PhUSE 2017

Paper SI09

Streamline SDTM Development and QC

Stephen Gormley, Amgen, United Kingdom

ABSTRACT

Amgen's Global Statistical Programming ("**GSP**") function have one centralised team (The CDISC Consultancy and Implementation Team - "**The CCI Team**"), using an in house Java Extract, Transfer and Load ("**ETL**") tool to produce all SDTM datasets and a number of associated submission deliverables (e.g. controlled terminology, xpt files, define.pdf) for all of Amgen's products. However, a different team within the GSP function ("**The Study Team**"), up until the middle of 2016, performed the Quality Control ("**QC**", i.e. verified the SDTM dataset is correct) of the SDTM datasets.

This approach had two areas of concerns for Amgen's GSP leadership team:

1. Two different sets of specifications maintained by the two different groups for both SDTM target structures and Controlled Terminology ("**CT**").
2. SAS used as the programming solution for the QC with 100% independent programming of each variable in every SDTM.

There were a number of problems identified as a direct consequence of the above approach, including the following four significant issues:

- First, the requirement on all of GSP (not just The CCI Team) to be experts in SDTM.
- Secondly, ambiguity in the accountability and decision making on SDTM.
- Thirdly, a large number of false-positives and/or minor differences that are raised when using independent programming in SAS with PROC COMPARE.
- Fourthly, a large duplication of resource effort.

Amgen's GSP function solved these problems by centralising the specification and QC of SDTM datasets within The CCI Team, together with taking a more risk based and automated approach (where possible) to the QC of SDTM.

INTRODUCTION

Amgen's GSP function changed the tools and processes for the specification and QC of SDTM datasets in Q1, **2016** into a more centralised service of SDTM specification and dataset creation, QC, and delivery.

Three key decisions were made:

- First, GSP would no longer QC using 100% independent programming of each variable in every SDTM.
- Secondly, GSP would introduce a more automated and calculated risk associated with SDTM dataset and variable QC.
- Thirdly, GSP would centralise the process within The CCI Team.

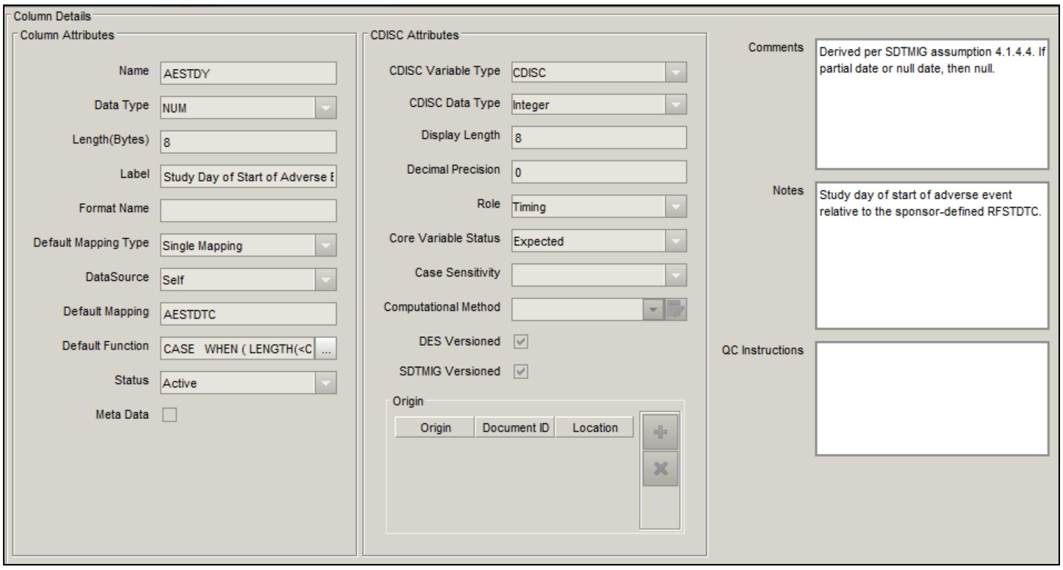
These three decisions were made in order to resolve the four significant issues highlighted previously (See **ABSTRACT** section). This paper shall go into detail regarding the problems with the previous non-centralised approach and also the re-designed GSP tools and processes. The Study Team and The CCI Team collectively shall be referred to as "**The Two Groups**" throughout this paper.

SDF[1]

The in-house Java ETL tool used within GSP is called the Submission Data File ("**SDF**") system and was designed to support the development, validation, maintenance and execution of SAS programs required to create SDTMs. A global library of metadata is stored in an Oracle database that describes each SDTM with the metadata stored for each SDTM IG version and each SDTM Amgen version. These metadata shall be referred to as "**SDF Global Library Templates**" throughout this paper and contain critical information regarding the SDTM, including, the source to target mapping, the target format, the QC instructions, the mapping type, functions required, controlled terminology and extended submission metadata (e.g. computational method, origin, core variable status, role). The SDF Global Library Templates are inherited at the study level for each study (or product) and these shall be referred to as "**SDF Study Level Templates**" throughout this paper.

