Paper DH06

Standardising The Standards – The Benefits of Consistency

Nathan James, Roche Products Ltd., Welwyn Garden City, UK

ABSTRACT

The introduction of the Study Data Tabulation Model (SDTM) has had a positive impact on the submission of clinical data to Regulators *and* greatly improved the transmission of data between Sponsors and Contract Research Organisations (CROs) but the implementation and interpretation of the model can vary between companies.

Identifying these discrepancies early, before any analysis programming has started, can reduce any unexpected complications during pooling of data at a later stage and also facilitate the development of standard programs on a project.

This paper provides details on the design and requirements for an in-house tool that was developed to check SDTM data between multiple studies and some of the benefits of monitoring the consistency of dataset structures before any analysis. This includes the impact on general pooling strategies, standard program development and program life cycle as data is refreshed.

INTRODUCTION

It is now common practice for certain aspects of the drug development lifecycle to be outsourced to a specialist CRO and the tabulation of the clinical database into SDTM is no exception.

This was the case for a recent project I was assigned to, for which, two of the Phase III studies had the SDTM construction outsourced to two *different* vendors. These studies were likely to be pooled in the future and so the structure of the two studies' data was of particular importance, as this would be paramount to the success and efficiency of the pooling later on. Essentially, datasets with comparable structure and metadata are easier to combine programmatically.

From an end-user's perspective of SDTM, and as a receiver and not a creator of the tabulation datasets, the SDTM model was assumed to have guaranteed that all data tabulations would have a standard framework and structure, regardless of the vendor involved. The conformance of the study data tabulation to this model was the responsibility of the vendors, so fully-compliant SDTMs were set to be received by the statistical programming team.

Upon receiving the SDTMs for both studies for the first time, it quickly became evident that the construction of the SDTM datasets differed in a variety of ways. There were to be some expected differences between the two sets of tabulation data, based on certain study-specific elements, but there also existed some other inconsistencies. Given that the model is standard, these were unexpected and could potentially be problematic for the pooling of the studies later on. The extent of the differences between the sets of the SDTMs was also an unknown and so a tool to programmatically identify these disparities was developed, detailed later in the paper. This tool would prove to have more benefits to a programming team than simply checking the consistency of SDTMs between studies.

EXPECTED DIFFERENCES BETWEEN THE STUDIES

When comparing the SDTMs of two different clinical studies there will always be expected differences in the tabulation data. The similarity of the tabulation will generally be proportional to the similarity of the collected data, so two oncology studies will have more comparable SDTMs than two clinical studies from two *different* therapeutic areas. Likewise, the tabulation of the safety domains will be more comparable than any efficacy domains as the same set of safety data is often collected for all clinical studies. Hence, studies will typically differ in terms of the actual SDTM domains that are created from the source data.

DOMAINS DELIVERED

In general, the same safety data is collected on all clinical studies, so domains such as adverse events [AE], laboratory tests [LB] and treatment exposure [EX] were expected to be part of the transfer from the two vendors. Similarly, most studies will collect information about the subjects participating in the clinical trial and any major completion or withdrawal events, so demography [DM] and disposition [DS] domains will exist too as a bare minimum. Trial Design domains are a regulatory requirement for any submission package to the Health Authorities; hence, the five domains [TA/TE/TI/TS/TV] were anticipated to be present for both studies. The main expected difference between the SDTM constructs were to be related to the efficacy data, as some of the endpoints between the studies differed which may result in them being mapped to separate findings domains.

STUDY-SPECIFIC ELEMENTS

The second expected difference between the two sets of SDTMs was related to the study-specific elements of each study. All clinical studies will have fields on the Case Report Form (CRF) that are specific to that study and data will therefore be collected that only pertains to that individual study. As the collected data is likely to be non-standard, it will be mapped to either a permissible variable in the SDTM domain, the appropriate supplemental qualifier (SUPPQUAL) or to a sponsor-defined domain. It must be noted that the SDTM datasets delivered from our vendors had the SUPPQUAL datasets already merged with the base SDTM domain, as variables, and not separate. It was therefore assumed that there would be some slight variable mismatches between the SDTM domains of each study.

