

Data Quality Checker

Ryan Finch, Roche Products Ltd, Welwyn Garden City, UK

ABSTRACT

Before starting to create reporting objects it is important to be confident that the data has reached the quality level required. Years of experience have taught us that whatever the number of Data Management checks and Science data review cycles, the standardization of your CRF/eCRF and database, you will always find specific scenarios that may cause your programs to fail.

This is why Roche has implemented a process to review key data before database lock to ensure that they match our expectations. This process is called the DQC “Data Quality Checker”. This process is even more critical for Roche, than perhaps for other companies, due to the fact that the data is extracted by Clinical Programming (Data Management) based on a specification document written by Statistical Programming (derived data as well as raw CRF data). The DQC tool is used to perform the user acceptance tests of the extracted datasets to ensure they match the specifications and that derivations are correct.

This paper will describe how the DQC is defined based on the standard structure of Roche data and how the programmers have the ability to use the DQC tool on the data – performing standard and non-standard checks. It will also contain a brief overview of the SAS macro-procedures available and its use. Finally it will expose some “real life” examples and how Statistics have worked with Data Management and Science at Roche to try resolve them, working as a cross-functional team to ensure the data reaches the highest possible level of quality.

INTRODUCTION

This paper will answer the following questions:

- *What* is the Data Quality Checker (DQC) tool?
- *Why* has this process been put in place, and why does Roche use it?
- *When* is this process used, and *how* do Statistics and Data Management interact with each other

BACKGROUND INFORMATION

Roche have a generic SAS dataset model (GDM) for CRF data (similar to the CDISC model). Statistical Programming write the specification document for how the data should be extracted by Clinical Programming (including basic derivations as well as the raw CRF data). Most standard Roche CRF pages collect the same data in the same format therefore the specifications follow a similar nature – for example personal data page, adverse event pages, etc. However each protocol has non-standard data depending on the therapeutic area of the protocol – for example a Rheumatoid Arthritis protocol collects Joint Count data whereas an Oncology protocol will collect Tumor data. This non-standard data is translated into either non-standard GDM datasets or as non-standard variables on existing GDM datasets.

WHAT IS THE DQC TOOL?

The DQC is Roche's global tool for checking the quality of protocol GDM data. It originated as a local tool for one of the sites, and as the company has moved to align processes from site-to-site the DQC has become the recognized tool used by the Statistics department. It is used to ensure the data is cleaned to the highest possible quality, and also to highlight any problem pieces of data which may cause other Roche statistics tools to fail (for example, MARS – Roche's safety reporting system).

The DQC tool is a collection of SAS macros that are used to perform a series of checks on the data. The tool is controlled by a control-program (meta-file) containing all the checks, and a calling-program which interfaces with the control-program to produce a list of the discrepancies of the data in the form of SAS output ("proc print"). The other macros sitting in the background are not used directly by the user but the tool interacts with them depending on the check it is performing. The DQC is used as the key tool for performing "error trapping" on all CRF data – collecting all discrepancies in a single location rather than having multiple other programs. The tool also retains useful information about how long discrepancies have been outstanding in the data, allowing the user to specify whether to see all existing discrepancies or just the new items since the last run. The sole purpose for the DQC is to track and raise data issues; the tool is not used for any derived dataset or report object creation programs.

The DQC contains standard Roche checks but it also offers the user the flexibility to tailor the checks to the protocol as required (protocol-specific checks). The majority of the standard Roche checks are applied to all protocols, but they can be switched off if they are not appropriate to the protocol in question.

The protocol-specific checks can be added to the tool at any time during the lifecycle of the protocol. It is best to try and front-load as many of them at the beginning during the study set-up phase when the Statistical Programmer is also writing the data specification document. This will help with testing the derivations in the data delivered by Clinical Programming (user-acceptance testing; UAT). However, there are checks that are likely to arise after UAT has been completed and after data review cycles that can be added to the tool any time up until Database Closure (DBC).

THE CHECKS

- What type of checks does the tool consist of?
 - REQUIRED (variables that need to be on datasets and populated correctly)
 - WHERE (SAS where statement processing)
 - STATEMENT (SAS if statement processing)
 - PROGRAM (more complex checks that need to be done in SAS programs i.e. more than one data step)
- How the checks are applied?
 - Standard checks on standard datasets
 - Standard checks on non-standard datasets (protocol-specific)
 - Non-standard checks (protocol-specific) on standard datasets
 - Non-standard checks (protocol-specific) on non-standard datasets (protocol-specific)

EXAMPLES

Standard Checks on Standard datasets

These are the default checks that are built into the tool to be applied to all protocols under the assumption that every protocol will be collecting the data and therefore should be applied. The user of the tool has the option to turn off these default checks and/or amend them according to the protocol requirements.