Note: the SDF Study Level templates are identical to the SDF Global Library Templates unless updates are made at the study level. A link is maintained between the two so compliance checks can be run at the study level to ensure compliance with standards (See **A MORE RISK BASED APPROACH TO QC** section).

An example of one variable (from the AE SDTM) in an SDF Global Template stored globally and inherited locally:



Column Details

Column Attributes

Name: AESTDY
Data Type: NUM
Length(Bytes): 8
Label: Study Day of Start of Adverse E
Format Name:
Default Mapping Type: Single Mapping
DataSource: Self
Default Mapping: AESTDTC
Default Function: CASE WHEN (LENGTH(<C
Status: Active
Meta Data: □

CDISC Attributes

CDISC Variable Type: CDISC
CDISC Data Type: Integer
Display Length: 8
Decimal Precision: 0
Role: Timing
Core Variable Status: Expected
Case Sensitivity:
Computational Method:
DES Versioned: ☑
SDTMIG Versioned: ☑

Origin

Origin	Document ID	Location

Comments: Derived per SDTMIG assumption 4.1.4.4. If partial date or null date, then null.

Notes: Study day of start of adverse event relative to the sponsor-defined RFSTDTC.

QC Instructions:

This is a brief description of SDF to give some context while reading this paper and does not go into the full details of the SDF system. For further details on SDF please refer to the PHUSE paper: http://www.phusewiki.org/docs/2006/AD12.pdf.

QC: VALIDATION AND VERIFICATION

A distinction that shall be implied throughout this paper:

- **Verification**: the SDTM has been programmed correctly (i.e. the SDTM is progammed in accordance with the specification).
- **Validation**: the SDTM is programmed as expected (i.e. the SDTM is programmed in accordance with the expecation of the consumer).

For example, a missing AE.AEDECOD may be verified as correct if the source AE term was missing, but would not pass validation as it is not expected to be missing by the FDA. Throughout this paper QC mainly refers to the verification of SDTM, as the validation of SDTM is still performed at Amgen by using Pinnacle 21[2] validation tools.

SPECIFICATION

A good specification should be (indeed, has to be):

- **Precise**: tell the "programmers" what they need to know so that they can deliver the output required.

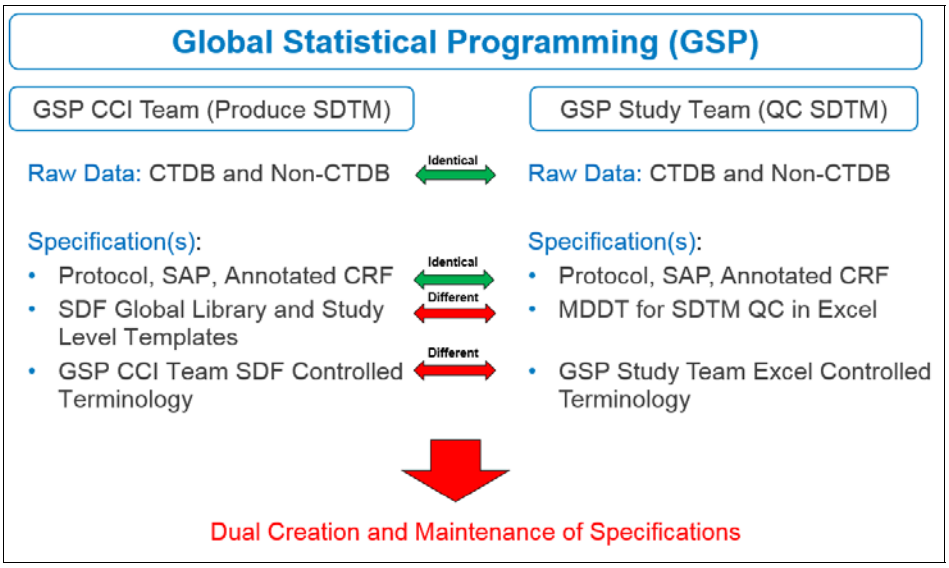
- **Correct**: describes what is really needed.
- **Clear**: easy to grasp and to understand without error for its intended users.

With The Two Groups producing a separate specification, ensuring these three key requirements were also not being fulfilled.

TWO AREAS OF CONCERN WITH THE PREVIOUS APPROACH

DUPLICATION OF SPECIFICATIONS

Problem One: two different sets of specifications being produced and maintained by The Two Groups with no centralisation and no standard approach across The Study Teams.