TRIAL DESIGN DOMAINS

The Trial Design domains, although required domains for all studies, were also projected to differ in terms of their content. These SDTM datasets contain metadata necessary to describe the design and planned conduct of the clinical trial to a reviewer. Two clinical studies of similar design will have Trial Design domains that are more alike, but as the content is very study-specific, there is a high possibility that the data within the domains is different. This is particularly true of the Trial Arms [TA] and Trial Elements [TE] domains since proposing a design for these datasets can be difficult. The fundamental definition that sets the foundation for the design of these domains is that of an ‘element’ of the study. An ‘element’ is described in the SDTM Implementation Guide as a basic building block in the trial design and is based on a planned invention, though the decision on what level of granularity is required is ultimately left to the person proposing the TE design. This definition is open to interpretation and despite the numerous examples in the SDTM Implementation Guide; there aren’t enough to cover every possible type of study design. For example, should there be individual elements defined for each change in the dosing regimen during the treatment period or just a single element to cover the whole treatment period? Both can be valid definitions.

The setup in the Trial Design domains is applied to subject level data across multiple SDTM domains. Variables such as Description of Planned Arm [DM.ARM] and Epoch [EPOCH] which contain information derived from the Trial Arms [TA] domain were therefore expected to contain differences too.

SDTM VERSIONS

Finally, working with data that has been mapped to different versions of SDTM may lead to expected differences. For the two studies in question, however, this was not an issue as it was ensured they were both being mapped to the same version of SDTM by the vendors to reduce the problems when pooling the data. The version selected was the version of the latest SDTM standards available at the start of the study, which was SDTM v1.2, including Amendment 1.

UNEXPECTED DIFFERENCES BETWEEN STUDIES

It has been established that there were differences to be expected between the sets of SDTMs delivered from the two vendors but the statistical programming team also encountered differences that were unexpected upon viewing the SDTMs for the first time. Most of these differences could be explained by two simple deductions.

SUBJECTIVE INTERPRETATIONS OF THE MODEL

The SDTM implementation is still open to a certain amount of subjectivity by the people assigned to create the SDTM datasets, predominantly in relation to any areas of the creation that involves performing derivations on the collected data. Considering one of the primary aims of the model is to standardise the tabulation of study data across the industry, it was not perceived that two different companies could take the SDTM standards and Implementation Guide and produce any variations in the enactment. A simple example of areas open to interpretation are the Trial Design domains, as mentioned previously, or how to construct the exposure domain [EX], as this is a domain derived from the collected exposure data.

At times, the implementation of the model requires input from an analysis perspective and this input from the statistical programming team was not sought by our vendors. Thus, differences existed between what was used as the Reference Start Date [DM.RFSTDTC] on each study. Reference Start Date is designated as study day 1 and so is the starting point for all study day [--DY] calculations, but the definition of what determines ‘study day 1’ is left to the clinical study teams. We found that one study used the date of randomisation as the Reference Start Date, whereas another used the date of first treatment. Unresolved, this would be problematic for pooling the SDTMs, as the definition and study day derivation should be consistent across all the pooled data. Another similar example relates to baseline flags [--BLFL] on the SDTMs. Again, this variable needs to be derived and indicates any baseline record within the domain. The definition of ‘baseline’ was subject to interpretation by those producing the SDTMs and so resulted in discrepancies between the SDTMs we were delivered.

MAPPING DECISIONS

The second reason for the differences, and linked somewhat to the first point, was that there will always be certain mapping decisions that have to be made when implementing the model. There are many cases when the data collected on a study does not have an obvious mapping to a standard or sponsor defined domain. Take, for example, deaths in a clinical study. A death is defined within the SDTM model as an outcome of an adverse event, rather than an event in itself and the SDTM model has been based on this assumption. Therefore, when a study collects deaths as an event instead, there is no clearly defined domain for this death data; it may justifiably reside in the Adverse Events [AE], Disposition [DS], Clinical Events [CE] or even a sponsor defined domain. There is no right or wrong way to map the data which will mean different vendors may take different approaches with the implementation.

AUTOMATED CHECKING TOOL DESIGN

It was known that the data for both studies would be pooled and so the extent of the differences between the two sets of SDTMs needed to be identified by the programming team. Initially, the SDTMs were being manually checked by means of visual comparison but it soon became clear that this would provide neither a thorough enough nor complete list of all the discrepancies. Hence, the development of an automated checking tool was required.

