Data Cleaning

Module 10 Topic 9

Cleaning Data

- Data cleaning or validation is a collection of activities used to assure validity & accuracy of data
- Logical & Statistical checks to detect impossible values due to
  - Data entry errors
  - Coding
  - Inconsistent data

Who Cleans Database?

Data Management Plan through SOPs clearly defines tasks & responsibilities involved in database cleaning

Cleaning Data

Activities include:

- Manual review of data
- Computer checks/validations designed to identify
  - Inaccurate or invalid data using ranges
  - Completeness
  - Protocol violations
  - Consistency checks
  - Aggregate descriptive statistics to detect strange patterns in data

Cody's definition of data cleaning

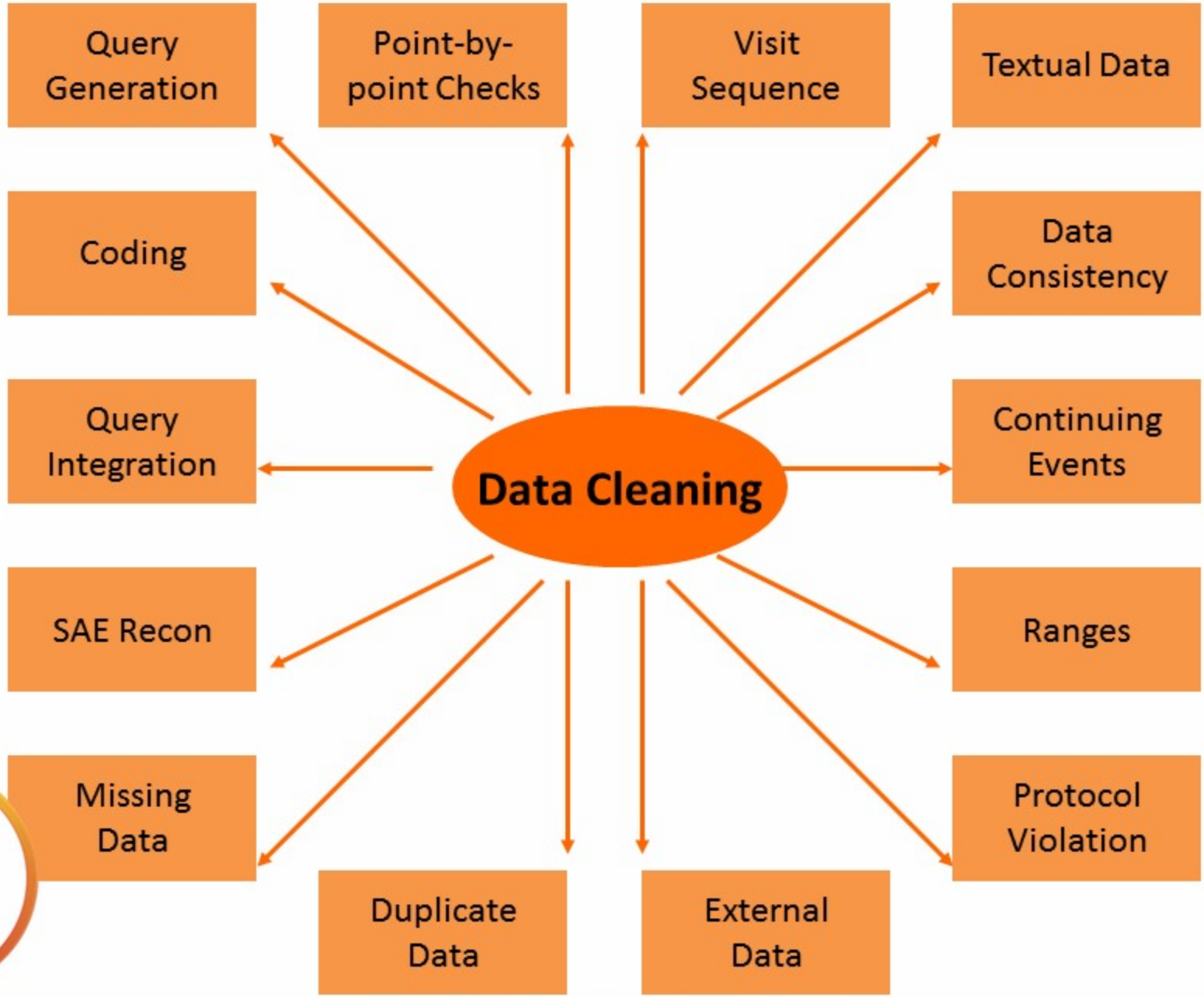
- Making sure that raw data were accurately entered into a computer-readable file
- Checking that character variables contain only valid values
- Checking that numeric values are within predetermined ranges
- Checking for & eliminating duplicate data entries
- Checking if there are missing values for variables where complete data are necessary