Global Statistical Programming (GSP)

GSP CCI Team (Produce SDTM)

Raw Data: CTDB and Non-CTDB

Specification(s):

- Protocol, SAP, Annotated CRF
- SDF Global Library and Study Level Templates
- GSP CCI Team SDF Controlled Terminology

GSP Study Team (QC SDTM)

Raw Data: CTDB and Non-CTDB

Specification(s):

- Protocol, SAP, Annotated CRF
- MDDT for SDTM QC in Excel
- GSP Study Team Excel Controlled Terminology

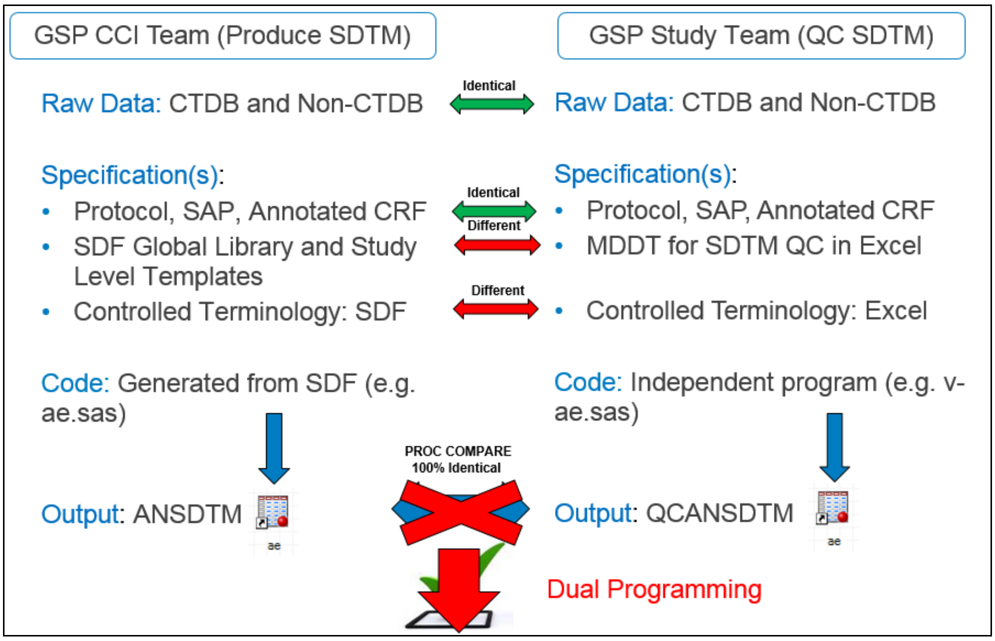
Dual Creation and Maintenance of Specifications

A number of the inputs used as the source for the creation of specifications were the same across The Two Groups (e.g. Analysis Plan, Protocol, Case Report Form). However, The CCI Team has an Oracle database with global libraries of source to target SDTM mappings (i.e. SDF Global Library Templates) and CT metadata. Conversley, The Study Teams had an Excel solution (different for each Amgen product) for the storage of both SDTM and CT metadata. A number of problems were identified with this approach, primarily a requirement on The Study Team to be experts in SDTM which led to:

- different interpretations of SDTM across both teams
- ambiguity in the accountability and decision making on the SDTMs
- ambiguity and inconsistency across the The Study Team Excel specifications
- a communication overhead trying to understand each of the two different sets of specifications
- problems determining the source of a QC issue, i.e. was the difference due to the specification or truly an error in the production program
- a lack of consistency in specifications and standards across The Study Teams (as a result of different Excel files being stored and maintained across products and therapeutic areas)

DUAL PROGRAMMING IN SAS

Problem Two: the primary QC method within each of The Study Teams was dual programming in SAS with no QC automation, no risk-based assessment of QC and the requirement that every observation in every variable had to match 100% with PROC COMPARE.



GSP CCI Team (Produce SDTM)

Raw Data: CTDB and Non-CTDB

Specification(s):

- Protocol, SAP, Annotated CRF
- SDF Global Library and Study Level Templates
- Controlled Terminology: SDF

Code: Generated from SDF (e.g. ae.sas)

Output: ANSDTM

GSP Study Team (QC SDTM)

Raw Data: CTDB and Non-CTDB

Specification(s):

- Protocol, SAP, Annotated CRF
- MDDT for SDTM QC in Excel
- Controlled Terminology: Excel

Code: Independent program (e.g. v-ae.sas)

Output: QCANSDTM

Dual Programming

There were a number of problems identified as a direct consequence of this approach, including:

- a large number of minor differences being found due to data issues that had not yet been resolved.
- an increase in the time it takes to determine the reason for any difference (i.e. is it the specification, the program, the raw data)
- the time and resource it takes to develop independent SAS code

Furthermore, this previous approach just felt like a missed opportunity for efficiency savings through a central, risk based and more automated QC approach.