REQUIREMENTS

The purpose of the tool was to return a list of all the discrepancies between the two sets of SDTMs. It was not required for the tool to determine whether the differences were valid or not, as this assessment would be performed by a reviewer of the tool output. It is important to note at this point that the program would solely focus on the consistency of data tabulation and not whether it is mapped correctly, which is more the purpose of a conformance tool such as OpenCDISC. The tool needed be written in SAS® and developed within a restricted timeline. There was also a requirement for the program to be dynamic so that it was capable of working with any number of SDTM domains, allowing the tool to be used again on future studies without producing difficult to debug errors. Finally, the program had to be simple to use so that anyone without SAS knowledge could run and use the tool.

The design of the tool would focus on checking three key areas, based on three levels of granularity; the domains themselves, the variables within the domains and finally, the data point values within the variables of the domains.

SETUP

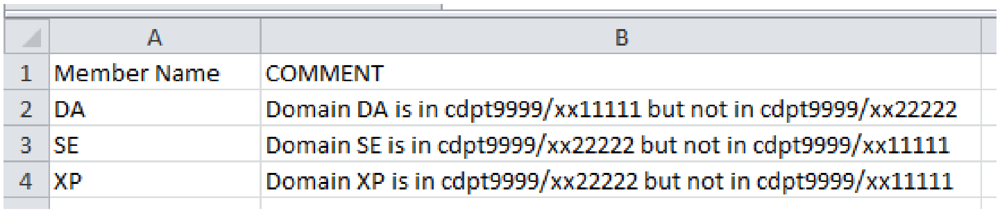
The tool had to be simple to use and dynamic. It was decided that all the user should need to input is the library location of the 2 sets of SDTMs. The tool would then determine which datasets and variables existed in each area and find any differences. Achieving this meant that there would be no hardcoding of terms or expected domains within the program, but instead the program would refer to the environment metadata at each stage of the checking to gather this information. The dictionary view sashelp.vcolumn was utilised throughout to find the domains that were present and also what variables existed on these domains.

The program was developed as a macro so it could be stored as a package of code at a central location and then called by anyone who wished to run the processing. Two library locations were the only input to the macro, ensuring independence from the data it was referencing. The macro could be stored anywhere it could be called by a SAS session or program.

After each stage of the consistency checks, the program outputs a CSV file of the findings to the location the macro was called from, ready for review.

DOMAIN CHECKS

The first level of checks performed by the tool highlighted any differences in the SDTM domains between the studies. This was a simple assessment on the existence of the domains between the two studies with no specific checking of the domain metadata or content. To achieve the list of differences, the dictionary view sashelp.vcolumn was used to create two datasets containing the names of the SDTM domains in each library. These two datasets were then merged by dataset name [MEMNAME] and where the SDTM domain only existed in one of the libraries, a comment about this discrepancy was output to the CSV file. An example of the output CSV file is below:

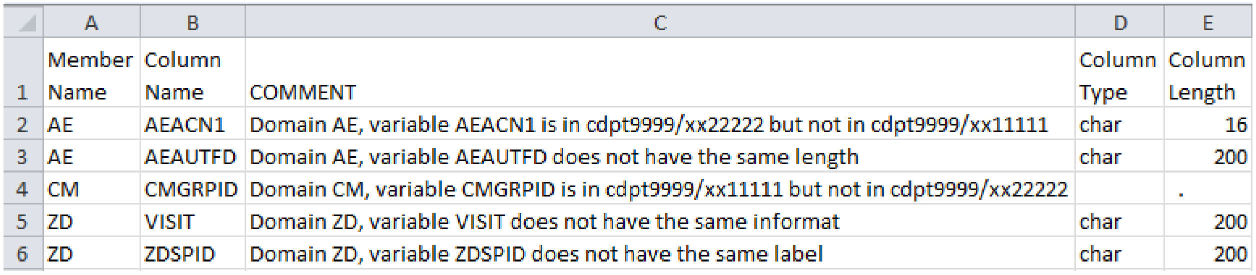


	A	B
1	Member Name	COMMENT
2	DA	Domain DA is in cdpt9999/xx11111 but not in cdpt9999/xx22222
3	SE	Domain SE is in cdpt9999/xx22222 but not in cdpt9999/xx11111
4	XP	Domain XP is in cdpt9999/xx22222 but not in cdpt9999/xx11111