PhUSE 2007

A standard check would be to ensure that patient identifying data such as protocol id, center number and patient number are populated in each piece of data. This sounds a very simple requirement, but one that is crucial to ensure that data can be combined together and used in analyses, ensuring that any subsequent data collected is not lost because it is not possible to identify where it belongs. This check itself is applied to all standard data (as well as for non-standard data). This particular check makes use of the “REQUIRED” type of check.

Another example of a standard check on standard data, but using a different type of check could be for gender. Gender is collected in almost every protocol as a categorical variable where the responses should be “MALE” or “FEMALE”. A “WHERE” type of check is performed on this variable to flag any occurrences where the value is anything else.

Standard Checks on Non-standard datasets

These types of checks are the ones that are specific to the protocol’s therapeutic area. As previously mentioned, a rheumatoid arthritis protocol collects joint count data in a non-standard dataset – this is regarded as a non-standard dataset because other protocols (such as an oncology study) don’t collect this same data. Although this is regarded as a non-standard dataset, it is worth mentioning that at Roche “knowledge sharing” is a commonly used phrase and other Roche projects in the rheumatoid arthritis therapeutic area try to collect the data in the same way, and database it in the same structure so as to share code from protocol to protocol where possible. The same type of standard checks can be applied to this data as before – protocol id, center number and patient number need to be populated and be consistent with other data. Other items such as dates are collected in a similar manner to those in standard datasets and the same type of standard checks are applied.

Non-standard Checks on Standard datasets

This section is for those pieces of data where they don’t conform to the standard Roche GDM data, but where the data fits in best for analysis if it is added to existing data as an extra variable rather than create a new dataset. These additional pieces of data are generally protocol-specific, but can also be new pieces of data that the regulatory authorities now require Roche and other sponsor companies to collect and where the Roche GDM data template specifications need updating – for example, ethnicity is now collected on all ‘new’ Roche protocols – or for protocol specific data such as date of Rheumatoid Arthritis diagnosis. As before, the same type of checks (i.e. REQUIRED, WHERE) are applied.

Non-standard Checks on Non-Standard datasets

These types of checks typically relate to what each statistics team create as additional checks to ensure that the therapeutic specific data is being collected and populated as expected in these additional datasets.

Type of checks

As referenced in the previous examples, the most commonly used checks applied on the data are the “REQUIRED” and “WHERE”. The “STATEMENT” and “PROGRAM” checks are reserved for more complex and detailed tasks.

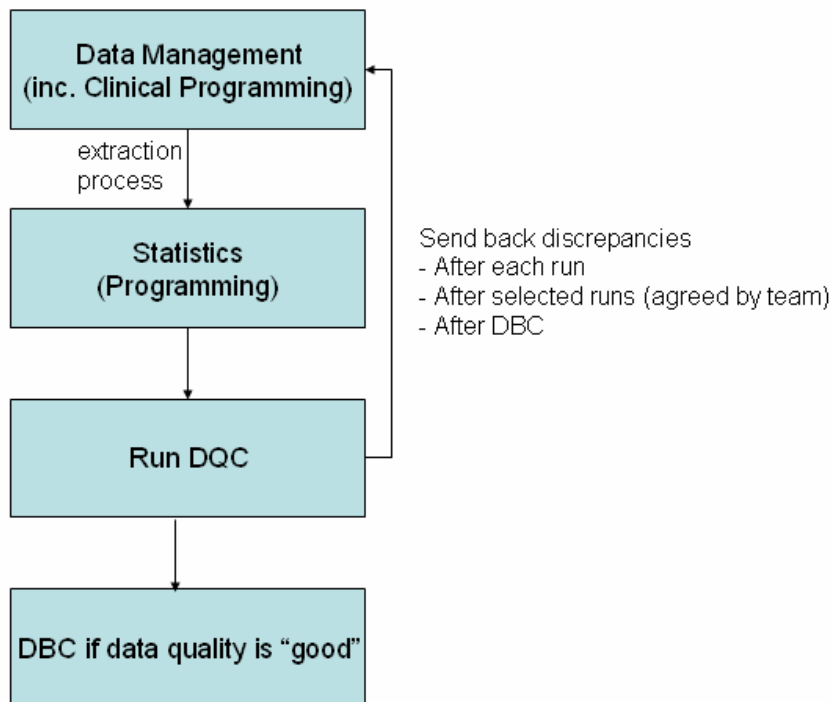
A “STATEMENT” check is used when a SAS WHERE statement is not sufficient and conditional IF processing is required – a piece of data that is dependent on another piece of data, and SAS IF processing is required instead of just a single WHERE statement. For example, if female reproduction details are collected for a protocol then this data should only be completed when gender has been selected as female. The check ensures that if gender is male then reproductive status is missing as well as checking the responses for female patients to ensure that the values are within the selected categories and not missing.

