PharmaSUG China 2018 - Paper DM-45

Automate Clinical Trial Data Issue Checking and Tracking

Dale LeSueur, Krishna Avula, and Qin Li, Regeneron Pharmaceuticals Inc.

ABSTRACT

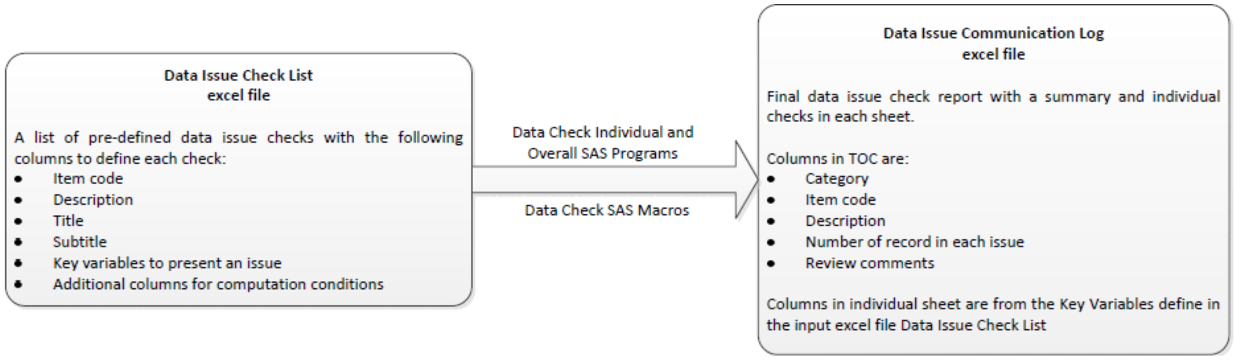
Well organized and properly cleaned data are fundamental for clinical trial data analysis. Programmatic data issue checks play an important role in data cleaning. Our tool is a set of SAS macro programs which automate data checking and issue tracking outside of data capture systems.

INTRODUCTION

Without quality data, the integrity of a study, safety and efficacy of a compound cannot be fairly evaluated. In practice, no matter how well a study is designed and conducted, errors can occur from different sources. Data cleaning is an essential process of identifying, resolving and documenting issues. It is achieved by implementing electronic checks in and outside data capture systems, and by monitoring and reviewing study data on regular basis. Programmatic issue checks outside the data capture system is a must for data reviewing, but issue updating and tracking is often manual and tedious.

Our tool is a set of macro programs designed to automate the checking and tracking of data issues. The programs call in a Data Issue Check List excel file which is pre-defined by study team at the beginning of a clinical trial, run each defined check by SAS macro program, and produce a final Data Issue Communication Log excel file recording and tracking all issues identified. The final excel file (Data Issue Communication Log) includes a Table of Contents summarizing all issues and counts, and individual tabs with subject/record level details for each issue. Hyperlinks are built for easy navigation from TOC to each tab of individual check and vice versa. The excel file is used to communicate and document data issues across functional areas and vendors. All remaining issues after database lock are documented with proper explanation in the excel file for future reference and documentation.

The simple flow chart below (Display 1, Process Flow Chart) describes the process of data checking and final report generating.



Data Issue Check List
excel file

A list of pre-defined data issue checks with the following columns to define each check:

- Item code
- Description
- Title
- Subtitle
- Key variables to present an issue
- Additional columns for computation conditions

Data Issue Communication Log
excel file

Final data issue check report with a summary and individual checks in each sheet.

Columns in TOC are:

- Category
- Item code
- Description
- Number of record in each issue
- Review comments

Columns in individual sheet are from the Key Variables define in the input excel file Data Issue Check List

Display 1, Process Flow Chart

THE OUTPUT: DATA ISSUE COMMUNICATION LOG

The most important component of a data issue checking is the output. We can programmatically perform important and difficult checks on the data, but if the output generated is not understandable or easy to follow, then the usefulness of the checks will be limited. We have created an output file that is easy to follow and use, but is also adaptable to the needs of different functions in a clinical team. Our output file is an excel file that includes one Table of Contents (TOC) tab as well as individual tabs for each check programmed.