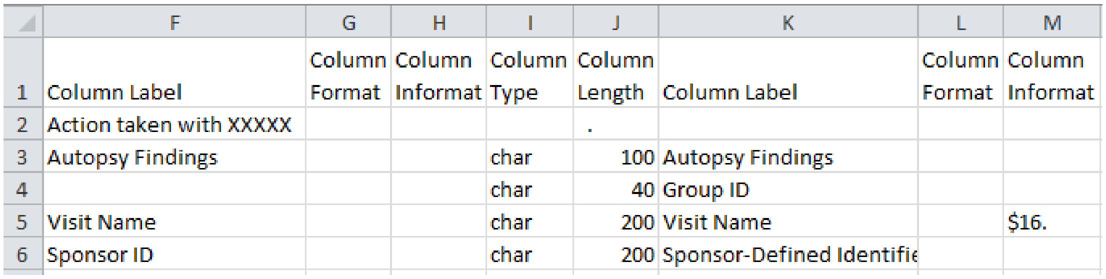
VARIABLE CHECKS

The next step of the program was to find the differences in the variables that existed in the two study areas. Obviously, if a domain existed in one library but not the other, then there will be a mismatch on all the variables in that domain. The reviewer has already been notified that the domain was missing in the previous set of checks and should seek to resolve this potential issue first. Therefore, the tool only performed the next stage of checks on SDTM domains that were present in both libraries.

After identifying the total number of matching domains and assigning each domain name to a macro variable, the tool loops through each domain and performs a similar merge to the first step, but on the variable name [NAME] instead. This enables the identification of variables that only exist in one of the libraries. At the same time, the attributes of variables are also compared to determine if there are any differences and this further check was only performed on common variables; those that existed on the common domains of both the studies. An example output from the program is shown below:



	A	B	C	D	E
1	Member Name	Column Name	COMMENT	Column Type	Column Length
2	AE	AEACN1	Domain AE, variable AEACN1 is in cdpt9999/xx22222 but not in cdpt9999/xx11111	char	16
3	AE	AEAUTFD	Domain AE, variable AEAUTFD does not have the same length	char	200
4	CM	CMGRPID	Domain CM, variable CMGRPID is in cdpt9999/xx11111 but not in cdpt9999/xx22222		.
5	ZD	VISIT	Domain ZD, variable VISIT does not have the same informat	char	200
6	ZD	ZDSPID	Domain ZD, variable ZDSPID does not have the same label	char	200



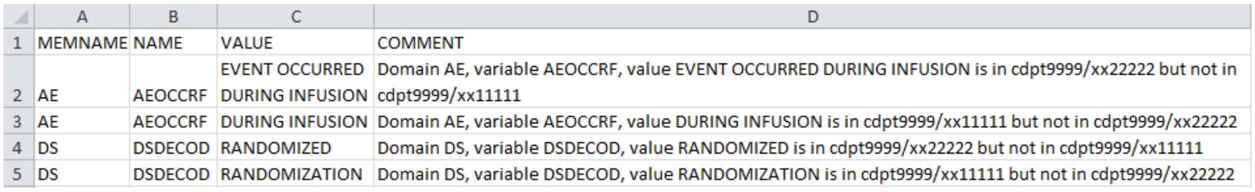
	F	G	H	I	J	K	L	M
1	Column Label	Column Format	Column Informat	Column Type	Column Length	Column Label	Column Format	Column Informat
2	Action taken with XXXXX				.			
3	Autopsy Findings			char	100	Autopsy Findings		
4				char	40	Group ID		
5	Visit Name			char	200	Visit Name		$16.
6	Sponsor ID			char	200	Sponsor-Defined Identifie		

The diagram above highlights a couple of cases when one variable was present on an SDTM domain for a study and not the other, such as the AE.AEACN1 and CM.CMGRPID variables shown. It also displays variables that were present in both SDTM domains, but had differing attributes. This is the case for variables AE.AEAUTFD, ZD.VISIT and ZD.ZDSPID, which mismatch on variable length, informat and label, respectively.