The use of “PROGRAM” checks are slightly more complex still, but it gives the user the ability to go away and do further processing on multiple pieces of data (several SAS data steps or SQL). The default checks using “PROGRAM” checks are the ones to ensure that data in every dataset has a corresponding record in DEMO for personal data information as the demographic data (one record per patient) is what is used to combine with all data for analysis.

PhUSE 2007

As mentioned, Clinical Programming derive some variables for Statistics based upon specifications provided by the Statistical Programmer. These are generally kept to a minimum and only contain basic derivations – anything more complex is done by the Statistical Programmer after the extraction process. For example these derivations include delivering date-time and study day variables. The use of “PROGRAM” checks interrogate the dates-times and study day variables to ensure that they are logical; for example the begin date of an Adverse Event is before the end date and is populated correctly, and when examining date-time variables it checks that the date and time have been combined correctly as well as checking the study day derivation has been calculated correctly using the correct variables as documented in the specifications.

PROCESS OF ENSURING DATA QUALITY – USING THE DQC



WHY USE THE DQC AND THIS PROCESS

The purpose for the statistics department using the DQC tool and this process is to ensure that the data delivered from Clinical Programming (Data Management) is to the highest possible level of quality. It helps statistics understand the data and any possible issues up-front instead of wasting valuable time later down the line investigating why programs or other tools fail.

The tool itself is used two-fold:

- firstly to UAT (user acceptance test) the data extracted and delivered by Clinical Programming to ensure it conforms to the specifications mapped out by Statistical Programming mapping. To check the data structure (i.e. attributes, labels, formats) as well as test the derivations.
- secondly as part of the ongoing/final process for reviewing data values for completeness/accuracy.

Data Management have the overall responsibility for “Data Cleaning” – the job of querying and cleaning the database to get the data as accurate and as complete as possible. However, the statistics department supports the “Data Review Plan” along with all the other clinical functions (Science, Labs, and Operations) which is used to determine the data that “must” be cleaned.

PhUSE 2007

Statistics get involved with reviewing and accepting the Data Management “Validation Specifications” which define which pieces of data *must* be cleaned come DBC time. These are the “edit checks” that Data Management are performing as part of the “data cleaning” process. Some of the DQC checks are repeats of Data Management’s “edit checks” but are included to ensure that the data discrepancies are resolved satisfactorily as statistics don’t see the edit check queries directly.

EXAMPLES – SCREEN SHOTS OF OUTPUT FROM THE TOOL

```

rckrun_demo.out - /opt/BIOSTAT/home/finchr1/cdp99999/
File Edit Search Preferences Shell Macro Windows Help
/opt/BIOSTAT/home/finchr1/cdp99999/rckrun_demo.out 9200 bytes L: 9 C: 0
CDP99999 - GDM data sets not defined 21:00 Monday, December 18, 2006 1
Obs TEXT
1 GDM data set AE does not exist!
2 GDM data set DIED does not exist!

```

This screen shows page 1 of the output which checks the standard datasets are available and being used.

```

rckrun_demo.out - /opt/BIOSTAT/home/finchr1/cdp99999/
File Edit Search Preferences Shell Macro Windows Help
/opt/BIOSTAT/home/finchr1/cdp99999/rckrun_demo.out 7197 bytes L: 27 C: 94
CDP99999 - Project-specific GDM variables 21:00 Monday, December 18,
Obs MEMNAME NAME TYPE LENGTH LABEL FORMAT
5 DEMO DIAGBGDC 2 9 Date of RA diagnosis
6 DEMO ESCDT 1 8 First Escape Datetime DATETIME20.
7 DEMO ETHNIC 2 12 Ethnicity
8 DEMO REGION 2 20 Geographic region
9 DEMO RNDDC 2 9 Randomization date
10 DEMO RNDNO 2 8 Randomization number
11 DEMO VISITDC 2 9 Visit Date
12 DEMO WGTGAT 2 20 Weight Category (kg)

```

This screen shows page 2 of the output which checks the data and highlights non-standard variables. These variables could be listed in your specifications provided to Clinical Programming and therefore acceptable or they could be temporary variables that have been forgotten to be dropped.