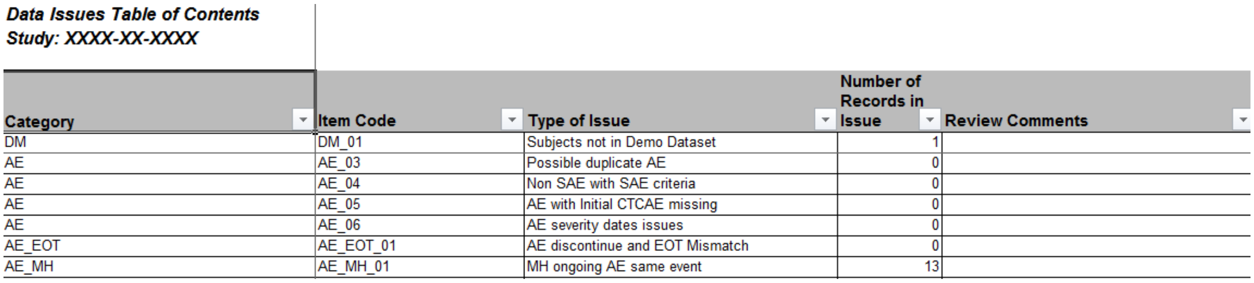
There are three key features in the excel file output that make the output easy to use and understandable:

- TOC tab: provide an overall summary of the issues being checked including the number of records identified with each issue.
- Hyperlinks: on the TOC tab there are hyperlinks to the individual tabs in the excel file for each check. The tab for each individual check also has a hyperlink back to the TOC.
- Review comments: We set up the output to include columns where comments can be entered in TOC tab and individual tabs. Comments for continuing issues will be carried forward when the output is updated on a new data transfer.

SUMMARY TABLE OF CONTENTS TAB

The Data Issue Communication Log output starts with the TOC tab (Display 2 shows an example) which includes five columns to give an overall picture of the data issues:

- Category - identifies a general category for each check (e.g. “AE”, “DM”, etc.) and can be used for filtering to specific types of checks.
- Item code - unique identifier for each check.
- Type of issue – short description of what was checked. This column includes a programmed in hyperlink, clicking on the text will open up the specific tab for the particular check.
- Number of records in each issue – a count of how many records from the clinical database were identified for a particular check.
- Review comments – column to collect general comments regarding an issue. Comments entered on the TOC page will be carried forward when the output is updated on new data transfer.



Data Issues Table of Contents
Study: XXXX-XX-XXXX

Category	Item Code	Type of Issue	Number of Records in Issue	Review Comments
DM	DM_01	Subjects not in Demo Dataset	1	
AE	AE_03	Possible duplicate AE	0	
AE	AE_04	Non SAE with SAE criteria	0	
AE	AE_05	AE with Initial CTCAE missing	0	
AE	AE_06	AE severity dates issues	0	
AE_EOT	AE_EOT_01	AE discontinue and EOT Mismatch	0	
AE_MH	AE_MH_01	MH ongoing AE same event	13	

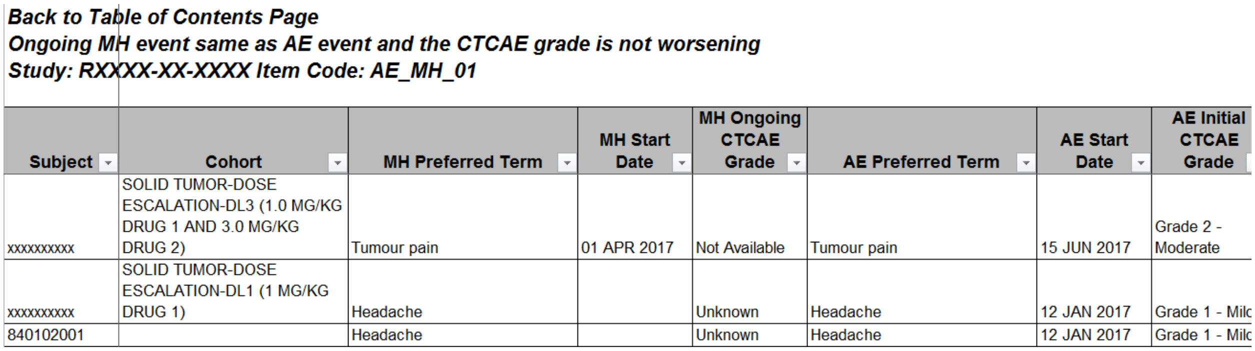
Display 2, Output Excel Report, Data Issue Communication Log - Table of Contents

INDIVIDUAL TABS FOR EACH ISSUE

Each programmed data check will have its own individual tab (Display 3 shows an example) in the output excel file. At the top of the individual tab there is a header that includes up to four lines:

- The first line of the header will have the text “Back to Table of Contents Page”. It serves as a hyperlink back to the TOC.
- The second line of the header will be the title for the individual check (from the Data Issue Check List excel file).
- The optional third line of the header will be a subtitle line that can be used to explain any highlighting in the output for the check (from the Data Issue Check List excel file).
- The last line of the header will be the Study number and the Item Code of the check.

