Standards for Clinical Data Quality and Compliance Checks

Sunil Gupta, Senior CDISC/SAS Consultant
Gupta Programming

INTRODUCTION

Standards in clinical data quality and compliance checks involve confirming the validity of critical data variables as well as early identification of health risks. These critical data might need to be non-missing, consist only of valid values, be within a range, or be consistent with other variables.

Proactive steps need to be taken to identify, isolate and report clinical data issues using a system that is flexible, easy to update and facilitates good communication with the team to help resolve data quality problems. In general, a best practices system considers the following issues: accuracy of the data, completeness of the information, and consistency of the data across studies. The two main categories of clinical data issues may be grouped as incorrect and incomplete data. In general, incorrect data issues consist of unexpected raw values, invalid raw values, incorrect conversion of raw values or inconsistent raw values with another variable or record. Also, incomplete data issues consist of missing values when required. (Gupta, 2008)

THE SOLUTION TO RESOLVE DATA ISSUES

The benefits of using an application with validated edit check macros includes increased productivity by quickly and easily applying the macros to other clinical studies, team endorsement to use the systematic approach, and the ability to communicate common issues/concerns in a consistent manner. Ensuring confidence in the raw data for the biostatistician is an essential ingredient in delivering a quality discrepancy management product. The standardization has improved our productivity of data cleaning by 80%. Being a global CRO, this automated standard suite of cleaning software is essential to maximize resources as well as to improve customer satisfaction on the delivery of sound data.

If we had continued customizing software and lengthy manual reviews of listings, the clients would not be satisfied our timelines and sought other vendors. These refinements in our processes limits cost overruns by only issuing complex data checks for a few data points that are of essential value to the client.

SPECIFYING REQUIREMENTS IN DATA MANAGEMENT PLAN (DMP)

The first critical step in standards and compliance checks is to specify the requirements in a Data Management Plan (DMP). Within the DMP, the requirements should be clear and complete for all possible data issues. It will be helpful for the subject matter expert to use the case report forms and protocol when developing the requirements. In addition, often, important variables used in tables, lists and graphs maybe included in the DMP. You should also confirm with the team that the procedures prescribed from reviewing the DMP fulfill their original goal.

The following three levels of data checks should be performed: general clinical data checks, CDISC standard domain checks, and protocol compliance checks. (Doles, 2004)

I. General Clinical Data Checks:

A. Key Variables - All unique key variables in each data set are required.

1. Demographics (DM): Subject identification number is non-missing and unique.

B. Range Values – Identify laboratory test results that exceed upper and lower range thresholds.

1. Demog (DM): valid age values within lower and upper range values.
2. Laboratory Data (LB): valid toxicity and hemoglobin values within lower and upper range values.

C. Data Values - Display all unique values of selected variables.

1. Confirm negative or missing values are acceptable
2. Confirm small percentage of missing values across all variables (sort by freq)

D. Complex Data Values - Display values of selected variables to meet specific database queries.

1. End Point: The primary and secondary variables are consistent with the prescribed parameters defined by the statistician.

E. Compare Data Values – Confirm the logic between two variables.

1. Adverse Events (AE): AE description, AE preferred term, and AE system organ class are required variables if any are non-missing.

F. Compare Date Values - Confirm the consistency between two clinical dates.

1. Most all followup dates should be <u>after</u> most all baseline dates (screening, inform, randomization, first dose date)
2. Most all followup dates should be <u>before</u> most all EOS dates (death date)

G. Unique Records - Check for duplicate records by key variables as well as all variables.

H. Compare common variables - Compare and identify differences of common variables between two data sets.

1. Raw AE data set and Analysis AE data set.

II. CDISC Standard Domain Checks:

A. AE (Adverse Events)

1. Logical and non-missing AE start and AE stop dates
2. AE start or AE stop dates after last dose or EOS date
3. Uncoded AE preferred or AE system organ class term for non-missing AE event
4. Serious AE without any AE action taken
5. Inconsistent 'No AE action taken' and at least one AE action code exists
6. Inconsistent 'Continuing = Yes' and AE stop date
7. Duplicates on AE start date, AE stop date, AE event, AE preferred term, and AE system organ class
8. Missing AE event for coded AE preferred or AE system organ class term