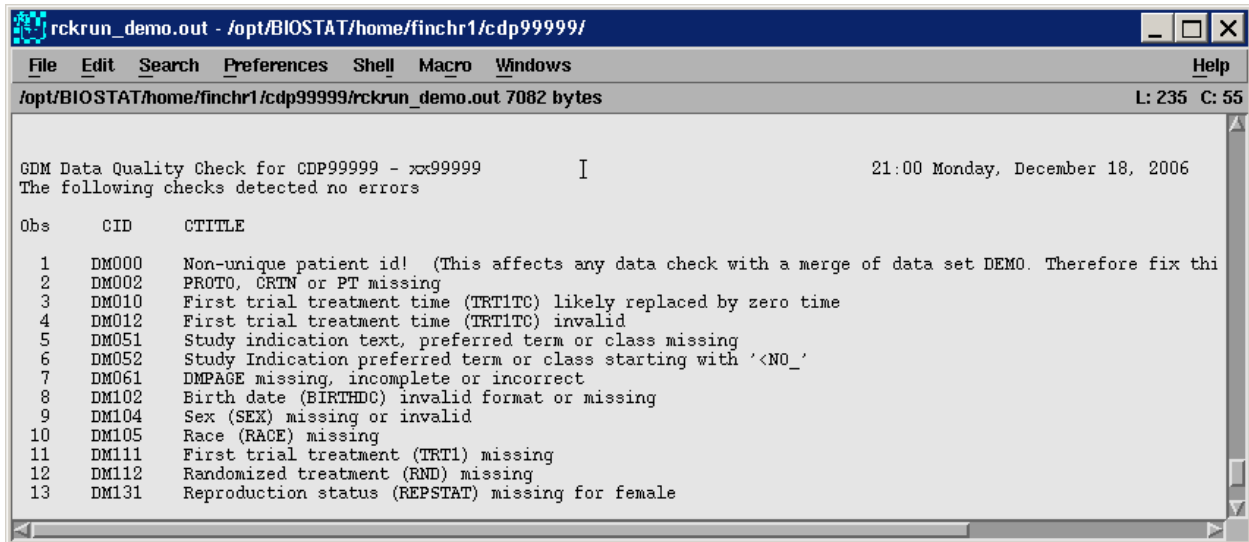
```

rckrun_demo.out - /opt/BIOSTAT/home/finchr1/cdp99999/
File Edit Search Preferences Shell Macro Windows Help
/opt/BIOSTAT/home/finchr1/cdp99999/rckrun_demo.out 7111 bytes L: 48 C: 1
GDM Data Quality Check for CDP99999 - xx99999 21:00 Monday, December 18, 2006
Check DM011 : First trial treatment date (TRTIDC) missing or invalid
Input Data Set : DEMO
[WHERE]: INPUT(TRTIDC,?? DATE9.) EQ .
Findings which emerged at any time are displayed.
CRTN PT TRTIDC _EMERGED
45621 3679 18NOV2006
86256 8652 NK NEW
..
..
2 of 256 observations printed.

```

PhUSE 2007

This screen shows page 3 (onwards) of the output which gives an example of a check. The check ID is unique for each check, and the information here tells the team which dataset has the issue and which variables are inaccurate. There are 256 discrepancies in the data for this specific check, however only 2 of them have been listed here. Printing all 256 is not necessary because a handful of records should be sufficient to show the user that there is a problem that needs to be addressed. The other records are available should the user want to view them all. The output also shows that one of the two discrepancies is "NEW" as of this run of the DQC.



```
rckrun_demo.out - /opt/BIOSTAT/home/finchr1/cdp99999/
File Edit Search Preferences Shell Macro Windows Help
/opt/BIOSTAT/home/finchr1/cdp99999/rckrun_demo.out 7082 bytes L: 235 C: 55

GDM Data Quality Check for CDP99999 - xx99999      I      21:00 Monday, December 18, 2006
The following checks detected no errors

Obs      CID      CTITLE
1      DM000      Non-unique patient id! (This affects any data check with a merge of data set DEMO. Therefore fix thi
2      DM002      PROTO, CRTN or PT missing
3      DM010      First trial treatment time (TRT1TC) likely replaced by zero time
4      DM012      First trial treatment time (TRT1TC) invalid
5      DM051      Study indication text, preferred term or class missing
6      DM052      Study Indication preferred term or class starting with '<NO_'
7      DM061      DMPAGE missing, incomplete or incorrect
8      DM102      Birth date (BIRTHDC) invalid format or missing
9      DM104      Sex (SEX) missing or invalid
10     DM105      Race (RACE) missing
11     DM111      First trial treatment (TRT1) missing
12     DM112      Randomized treatment (RND) missing
13     DM131      Reproduction status (REPSTAT) missing for female
```