OPTIONS CONSIDERED

Amgen's GSP department discussed at length a number of options to resolve the problems described in the previous sections. A number of meetings were held with key senior employees who were asked their opinion on the following four approaches:

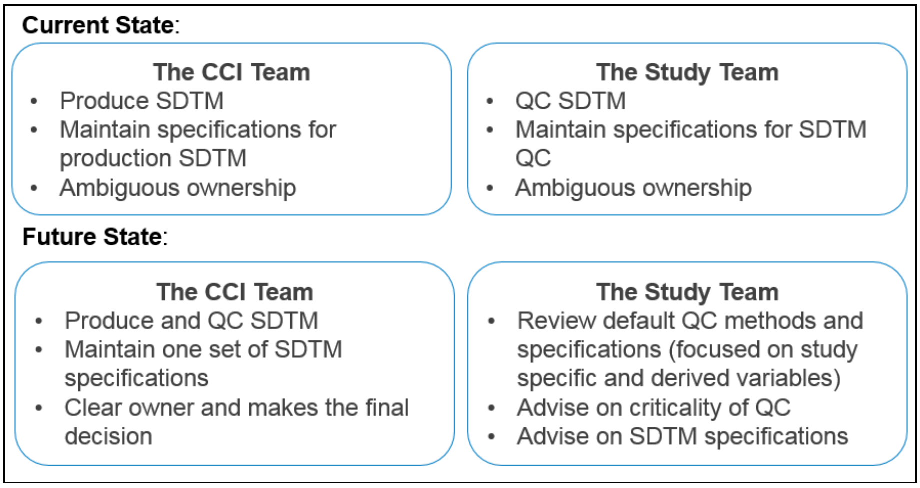
- **Option 1:** Keep the status quo, although there are a number of problems Amgen do produce a high quality SDTM package.
- **Option 2:** Outsource SDTM production and QC completely to a CRO.
- **Option 3:** The Study Teams both produce and QC SDTM.
- **Option 4:** The CCI Team both produce and QC SDTM.

The GSP Leadership Team selected option four, as the best solution to alleviate the four significant issues (See **ABSTRACT** section), for three main reasons:

- First, it was generally felt to be more efficient to keep both production and QC within one Amgen team.
- Secondly, the SDF system already had the SDF Global Library Templates centralised across all products with global librarians maintaining this generic metadata.
- Thirdly, The CCI Team had built up a wealth of expertise in SDTM with five employees whose specific role is of a CDISC SDTM Consultant.

THREE STEP SOLUTION

The solution centralised the creation of the specification and the QC of SDTM from within the SDF system, which was to be owned by The CCI Team with key advice required from The Study Team on study specific variables and the assignment of QC criticality.



Current State:

The CCI Team
- Produce SDTM
- Maintain specifications for production SDTM
- Ambiguous ownership

The Study Team
- QC SDTM
- Maintain specifications for SDTM QC
- Ambiguous ownership

Future State:

The CCI Team
- Produce and QC SDTM
- Maintain one set of SDTM specifications
- Clear owner and makes the final decision

The Study Team
- Review default QC methods and specifications (focused on study specific and derived variables)
- Advise on criticality of QC
- Advise on SDTM specifications

Furthermore, the QC was also automated, where possible, and risk based (i.e. the criticality and complexity of each variable in each SDTM would be assessed with an appropriate level of QC assigned).

