag in SUPPAE
- Missing Study Day variables
- Missing Trial Design domains

Inconsistency in Death data



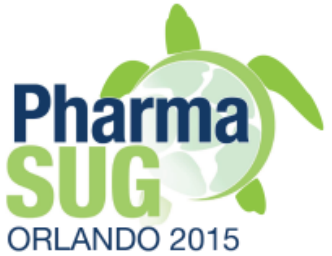
- Subjects death info is very important
- FDA asks to populate subject death as a last record in DS domain
- Common death reporting inconsistencies:
  - Inconsistency between DM and DS death information
  - Missing death dates in DM and DS domains
  - Invalid coding of DS terms
    - DSTERM=”Death”
    - DSDECOD=”ADVERSE EVENT” or “OTHER” instead of “DEATH”

Inconsistency in Death data



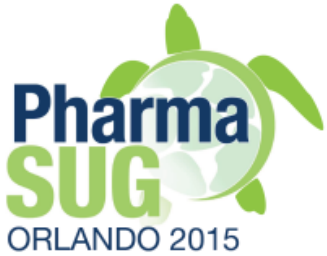
- Death information in DM and DS vs. other domains
  - Example 1. Subject death is not listed in DM and DS, but subject has
    - FATAL Adverse Event
    - Comment record like “After subject death …”
    - Protocol deviation record like “Due to subject death …”
    - Date of Autopsy record in SUPPQUAL domain
  - Example 2. Inconsistency in subject death date in DS domain and FATAL adverse event end date

Missing Disposition Dates



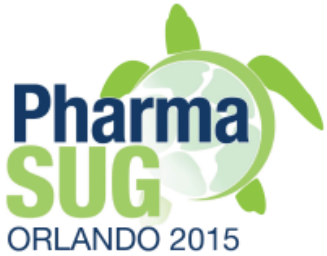
- Analysis of Disposition data is very important
  - E.g., high rate of early termination for study treatment may be due to lack of efficacy or safety issues
- There is a lack of formal regulatory requirements on collection of DS dates
- Dates which should be collected
  - Informed Consent
  - Study/Treatment Termination
  - Last F/U Contact with Subject
- Utilize Risk Based Data Verification

Duplicate records



- Multiple test result or event records on the same time point
- May be due to different reasons
  - "Pure duplicates". Difference in --SEQ only
  - Different results for the same time-point
  - Same Original Result, but different Original Units
  - One record with actual result, another NOT DONE
  - The only difference in --SPID, which is not described in define.xml
- Duplicate records in SUPPQUAL domains
  - Multiple records with the same USUBJID, IDVAR, IDVARVAL and QNAM is a data integrity issue!

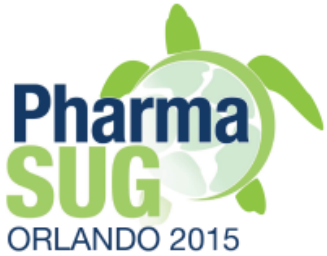
Incorrect RACE "OTHER"



- Most commonly used extended terms for RACE CT: "MULTIPLE", "UNKNOWN", "OTHER"
- RACE="OTHER"
  - "Caucasian" (should be mapped to "WHITE")
  - "Hispanic" (it's Ethnicity, not Race)
  - "United Kingdom" (it's Nationality, not Race)
  - "Not Reported" (should be mapped as "UNKNOWN")
- RACE="MULTIPLE"
  - "White and Hispanic" (split to Race and Ethnicity)
- Anything collected as a free text requires data cleaning

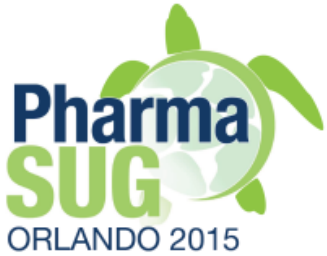
Programming and Mapping Errors

Missing values for Required variables



- Structural data consistency
- Examples
  - HOTERM
  - EXTRT
  - LBTEST
- Ensure that data is collected
- Consider using special terms
  - “Unknown”
  - “All Labs”