This screen shows the last page of the output which returns information to the user stating which checks have been performed on the data and have detected no issues in the data. This is always nice to know that the checks have been performed and the data is looking clean.

WHEN IS THE DQC RUN AND HOW DO STATISTICS INTERACT WITH DATA MANAGEMENT

The DQC is recommended to be run after each data delivery received from Clinical Programming/Data Management. The reason being that if a new delivery contains potential issues it is possible to *reject* the delivery and to allow statistics to continue working on front-loading the analysis for the next reporting event. The issues can be passed back to Clinical Programming/Data Management to be resolved; some of these may be Clinical Programming derivation issues or Data Entry errors which can be more easily addressed, whereas others could result in queries at investigator sites which generally take a little longer to resolve. But having this process in place ensures that the statistics team can continue to work without being interrupted by problem data if it's a temporary issue that could be resolved by the next data delivery.

At Roche, project teams either submit findings from the DQC after each data delivery or at selective times during the lifecycle of the protocol. These decisions are usually based on the type of protocol (development phase; II / III), duration of the study (6 months, 2 years), and the frequency of data deliveries (weekly, monthly). Statistics may execute the DQC on every data delivery, but feedback on the findings is only provided to Data Management as agreed with the team in the Data Review Plan – the minimum occurrence is at Database Closure (DBC) time, but in practice it is more likely to be on a regular occurrence leading upto DBC or leading upto interim reporting events (DSMB/DSC, internal review meetings).

The Data Review Plan formalizes the DQC process and the SAS output with the list of discrepancies generally becomes the deliverable from the Statistics department. The output is self-explanatory to colleagues from Statistics, but requires some basic training with Data Management to help them read the output or alternatively some project teams translate the output into brief narratives and pass on the information.

PhUSE 2007

EXAMPLES

Project team A are working on a phase III Rheumatoid Arthritis trial that has a duration of 6 months. The protocol will consist of 750 patients. First review meeting will be when 50% of the patients have completed 3 months, data review at this stage. Second review meeting will be when 50% of the patients have completed all 6 months, data review at this stage. Final review meeting will be 4 weeks prior to last patient last visit, data review at this stage and weekly up until DBC. The DQC findings will be provided back to Data Management after each review. Data Management will follow-up on discrepancies, and relay information back to Statistics for data that is not possible to verify (this data could be patient data that was too long ago, or lab data that has been lost).

Project team B are working on a phase II Renal trial that has a duration of 20 weeks. The protocol consists of 190 patients. First review meeting will be when 70% of the patients have completed the study, data review at this stage. Final review meeting will be 2 weeks prior to last patient last visit and at DBC time. Data Management will receive DQC output and read it themselves, following up on discrepancies.

CONCLUSION

It is not possible to have the database 100% accurate as there are always going to be pieces of data that you can't rectify for one reason or another, but the Statistics team play a role within the cross-functional team to help ensure that the data is finalized to the highest possible standard.

The DQC tool is automated in the sense that all the checks are performed programmatically rather than visually looking at the data, therefore it is quicker to run and more efficient. Having all the checks in a single location makes maintenance easier too.

The benefits of using standard data (GDM) and a standard tool (DQC) globally at Roche promotes the use of "knowledge sharing". Within Roche there is a global working group that get together and look into issues with the DQC tool – adding new checks and amending existing checks as appropriate. These changes are driven by user findings as well as the GDM data model evolving (largely down to changes with how the regulatory authorities require the data).

CONTACT INFORMATION

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

Ryan Finch
Roche Products Ltd
Hexagon Place
6 Falcon Way
Shire Park
Welwyn Garden City
Hertfordshire, AL7 1TW
Phone: 01707 365630 (direct line)
Fax: 01707 383145
Email: ryan.finch@roche.com
Web: www.roche.com