VALUE CHECKS

The final and most demanding step of the program is the consistency check on the values within the studies. This is problematic because there will be expected differences within the data and also variables that contain data that will never match between studies. Hence, a hardcoded list of variables to be excluded from the check was incorporated. This included subject identifier variables [USUBJID], any date variables ending in –DTC and ID variables such as those ending in –SPID and –GRPID, to name a few. As it was likely that this list was not exhaustive, a macro parameter was added to further exclude a list of user-specified variables from this check. Also, all numeric variables were removed. The intention was that this check would help to identify any differences in the use of controlled terminology without having to enter all the code lists.

This value check stage of the program is only performed on common variables on common domains in the two libraries. The processing works by looping through the common domains with a nested loop through the common variables in that domain. For each variable, unique values from both libraries are then merged together with a library identifier for where the unique value came from. If the value only existed in one of the libraries, then it was output to the exported CSV file. An example output from the tool is below, which depicts one of the cases found where the vendors deviated in their code list extensions.



	A	B	C	D
1	MEMNAME	NAME	VALUE	COMMENT
2	AE	AEOCCRF	EVENT OCCURRED DURING INFUSION	Domain AE, variable AEOCCRF, value EVENT OCCURRED DURING INFUSION is in cdpt9999/xx22222 but not in cdpt9999/xx11111
3	AE	AEOCCRF	DURING INFUSION	Domain AE, variable AEOCCRF, value DURING INFUSION is in cdpt9999/xx11111 but not in cdpt9999/xx22222
4	DS	DSDECOD	RANDOMIZED	Domain DS, variable DSDECOD, value RANDOMIZED is in cdpt9999/xx22222 but not in cdpt9999/xx11111
5	DS	DSDECOD	RANDOMIZATION	Domain DS, variable DSDECOD, value RANDOMIZATION is in cdpt9999/xx11111 but not in cdpt9999/xx22222

TOOL BENEFITS

By programmatically automating the comparison of the two sets of SDTMs, there has been a definite time saving over manual checking of the SDTM deliveries. When manually performing these checks, time is spent acknowledging that two ‘elements’ are the same. For instance, that a certain variable or controlled term is present for both studies. The tool disregards these similarities and provides a report solely based on the discrepancies, allowing the reviewer’s time to be focused only on the differences between the studies. There are some limitations with this approach, however, as the program is performing structural checks on the datasets and is not intelligent enough to ascertain whether a variable should be present or not. It could be the case that the Visit variable is incorrectly on the Adverse Event domain for both studies, but the program would not flag this as erroneous. It works on the principle of ‘majority rules’ whereby if an element is present in both studies then it assumed that it should be there.

The main benefit of using a tool like this is that the resulting sets of SDTMs, after all issues are resolved, will be more consistent between studies. This has a positive impact on many programming areas, but one of the key aspects is that of pooling data. Datasets with identical metadata, whether it is the variables on those datasets or the attributes of the variables themselves, can be easily stacked without any pre-processing or concern around potential value truncation due to a variable length mismatch. By resolving the discrepancies from the tool, the reviewer is almost unknowingly improving the ease of pooling the SDTMs.

Aside from making the datasets from different clinical studies easier to combine, it goes one step further by enabling pooling strategies to be considered at the very start of the study. The tool can be run as soon as the two studies that are likely to be pooled have data available in SDTM format. More often than not, pooling strategies for clinical studies are an afterthought and are considered after the individual study analyses has been performed. It is at this stage that the SDTMs for the study that were reported on can no longer be modified without potential impact on the study program and analyses already reported. Hence, the programming to align the SDTMs between studies is generally performed in a pre-processing step before the datasets are pooled. By resolving these inconsistencies between studies at an earlier stage in the study lifecycle when the SDTMs are being constructed, the amount of programming effort as part of the pooling strategy is reduced.

The benefits of performing SDTM consistency checks between studies at the start of a study continue. In general, no analysis programs would have been written at this stage, so changes to the SDTMs can be raised and implemented without any concerns for the impact on analysis programs. Further to this, when analysis programming does begin, development of standard programs that operate across studies can be enacted with the confidence that fewer study-specific changes due to SDTM variations will need to be incorporated.