Below the header will be the records found in the raw data with the particular issue. The columns included in the output are specific to each individual check. Enough columns are included to clearly display the issue and identify record in the clinical database where the issues come from. Additional columns can easily be added if the clinical team reviewing the output determines that more information is needed to demonstrate the issue. In situations where multiple values from the same data source are included (e.g. vital signs – blood pressure, heart rate, respiratory rate, etc.), highlighting of cells can be used to indicate which specific value is the one with an issue.



Back to Table of Contents Page
Ongoing MH event same as AE event and the CTCAE grade is not worsening
Study: RXXXX-XX-XXXX Item Code: AE_MH_01

Subject	Cohort	MH Preferred Term	MH Start Date	MH Ongoing CTCAE Grade	AE Preferred Term	AE Start Date	AE Initial CTCAE Grade
xxxxxxxxxx	SOLID TUMOR-DOSE ESCALATION-DL3 (1.0 MG/KG DRUG 1 AND 3.0 MG/KG DRUG 2)	Tumour pain	01 APR 2017	Not Available	Tumour pain	15 JUN 2017	Grade 2 - Moderate
xxxxxxxxxx	SOLID TUMOR-DOSE ESCALATION-DL1 (1 MG/KG DRUG 1)	Headache		Unknown	Headache	12 JAN 2017	Grade 1 - Milc
840102001		Headache		Unknown	Headache	12 JAN 2017	Grade 1 - Milc

Display 3, Output Excel Report, Data Issue Communication Log - Individual Issue Tab

PRE-DEFINED LIST OF DATA ISSUE CHECKS

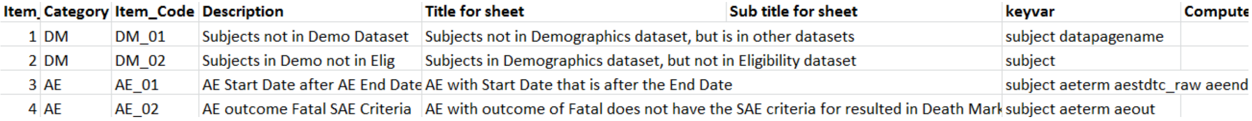
The first step towards generating the output is determining what to check? Ideally the checks are decided upon with the clinical team at the beginning of a clinical trial. These checks could include checks on values (e.g. do any values fall outside an expected range or values for a particular parameter), checks for consistency of data across domains (e.g. is information about AE action taken on the Adverse Event CRF page consistent with information entered on a disposition CRF page), and checks based on issues identified when validating SDTM datasets using Pinnacle 21.

The agreed upon checks are placed in a Data Issue Check List excel file. This file contains two tabs, a CheckList tab and a ProgSpecs tab.

CHECKLIST TAB

The CheckList tab (Display 4 shows example) identifies each of the checks to be performed and is structured with one row for each check. The following columns are included:

- Item code - unique identifier for each check.
- Description – short description of what will be checked.
- Title - will show up on the output for a particular check.
- Subtitle - optional, used to provide additional information about a check. For example, if highlighting is used the subtitle can be used to explain what the highlighting represents.
- Key variables to present an issue - set of variables to uniquely identify the record with an issue in the clinical database.
- Additional columns for computation conditions - compute statements in PROC REPORT, can be used to highlight particular cells in the output.



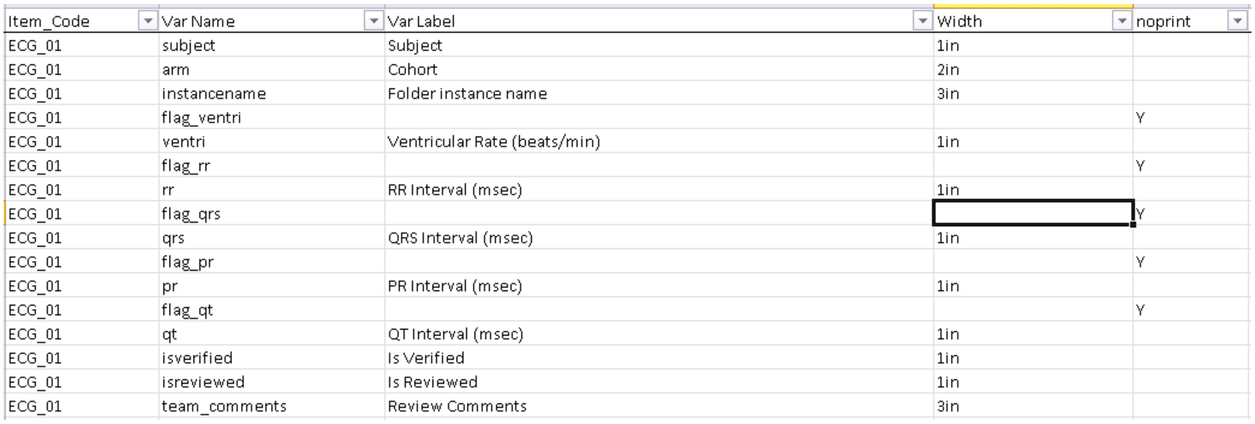
Item_	Category	Item_Code	Description	Title for sheet	Sub title for sheet	keyvar	Compute
1	DM	DM_01	Subjects not in Demo Dataset	Subjects not in Demographics dataset, but is in other datasets		subject datapagename	
2	DM	DM_02	Subjects in Demo not in Elig	Subjects in Demographics dataset, but not in Eligibility dataset		subject	
3	AE	AE_01	AE Start Date after AE End Date	AE with Start Date that is after the End Date		subject aeterm aestdtc_raw aeend	
4	AE	AE_02	AE outcome Fatal SAE Criteria	AE with outcome of Fatal does not have the SAE criteria for resulted in Death Mark		subject aeterm aeout	