Cody's definition of data cleaning (contd)

- Checking for uniqueness of certain values, such as subject IDs
- Checking for invalid data values & invalid date sequences
- Verifying that complex multi-file (or cross panel) rules have been followed. For e g., if an AE of type X occurs, other data such as concomitant medications or procedures might be expected

Clean Data Checklist

- Refers to a list of checks to be performed by data management while cleaning database
- Checklist is developed & customized as per client specifications
- Provides list of checks to be performed both on
  - Ongoing/periodic basis
  - Towards end of study
- Strict adherence to checklist prevents missing out on any of critical activities

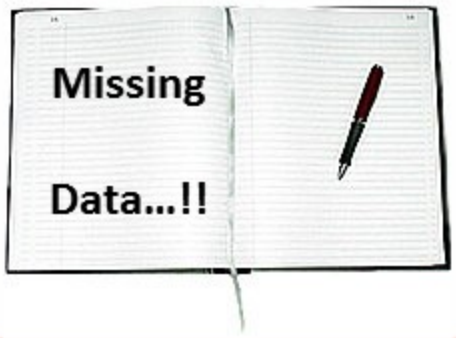

Query Generation
Point-by-point Checks
Visit Sequence
Textual Data
Coding
Data Consistency
Query Integration
Data Cleaning
Continuing Events
SAE Recon
Ranges
Missing Data
Protocol Violation
Duplicate Data
External Data

Point-by-Point Checks

- Refers to cross checking between CRF & database for every data point
- Constitutes a “second-check” apart from data entry
- Incorrect entries/entries missed out by Data Entry are corrected during cleaning
- Special emphasis to be given for
  - Dates
  - Numerical values
  - Header information (including indexing)

Missing Data Checks

- Missing responses to be queried for, unless indicated by investigator as
  - Not done
  - Not available
  - Not applicable
- Validations to be programmed to flag missing field discrepancies



Missing Page Checks

- Expected pages identified during setup of studies
- Tracking reports of missing pages to be maintained to identify
  - CRFs misrouted in-house
  - CRFs never sent from Investigator's site

Protocol Violation Checks

- Protocol adherence to be reviewed & violations, if any, to be queried
- Primary safety & efficacy endpoints to be reviewed, to ensure protocol compliance

Key Protocol Violations

- Inclusion & exclusion criteria adherence
  - Age
  - Concomitant medications/antibiotics
  - Medical condition
- Study drug dosing regimen adherence
- Study or drug termination specifications
- Switches in medications

Continuity of Data Checks

- Refers to checking continuity of events that occur
  - Across study
  - Across visits
- Includes
  - Adverse Events
  - Medications
  - Treatments/Procedures
- Overlapping Start/Stop Dates & Outcomes to be checked across visits

Continuity of Data Checks (contd)

Overlapping dates across visits:

- Scenario: Per protocol, AEs are to be recorded on Visits 1, 2 & 3
- "Headache" is recorded as follows:

Visit	Start Date	Stop Date	Outcome
1	01-Jan-2004	12-Jan-2004	Continues
2	01-Jan-2004	12-Jan-2004	Resolved
3	20-Jan-2004	20-Jan-2004	Resolved

Consistency Checks

- Designed to identify potential data errors by checking
  - Sequential order of dates
  - Corresponding events
  - Missing data (indicated as existing elsewhere)
- Involves cross checking between data points
  - Across CRFs
  - Within same CRF

Consistency Checks (contd)

Cross check across different CRFs:

- AE reported with action "concomitant medication" (AE Record)
- Ensure corresponding concomitant medication reported in appropriate timeframe (Concomitant Medication Record)

Event	Start Date	Stop Date	Outcome
Fever	13-Jun-2005	20-Jun-2005	Continues

Event	Start Date	Stop Date	Outcome
Paracetamol	13-Jun-2005	20-Jun-2005	Continues

Consistency Checks (contd)

Cross check within same CRF:

- 1st DCM: Report doses of antibiotics taken "before" intake of first dose of study drug
- 2nd DCM: Report doses of antibiotics taken "after" intake of first dose of study drug:

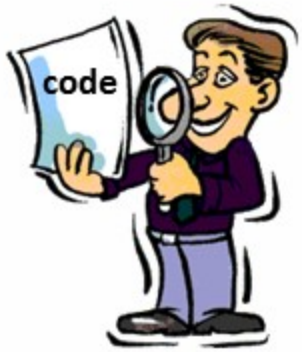
NOTE: First dose of study drug is taken on 15-May-2001

Antibiotic	Dose	Route	Start Date	Stop Date
Amoxicillin	5 mg	Oral	11-May-2001	14-May-2001