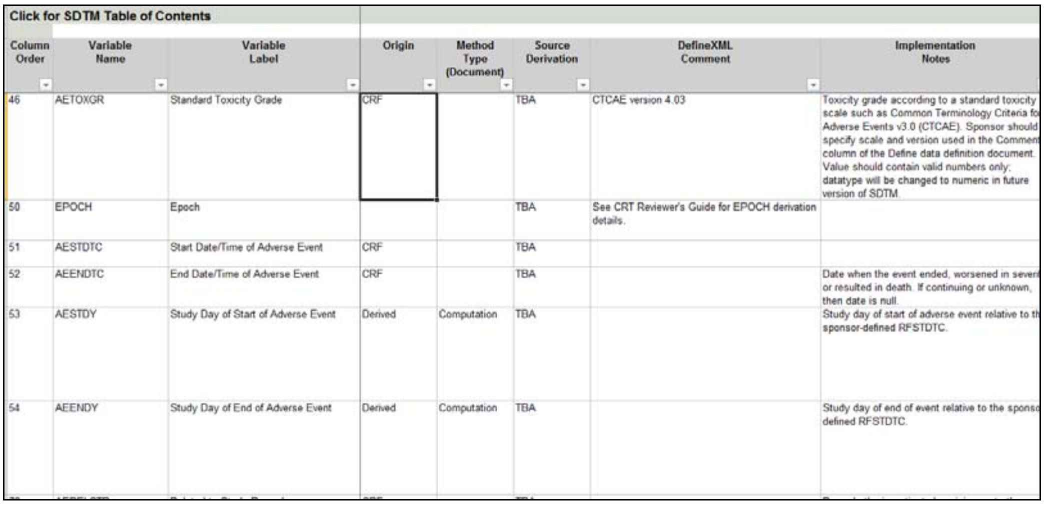
STEP 1: CENTRALISE THE SPECIFICATION

To alleviate the requirement on all of GSP (not just The CCI Team) to be experts in SDTM and to ensure no ambiguity in the ownership and decision-making on SDTM (i.e. significant issue one and two, see **ABSTRACT** section), the creation of the specifications were centralised into SDF and owned by The CCI Team. Nevertheless, in order to meet the three conditions of a good specification (See **INTRODUCTION > SPECIFICATION** section), for study specific items, The Study Team are still required to provide input into a few of the decisions on key SDTM variables (e.g. XXBLFL, EPOCH, EXDOSE, endpoints of interest that require a higher level of QC).

With the above in mind, the Oracle database containing the metadata used to create the SDTM from within SDF was realised in Excel for study team review and input. An example of which can be seen below, i.e. the example SDF Global Template (See **INTRODUCTION > SDF** section) displayed in an Excel format for consumption and review by The Study Team.

Note: The Study Team do not have access to the SDF System and are not trained in SDF or the SDF Global Library Templates, so a convenient solution had to be found for review and Excel was the most appropriate choice. Also, the new Excel output is almost identical to the previous Excel file the study teams used for their SDTM specification.

An SDF Template in Excel format for The Study Team review:



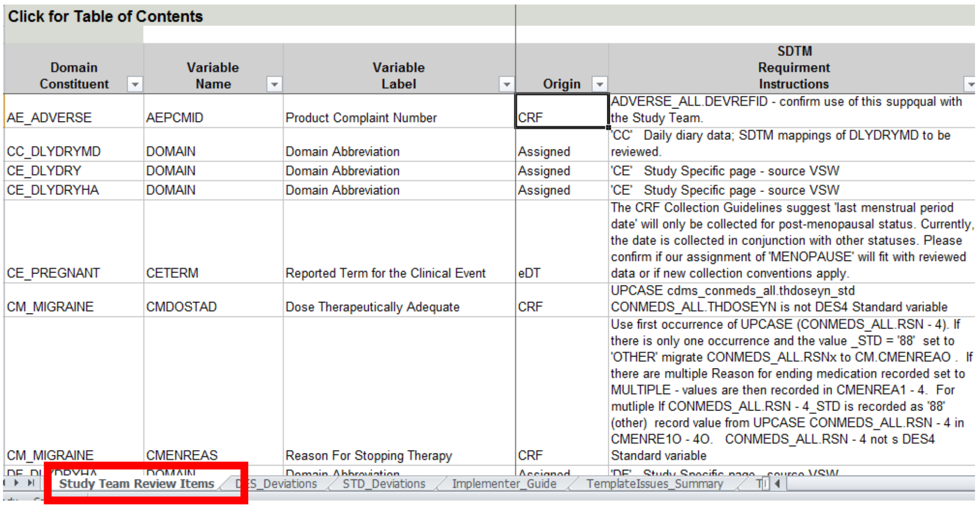
Click for SDTM Table of Contents

Column Order	Variable Name	Variable Label	Origin	Method Type (Document)	Source Derivation	DefineXML Comment	Implementation Notes
46	AETOXGR	Standard Toxicity Grade	CRF		TBA	CTCAE version 4.03	Toxicity grade according to a standard toxicity scale such as Common Terminology Criteria fo Adverse Events v3.0 (CTCAE). Sponsor should specify scale and version used in the Comment column of the Define data definition document. Value should contain valid numbers only; datatype will be changed to numeric in future version of SDTM.
50	EPOCH	Epoch			TBA	See CRT Reviewer's Guide for EPOCH derivation details.	
51	AESTDTC	Start Date/Time of Adverse Event	CRF		TBA		
52	AEENDTC	End Date/Time of Adverse Event	CRF		TBA		Date when the event ended, worsened in severi or resulted in death. If continuing or unknown, then date is null.
53	AESTDY	Study Day of Start of Adverse Event	Derived	Computation	TBA		Study day of start of adverse event relative to th sponsor-defined RFSTDTC.
54	AEENDY	Study Day of End of Adverse Event	Derived	Computation	TBA		Study day of end of event relative to the sponso defined RFSTDTC.