B. CM (Concomitant Medications)

1. Logical and non-missing CM start and CM stop dates

2. CM start or CM stop dates after last dose or EOS date
3. Uncoded CM preferred term for non-missing CM event
4. Missing unit for non-missing dose

C. DM (Demog)
1. Protocol compliance on patient population – min/max age, gender, race, lab levels, etc.
2. Inconsistent patient in data set and not in DM

D. EX (Exposure)
1. Missing unit for non-missing dose
2. Inconsistent EOS reason with early termination reason
3. Confirm drug dose calculation

E. EOS (End of Study)
1. Inconsistent 'Complete study' and 'Reason for study termination' non-missing
2. Inconsistent 'Reason for study termination' and death date, AE crf, etc.

F. LB (Lab - Vertical structure)
1. Correct conversion of each lab value from reported units to standard international units
2. Valid lab tests for ltestcd value
3. Consistent and valid units for ltestcd value
4. Missing unit for non-missing lab value
5. Valid normal range flags
6. Valid unique visit flag for multiple labs on same day

III. Protocol Compliance Checks:
1. Confirm Demog patient counts across all data sets.
2. Are there any protocol violations that should be excluded from analysis?
3. Are the treatment groups randomly distributed based on safety subset population?
4. For each lab, are there major deviations in value from baseline over time?
5. For each lab data transfer, are patients correctly identified?
6. Are the top 10 adverse events expected?
7. Are patient follow-up visit windows in compliance with the protocol? Check for differences between two clinical dates. Ex. Lab dates should be after 1 week of dose dates.
8. For any critical variable, are there any statistically significant outliers?

DEVELOPING EDIT CHECK MACROS

The challenge is to develop a system that is flexible enough to run selected data checks, allow the consumer the ability to modify data checks and provide a method of adding new data checks as requested. In addition, it is important for the system to display the message 'No records found' to confirm that the data quality check was performed and that data assumptions were met. Finally, feedback from the team needs to be incorporated within the edit checking and reporting process for each data issue identified. This is important to prevent 're-inventing the wheel'. (English, 2005) By using common options and statements of selected SAS procedures and standard titles and footnotes, a macro based system was developed with universal application that could be used across all teams.

To meet the requirements of the DMP, data issues are categorized so that edit check macros would have specific functions. In general, the types of data issues in table 1 can be addressed by edit check macros in table 2.

Table 1. Types of Data Issues

Type of Data Issue	Brief Description
Acceptable Values	Values are one of the valid values for variable
Character Formats	Format of values within character variables are as expected, ex. XXX-XXXX
Consistency Across Variables	Values are consistent across multiple variables
Consistency Across Data sets*	Values are consistent across multiple data sets
Non-Duplicate Records	Each record is unique and not duplicated
Overlapping Records	Records that inappropriately overlap. For example, treatment records that overlap cycles in an oncology study.
Protocol Compliance Rules	Study specific logic-based check to confirm data compliance, ex. lab conversion
Range Check	Values are within a specified range
Required Value	Value is non-missing
Unique Value	Values are unique

* May require extra programming step since most all edit check macros require single data set.

Table 2. Brief Description of Selected Edit Checks from the SAS Macro Library

Macro	Brief Description
%crt_comp*	Compare data values and variable attributes of two data sets (Proc Compare).
%exprpt	Exception Reporting: data set exists?, variable exists?, records exists?, non-missing values exist?
%negval	Check for negative values.
%subqry*	Display selected variables in one data set based on condition in another data set, useful for checking data across data sets (Proc SQL with subquery).
%u_eccust	Display selected variables based on customized user conditions using a single IF statement. Note that this is an exception macro since most all other edit check macros use the WHERE statement.
%u_ecdup	Check for duplicate records.
%u_ecfreq	Display frequency of selected single or multiple crossed variables (Proc Freq).
%u_ecmens	Display descriptive statistics including range values of continuous variables (Proc Means).
%u_ecprnt	Display selected variables with selected conditions (Proc Print).