Antibiotic	Dose	Route	Start Date	Stop Date
Streptomycin	7 mg	IV	16-May-2001	17-May-2001

Coding Checks

- Textual or free text data collected & reported (AEs, medications) must be coded before they can be aggregated & used in summary analysis
- Coding consists of matching text collected on CRF to terms in a standard dictionary
- Items that cannot be matched, or coded without clarification from site
  - Ulcers, for example, require a location (gastric, duodenal, mouth, foot, etc.) to be coded



Range Checks

- Designed to identify
  - Statistical outliners
  - Values that are physiologically impossible
  - Values that are outside normal variation of population under study
- Ensure that appropriate range values are applied
  - For e.g., ranges for WBCs can be applied either in 'percentage' or in 'absolute'
- Ensure that appropriate ranges are applied depending on whether lab used is
  - Primary
  - Secondary

Range Checks (contd)

Cross check between Hematology record & AE record:

Hematology Test	Date	Result	Normal Range
WBC	05-Jan-2006	13,710 cells/µL/cu mm	4,300 - 10,800 cells/µL/cu

Event	Start Date	Stop Date	Outcome
Streptococcal Infection	04-Jan-2006	07-Jan-2006	Resolved

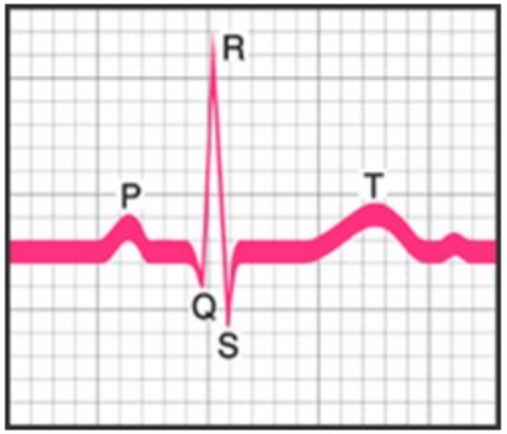
External Data Checks

- Ensure receipt of all required external data from centralized vendors:
  - Laboratory Data
  - Device Data (ECG, Bioimages)
- Missing e-data records to be tracked & requested from vendor on a periodic basis
- Missing data to be noted & corresponding values to be ‘re-loaded’ by vendor

External Data Checks (contd)

Examples of missing data/values:

- Missing collection time of blood sample
- Missing date of ECG
- Missing location of chest radiograph
- Missing systolic blood pressure
- Missing microbiological culture transmittal ID



External Data Checks (contd)

Examples of invalid data/values:

- Incorrect loading of visit number
- Incorrect loading of subject number
- Incorrect loading of date/time of collection

Duplicate Data Checks

- Refers to duplicate entries
  - Within a single CRF
  - Across similar CRFs
- Duplicate entries & duplicate records to be deleted per guideline specifications
- Examples:
  - Treatment ‘physiotherapy’ on ‘30-Aug-2001’ reported twice on either same Treatment Record or across two different Treatment Records

Duplicate Data Checks (contd)

- Examples:
  - Both Visit 4 & Visit 10 Blood Chemistry CRFs (with different collection dates) are updated with same values for all tests performed
  - Both 'primary' & 'additional' Medical History CRFs at Screening are reported with same details of abnormalities

Textual Data Checks

- All textual data to be proofread & checked for spelling errors
- Obvious mis-spellings to be corrected per Internal Correction (as specified by guidelines)
- Common examples of textual data:
  - Abnormalities/pre-existing conditions in Medical History record
  - Adverse Events
  - Medications/Antibiotics
  - Project & study-specific data

Visit Sequence Checks

- Sequence of visits should be reviewed & if out of sequence, should be either
  - Queried
  - Corrected per Internal Correction (as per guidelines)
- Either a single CRF or a group of CRFs could be out of sequence with that particular visit

Visit Sequence Checks (contd)

Visit	Visit date
1	01-Jan-2000
2	02-Jan-2000
3	03-Jan-2000
4	04-Jan-2000

Visit	Vitals Record Date of Vitals
1	01-Jan-2000
2	**03-Jan-2000**
3	**02-Jan-2000**
4	04-Jan-2000

Screening	
Record	**Visit date**
Demography	20-Feb-2006
Med. History	20-Feb-**2005**
Inclusion Criteria	20-Feb-2006
AE	20-Feb-2006

SAE Reconciliation Checks

- All SAEs reported on CRFs should be reconciled with those reported on SAE Reports & vice versa
- Communication to be maintained with
  - Sponsor
  - Clinical Scientist

Documents to be Followed

- Protocol
- Guidelines – General & Project-Specific
- SOPs
- Subject Flowcharts
- Clean Patient Check Lists
- Tracking Spreadsheets