A high-level overview of key items that form part of the new specification process:

- The Study Team reviews each column in the Excel file that is deemed, by The CCI Team, to require further study team input before the specification can be finalised (generally complex variables/observation - e.g. EXDOSE, EPOCH, XXBLFL - and/or study specific source to target mappings - e.g. SUPPQUAL Variables, new PARAMCDs).
- The SDF Study Level Templates, inherited from the SDF Global Library Templates, contain the QC level to be assigned to each variable/observation (See **A MORE RISK BASED APPROACH TO QC** section).
- The SDF Study Level Templates are then updated based on The Study Team review of the Excel file.
- Variables that are critical for the study and/or those variables/observations that The Study Team require a higher level of QC (e.g. a primary endpoint lab value, LBTESTCD="HGB") can be highlighted in the Excel file by The Study Team.
- Simple source to target mappings with basic transformation logic that are in the SDF Global Library Templates and do not deviate from standards would not generally require The Study Team review.
- No more than 10% of variables should require further input by The Study Team.
- It is still incumbent on The Study Team to ensure that they are comfortable with the QC method assigned for each variable/observation as part of this review process.
- To aid the review process a separate sheet is created in the Excel file to list out all of the variables/observations that requires study team input.

An example of the sheet in Excel that lists all variables requiring review and input by The Study Team (collating all individial SDF Study Level Templates):



Click for Table of Contents

Domain Constituent	Variable Name	Variable Label	Origin	SDTM Requirment Instructions
AE_ADVERSE	AEPCMID	Product Complaint Number	CRF	ADVERSE_ALL.DEVREFID - confirm use of this suppqual with the Study Team.
CC_DLYDRYMD	DOMAIN	Domain Abbreviation	Assigned	'CC' Daily diary data; SDTM mappings of DLYDRYMD to be reviewed.
CE_DLYDRY	DOMAIN	Domain Abbreviation	Assigned	'CE' Study Specific page - source VSW
CE_DLYDRYHA	DOMAIN	Domain Abbreviation	Assigned	'CE' Study Specific page - source VSW
CE_PREGNANT	CETERM	Reported Term for the Clinical Event	eDT	The CRF Collection Guidelines suggest 'last menstrual period date' will only be collected for post-menopausal status. Currently, the date is collected in conjunction with other statuses. Please confirm if our assignment of 'MENOPAUSE' will fit with reviewed data or if new collection conventions apply.
CM_MIGRAINE	CMDOSTAD	Dose Therapeutically Adequate	CRF	UPCASE cdms_conmeds_all.thdoseyn_std CONMEDS_ALL.THDOSEYN is not DES4 Standard variable
CM_MIGRAINE	CMENREAS	Reason For Stopping Therapy	CRF	Use first occurrence of UPCASE (CONMEDS_ALL.RSN - 4). If there is only one occurrence and the value _STD = '88' set to 'OTHER' migrate CONMEDS_ALL.RSNx to CM.CMENREAO . If there are multiple Reason for ending medication recorded set to MULTIPLE - values are then recorded in CMENREA1 - 4. For mutliple If CONMEDS_ALL.RSN - 4_STD is recorded as '88' (other) record value from UPCASE CONMEDS_ALL.RSN - 4 in CMENRE1O - 4O. CONMEDS_ALL.RSN - 4 not s DES4 Standard variable
DE_DLYDRYHA	DOMAIN	Domain Abbreviation	Assigned	'DE' Study Specific page - source VSW

Once the specification is finalised for an SDTM then the production and QC of that SDTM can start. In the first few studies to go through this new process, this was a time consuming task. However, it is an important step: First, to ensure a good specification was in place before programming and QC begins; and secondly, a thorough review of the QC level is required, as those **not** flagged requiring a higher level of QC are just checked using data and metadata compliance checks (See **A MORE RISK BASED APPROACH TO QC** section).

STEP 2: A MORE RISK BASED APPROACH TO QC

To alleviate the large number of false-positives and/or minor differences (e.g. a different sort order, a different handling of a data issue) that are raised when using independent programming, in SAS with PROC COMPARE (i.e. Significant issue three, see **ABSTRACT** section), a more risk based approach was taken to the QC of SDTM's. The complexity and criticality of each variable/observation is reviewed by The Two Groups during the specification creation process (See **CENTRALISE THE SPECIFCATION** section), and an appropriate level of QC assigned (i.e. if the QC level is not already in the SDF Global Library Template or there is a deviation to the QC level in the SDF Global Library Template)

Note: There are still two roles within The CCI Team for the production and QC of SDTM, one member of the team is responsible for the production of SDTM and another is responsible for the QC. Both are heavily involved in the specification process to ensure specifications meet the three rules for a good specification (See **INTRODUCTION > SPECIFICATION** section), and that there is appropriate QC assigned to each SDTM.