Data consistency issues



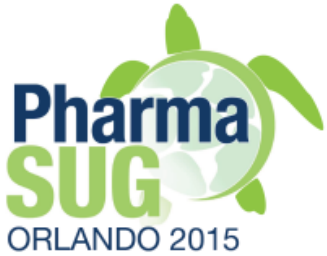
- Inconsistency between Trial Visits (TV) and Subject Visits (SV) data vs. other domains
- Inconsistent Standard Units
- Inconsistent terms within CT
  - EGTESTCD=”QTC”
  - EGTEST=”QT Uncorrected”
  - NCI Code must be the same
- RELREC or SUPPQUAL domains with reference to non-existing records

Other issues

- --STRF and --ENRF variables used for subjects without RFSTDTC and RFENDTC
- Comment in SUPPQUAL domains
- Leading space and special characters like <LF> and <CR>
- Study Day imputation for partially missing dates
- Incorrect calculation of Study Day

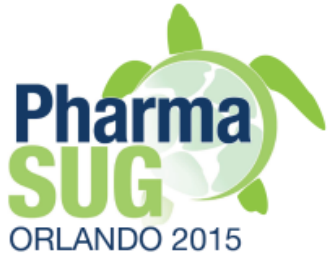
Control Terminology Issues

Control Terminology



- Usage of standard Control Terminology (CT) is required for regulatory submission
- Standard tools rely on CT
- Standardized data has
  - Standard structure (SDTM, SEND, ADaM)
  - Standardized content (CDISC CT, MedDRA, etc.)

Common issues with CT



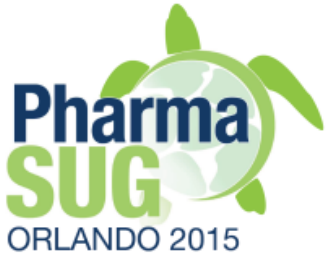
- Ignoring existing terms in extensible codelists
  - New terms can only be added if they are not already represented in standard codelist
- Modification of standard terms by conversion into Upper Case or misspelling
- Not following new CDISC CT codelists
  - SDTM and Terminology are separate standards
  - Terminology is assigned in SDTM IG
  - Published IGs are not updating with new CT codelists
  - Monitor new versions of CT for new codelist

CT issues due to data collection

- Data collection as free text leads to problems with implementation of standard terminology
  - Mapping issues
  - Invalid data
    - E.g., CMDOSU as “000”, “1 Patch every four days”, “Table”, etc.
- Invalid data collection design
  - E.g., AEACN
    - collected as action taken for AE, rather than with study drug
      - “Hospitalization”, “Additional Medication”
    - collected as “DOSE MODIFIED”, rather than “DOSE INCREASED” or “DOSE REDUCED”

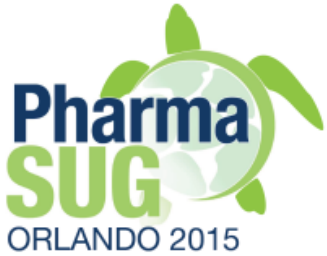
Summary

Summary

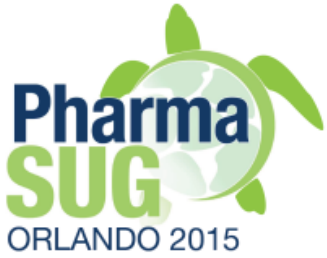


- High quality data in standardized format is required for regulatory submissions
- The industry's pace of standards adoption has greatly accelerated in the last few years
- Be aware and be prepared

References



- http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm
- http://cdisc.org/standards-and-implementations
- http://www.opencdisc.org
- http://www.pinnacle21.net



Name: Sergiy Sirichenko
Organization: Pinnacle 21
Address: 531 Plymouth Road, Suite #508
City, State ZIP: Plymouth Meeting, PA 19462
Work Phone: 908.781.2342
E-mail: ssirichenko@pinnacle21.net
Web: pinnacle21.net