Display 4, Input Excel File: Data Issue Check List CheckList tab

PROGSPECS TAB

The ProgSpecs tab (Display 5 shows an example) identifies the columns to be included in the individual output tab. This tab includes one row for each column that is to be included in individual output tab of the excel file for each check. The following columns included in this tab:

- Item code - unique identifier for each check.
- Var Name - name of the variables to be included as columns in the output.
- Var Label - label of the variables for a column in the output.
- Width - specify width of a column in the output.
- Noprint - marked as “Y” for variables which are used to flag issues to be highlighted but the variable won’t be shown in the output.



Item_Code	Var Name	Var Label	Width	noprint
ECG_01	subject	Subject	1in	
ECG_01	arm	Cohort	2in	
ECG_01	instancename	Folder instance name	3in	
ECG_01	flag_ventri			Y
ECG_01	ventri	Ventricular Rate (beats/min)	1in	
ECG_01	flag_rr			Y
ECG_01	rr	RR Interval (msec)	1in	
ECG_01	flag_qrs			Y
ECG_01	qrs	QRS Interval (msec)	1in	
ECG_01	flag_pr			Y
ECG_01	pr	PR Interval (msec)	1in	
ECG_01	flag_qt			Y
ECG_01	qt	QT Interval (msec)	1in	
ECG_01	isverified	Is Verified	1in	
ECG_01	isreviewed	Is Reviewed	1in	
ECG_01	team_comments	Review Comments	3in	

Display 5, Input Excel File: Data Issue Check List ProgSpecs tab

SAS PROGRAMS AND MACROS

The SAS programs and macros tie together the input excel file with the output excel file.

INDIVIDUAL CHECK PROGRAMS

Each check is performed by an individual SAS program which produces an output dataset containing all the records identified as a potential issue for a specific check as defined on the Data Issue Check List excel file. Both the name of the program and the name of the dataset will correspond to the Item Code. This naming convention facilitates the use of one main program to run all the individual programs and compile all the output. Separate programs for each check simplifies the process for adding new checks as existing programs do not need to be modified.

MAIN CHECK PROGRAM

The idea of the main check program is straightforward – determine which checks to run, run them, and output the results. Macros are used to simplify the process of doing repetitive similar tasks (e.g. running all the individual check programs). The basic flow is as follows:

- Import Data Issue Check List excel file. This import will identify which checks to run and also provide specifications for the output file.
- The clinical team has the option of putting comments into the Data Issue Communication Log output file. In the case the program will import comments from the previous review cycle. These comments will be merged with the individual check datasets. For any issues that remain from the previous cycle the comments will be carried forward into the new output.
- Based on the individual check datasets, determine the numbers of records with an issue identified for each check, this will be shown on the TOC tab.
- Generate the Data Issue Communication Log output file. The TOC tab is generated first. A macro is then used to create the tabs for each of the individual checks.

CONCLUSION

Reporting data issues is an important part of a statistical programmer’s job in the pharmaceutical industry. Identifying, communicating and documenting data issues can be a tedious and time consuming process. Our tool for automating clinical trial data issue checking and tracking has significantly cut down on the time that we spend searching for and documenting data issues as we work with clinical trial data. It has also helped us be more proactive in carrying forward checks from one study to other similar studies. The clinical teams have also found these outputs to be very helpful as they review and clean the data for clinical trials, and communicate and document issue and resolutions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dale LeSueur
Regeneron Pharmaceuticals Inc.
Dale.lesueur@regeneron.com