The following four approaches are then used as the QC method for the SDTM, dependent on the criticality and complexity, as agreed at the specification stage:

1. **Basic: Metadata Checks**
    - **Variable Types**: Low risk, direct mappings from source to target, standard function conversions.
    - **Example Checks**:
        - Study Level Templates vs SDF Global Library Templates: e.g. Conformance to standards, lengths, sort orders, source data point
        - Submission Deliverables: e.g. Are computational algorithms populated, define produced, origin populated.
        - Analysis Level Checks: e.g. locked analysis for deliverable, all domains present in analysis folder, verification documentation complete.
    - **QC Approach:** Generic code used by all products and studies that is run to interrogate the SDF Global Library Templates and compared with the study level metadata to ensure compliance. Excel is the output used to illustrate any deviation to the standards and a disposition has to be entered, by the production programmer (reviewed by the QC programmer), into the Excel file if a deviation is found.

2. **Basic: Data Checks**
    - **Variable** Types: Low risk, direct mappings from source to target, standard function conversions.
    - **Example Checks**: Log checks, CT, uppercase, ranges, value level metadata, date errors, blank values.
    - **QC Approach:** Generic code used by all products and studies that is run to interrogate the study level SDTM data and ensure no issues are found. Excel is the output used to illustrate any problems found with the data and a disposition has to be entered, by the production programmer (reviewed by the QC programmer), into the Excel file if a data problem is found.

3. **Complex and Generic: SAS with PROC COMPARE**
    - **Variable Type:** High Risk, Critical and/or Complex.
    - **Example Variables:** General SDTM algorithms, e.g. ***XXBLFL, RFSTDTC, EPOCH***, transpose, observation checking.
    - **QC Approach:** PROC COMPARE using SAS macros specified, developed, tested and deployed generically which are then attached to the appropriate SDF Global Library Template, to be run at the study level (with study specific parameter settings, if needed). Excel is the output used to illustrate any differences in the PROC COMPARE.

4. **Complex and Study Specific: SAS with PROC COMPARE**
    - **Variable Type:** High Risk, Critical and/or Complex.
    - **Example Variables:** Study specific independent programs, e.g. ***EXDOSE, EXTRT, ARMCD, EPOCH, POPFLAGS***, ***XXSTRESC*** *(WHERE=PRIMARY ENDPOINT)*
    - **QC Approach:** PROC COMPARE using SAS macros specified, developed, tested and deployed locally for the study, which are then attached to the SDF Study Level Template to be run at the Study level. Excel is the output used to illustrate any differences in the PROC COMPARE.

Note: One Excel file is output per SDTM domain and includes all of the above checks if the domain requires all four different types of QC.

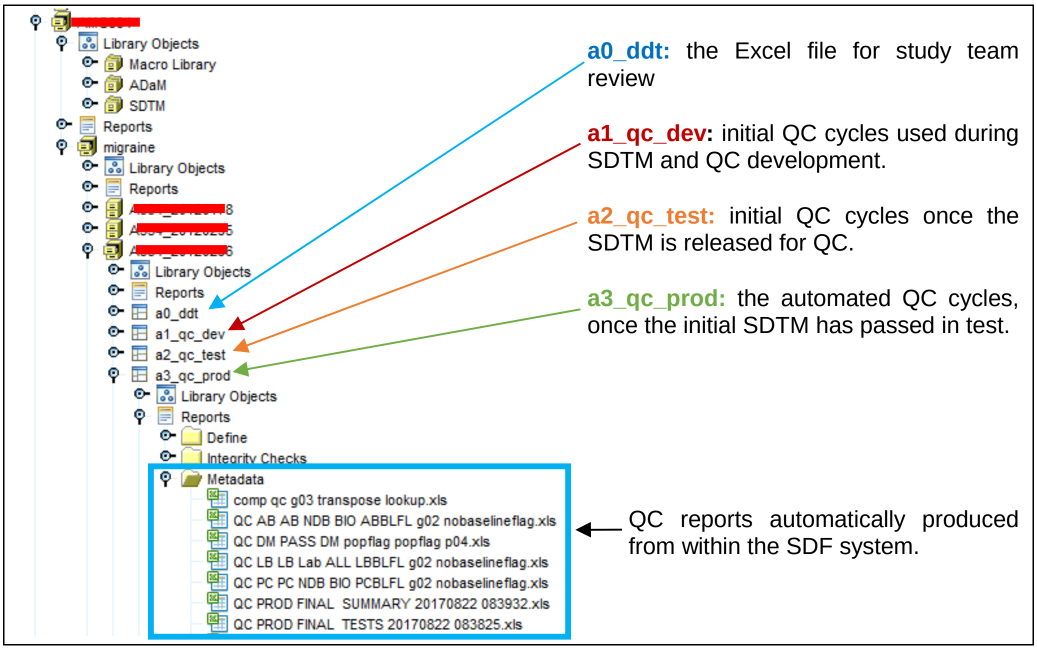
Once all of the following are complete, the QC can then be set up to run automatically:

- Complex and study specific QC SAS macro development has completed.
- Complex and generic QC SAS code is running with the correct parameter settings.
- Metadata and data checks are running and producing the compliance reports.
- Each Excel sheet for the four methods of QC are output as expected.