A further benefit of having such a tool has been in relation to the program lifecycle on a study. During study conduct it is likely that analysis programs will be developed whilst data is still being collected and therefore the SDTM datasets are constantly being refreshed with new data. If transfers of the SDTM data are archived, then the consistency tool can be run on sequential transfers of an individual study’s SDTMs. This would allow for any changes in the mapping, possibly due to previously incorrect mapping or new supplemental qualifier data, to be detected and analysis programs can be updated accordingly. This concept can extend beyond SDTM datasets and the same principle could be applied to analysis datasets too, with the tool being used to monitor any changes within or between studies. Additionally, two sets of SDTMs built from two different versions of the standards could also be compared for differences that a statistical programming team should be made aware of.

Building on this, when standard programs have been developed for a number of existing studies and a new study is introduced, this tool can be used to identify where the standard programs need to be modified for this new study. Say a standard program has been developed to analyse laboratory data and given previous studies, the Method of Test [LB.LBMETHOD] terms accounted for included ‘PHOTOGRAPHY’. The tool would then detect if any new terms are present in the new study being introduced, such as ‘PHOTOGRAPH’. It would then be evident that any programs using the LBMETHOD variable would need to be re-evaluated for this new term and potentially updated. This will be a likely scenario for any variables with extensible code lists, such as the collection method variable [--METHOD], as vendors can freely add to the code lists based on the available data collected. Changes in variables can also be detected by the tool. For example, the new study may allow for investigators to record multiple race categories, so the Race variable [DM.RACE] contains the valid term ‘MULTIPLE’ and the applicable race categories are mapped to the supplemental qualifier under the parameters ‘RACE1’ and ‘RACE2’. As long the supplemental qualifiers are combined with the base SDTM as a prerequisite to the consistency tool being run, then the detection of the new variables RACE1 and RACE2 should be present in the tool findings, thus provoking an update to any analysis programs that handle the RACE variable.

CONCLUSION

Planning for pooling activities can never happen early enough and it would be advisable to start checking for the consistency of the data tabulation between studies as soon as SDTM datasets are received. Data that is mapped consistently is not only easier to combine, but also can lead to better development of standard analysis programs across studies.

Production of a tool to programmatically check the consistency of data tabulation would be advisable and a basic tool can be achieved within a relatively short time, thus providing a means of ensuring the data is fit for reporting purposes. Once developed, this consistency tool can be more advantageous to statistical programming teams than just ensuring SDTMs are consistent between clinical studies. By knowing the structural differences between datasets of existing and new studies, standard analysis programs can quickly be modified to accommodate the new studies. Also, the tool can be run on just a single study and used to compare subsequent transfers of SDTM datasets as data is refreshed, so any changes in the SDTM structure can be acknowledged and associated changes implemented in the study's analysis programs quickly and efficiently, in a targeted manner.

The program was developed in response to identifying discrepancies in the SDTMs delivered by two separate vendors but there are deeper questions around *why* such discrepancies should exist. Given two similar studies for the same investigational drug and indication as well as a well-documented, relatively mature tabulation standard such as SDTM, differences in the mapping should not exist.

The exact reason may never be known, but it is certainly recommended to never assume that the SDTMs delivered by an in-house or outsourced transformation team will be structurally identical. For most part, the SDTM model is robust and does guarantee standard mappings but certain aspects still require further education and clarification. This is especially true for any derived variables as well as any non-standard, study-specific data that is collected and must be mapped into the model. Anywhere where the model is open to interpretation it is also open to deviations in the mapping decisions.

REFERENCES

Study Data Tabulation Model v1.2
Study Data Tabulation Model Implementation Guide v3.1.2
(www.cdisc.org/sdtm)

ACKNOWLEDGMENTS

I would like to thank:

- Alex Hughes and Yvette Baptiste for reviewing this paper.

CONTACT INFORMATION

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

Nathan James
Roche Products Limited
6 Falcon Way
Shire Park
Welwyn Garden City / AL7 1TW
United Kingdom
Work Phone: +44 (0)1707 36 5926
Email: nathan.james@roche.com