* Requires two data sets instead of one.

SUMMARY

By implementing a simple, yet effective solution to this major problem, CDM is able to allocate minimum resources to discrepancy management. In addition, enhancements to data validation procedures for the study can be carried out with minimum SAS statistical programming experience. With the addition of the e-mail notification, the team is able to take action more quickly and communicate the data issues to the team more effectively.

SAS Programming hours using edit check macros for one study	8 hours
SAS Programming hours not using edit check macros for one study	40 hours

At a very high level, standards and compliance checks for data can be considered similar to the traditional user acceptance testing for system applications. The same principles of good requirements, valid data and coding, and comprehensive testing should be applied. SOPs and guidelines need to be written to assure proper steps are in place for processing and cleaning clinical data.

REFERENCES

Doles, W. *Managing Laboratory Data*, Data Basics, Society for Clinical Data Management, Spring 2004 newsletter, available at: http://www.scdm.org

English, L. *Don't just comply; Prevent*, SAS.com, available at: http://www.sas.com/news/sascom/2005q3/column_guestenglish.html

Gupta, S. *Clinical Data Acceptance Testing Procedure*, Pharmaceutical Programming, 2008, Vol. 1, No 2, pages 107-117

RECOMMENDED READING

Brunelle, R, Kleyle R., A Database Quality Review Process with Interim Checks, Drug Information Journal, 2002, 36: 357-367.

Ciambrone, K R. *Fit for Reporting*, Data Basics, Society for Clinical Data Management, Summer 2004 newsletter, available at: http://www.scdm.org

Collins, SH. *Ensuring Quality Data*, Data Basics, Society for Clinical Data Management, Summer 2007 newsletter, available at: http://www.scdm.org

Eckerson, W. *Data Quality and the Bottom Line*, The Data Warehousing Institute, 2002 available at: http://www.dataflux.com/Resources/file-stream.asp?rid=37

Fendt, KH. *The Case for Clinical Data Quality*, Data Basics, Society for Clinical Data Management, Summer 2004 newsletter, available at: http://www.scdm.org

Nahm, M., *Data Gone Awry*, Data Basics, Society for Clinical Data Management, Fall 2007 newsletter, available at: http://www.scdm.org

Nahm, M, Dziem, G, Fendt K, Freeman L, Masi, J, Ponce Z, *Data Quality Survey Results*, Data Basics, Society for Clinical Data Management, Summer 2004 newsletter, available at: http://www.scdm.org

CONTACT INFORMATION

Sunil K. Gupta
Senior CDISC/SAS Consultant
Gupta Programming
Phone: (805)-577-8877 E-mail: Sunil@.GuptaProgramming.com

Sunil is an international speaker, best-selling SAS author, and a global corporate trainer. Sunil is the Principal SAS Consultant at Gupta Programming since 1994. Most recently, he completed both of his CDISC online classes with University of California at San Diego and SAS Institute India and will start to teach Sharpening Your SAS Skills online class for UCLA Extension in 2015. In 2011, Sunil launched his unique SAS resource blog, SASSavvy.com, for smarter SAS searches. Currently, SAS Savvy's membership consists mostly of SAS programmers, university students and corporate accounts.

Most recently, Sunil was recognized by SAS Institute's Circle of Excellence for 20 years of service. Last year, Sunil was an invited presenter at WUSS, NESUG and SESUG for his 'highly acclaimed' Proc SQL Hands-on workshop. He has been using SAS® software for over 20 years and is a SAS Base Certified Professional. He is also the author of Quick Results with the Output Delivery System, and Sharpening Your SAS Skills.