STEP 3: AUTOMATE THE QC

To further reduce the resource burden (i.e. Significant issue four, see **ABSTRACT** section) the SDF system was updated to allow the QC code to be run automatically (See example below of a directory tree within SDF) to allow for earlier detection and disposition of issues. Using a built in Java scheduler within SDF, each time the raw data is extracted, the SDTM is re-run and the QC code and compliance checks also run on schedule with each of the compliance reports and Excel files re-produced (See **A MORE RISK BASED APPROACH TO QC** section). Furthermore, all of the cycles of QC and disposition of issues between the production and QC programmer are contained within the Excel files, which are used as the evidence that the SDTM has formally verified in addition to the validation checks run using Pinnacle 21[2].

An example of the QC outputs within SDF organised by TA, Product, Indication, Study:



```
[redacted]
  Library Objects
    Macro Library
    ADaM
    SDTM
  Reports
  migraine
    Library Objects
    Reports
    [redacted]
    [redacted]
    [redacted]
      Library Objects
      Reports
      a0_ddt
      a1_qc_dev
      a2_qc_test
      a3_qc_prod
        Library Objects
        Reports
          Define
          Integrity Checks
          Metadata
            comp qc g03 transpose lookup.xls
            QC AB AB NDB BIO ABBLFL g02 nobaselineflag.xls
            QC DM PASS DM popflag popflag p04.xls
            QC LB LB Lab ALL LBBLFL g02 nobaselineflag.xls
            QC PC PC NDB BIO PCBLFL g02 nobaselineflag.xls
            QC PROD FINAL  SUMMARY 20170822 083932.xls
            QC PROD FINAL  TESTS 20170822 083825.xls
```

a0_ddt: the Excel file for study team review

a1_qc_dev: initial QC cycles used during SDTM and QC development.

a2_qc_test: initial QC cycles once the SDTM is released for QC.

a3_qc_prod: the automated QC cycles, once the initial SDTM has passed in test.

QC reports automatically produced from within the SDF system.

ROLL OUT

The solution was piloted on two studies in Q1 2016, and although it was a steep learning curve, The Two Groups adapted to the change relatively quickly and three critical database locks were taken with the new approach during the pilot. Furthermore, The Study Team were also asked to continue using SAS to QC the SDTMs during the pilot to ensure no loss in quality with the new approach. No major problems were identified with the QC of SDTM using the new approach in these three database locks. However, refinements to the process, primarily the communication between the two teams and how to track an update, were made.

Unfortunately, it is too early in the roll out to determine if resourcing will be reduced, as envisaged by the GSP Leadership Team at the outset.

CONCLUSION

Through the centralisation of the specification, production and QC of SDTM the Amgen GSP function have managed to alleviate three out of four of the major problems. i.e. The requirement on all of GSP (not just The CCI Team) to be experts in SDTM; the ambiguity in the accountability and decision making on SDTM; and the large number of false-positives that are raised when using independent programming in SAS with PROC COMPARE. However, the fourth and possibly one of the more important of the four, the impact on resource, has yet to be quantified.

CONSIDERATION

It has not been raised or discussed in this paper. However, to the author's knowledge, there is no formal guidance to the life sciences industry on specifically how to verify that the SDTM (or indeed ADaM) has been programmed correctly and in accordance with the specification. Although, the FDA does provide general principles on software validation[3] clearly indicating that there should be an assessment of risk when determining the level of testing any software. With both of these points in mind, during and after the presentation that accompanies this paper, it would be interesting to hear the strategy other companies adopt to the QC of their standardised datasets. As an industry do we spend too much time or too little time on the QC of SDTM? Should CDISC define a standard for the QC of SDTM and/or ADaM? What would be the impact of no QC at all (Just rely on validation using Pinnacle 21[2])? Indeed, do any companies rely solely on Pinnacle 21[2] as their sole QC method?

RECOMMENDED READING

[1] "**The Submission Data File System Automating the Creation of CDISC SDTM and ADaM Datasets**", [online], Available at http://www.phusewiki.org/docs/2006/AD12.pdf, [Accessed **20 August 2017**]

[2] **"Pinnacle 21 Validation Checks",** [online], Available at https://www.pinnacle21.com/products/validation, [Accessed **20 August 2017**]

[3] **"General Principles of Software Validation; Final Guidance for Industry and FDA Staff",** [online], Available at https://www.fda.gov/MedicalDevices/ucm085281.htm#_Toc517237938 , [Accessed **20 August 2017**]

CONTACT INFORMATION

Your comments and questions are valued and encouraged, feel free to contact:

Stephen Gormley
Amgen Ltd.
1 Sanderson Road
Uxbridge
UB8 1DH
Email: sgormley@amgen.com