Paper RS04

Experiences Submitting CDISC SDTM and Janus Compliant Datasets

Carol Vaughn, Gregory Ridge, and William Friggle

sanofi aventis

ABSTRACT

The primary resource for developing datasets compliant with the Clinical Data Interchange Standards Consortium (CDISC) Standard Data Tabulation Model (SDTM) is the CDISC SDTM Implementation Guide (SDTMIG). With the wide range of strict rules, a sponsor has much work to ensure complete compliance. However, mere compliance with the SDTMIG is not sufficient to ensure the data will be able to be loaded into the FDA's Janus data warehouse – the SDTM standards based clinical data repository which is the future destination of all submitted SDTM datasets. As part of the Janus loading process, an application within Janus runs "rule violation" checks on the datasets. Beyond straightforward checks for compliance with the SDTMIG, there are many other checks that will produce rule violations even with strict SDTMIG compliance. Noncompliance can range from dirty data problems, to failure to meet undocumented requirements such as the need for certain controlled terminology or the prohibition of adverse event coding other than MedDRA. Janus categorizes these rule violations as Low, Medium, or High. Violations that are classified as High will prevent the data from successfully loading into Janus. This paper details the sanofi-aventis experience in the development of SDTM submissions which are not only CDISC SDTM compliant, but Janus compliant as well.

INTRODUCTION

This paper will provide a background of the current regulatory environment, provide information about the FDA's Janus data warehouse, and describe the rule violation checks run by the Janus smart tools application, WebSDM™. The paper will then give a background of the sanofi-aventis SDTM submission which initially failed the Janus load attempt. The rule violations triggered will be described, as well as the measures taken to resolve the violations (if appropriate). Also covered will be the programmatic method used to make the revisions, the iterations of loading the revised data into a trial version of WebSDM™, the Define.xml revisions made, and our collaboration with Phase Forward's Lincoln Technologies safety group, the developers of WebSDM™, and the FDA.

BACKGROUND OF REGULATORY ENVIRONMENT

In December of 2006 the FDA announced their intention to make compliance with the CDISC submission data standards required by regulation. Currently, the CDISC SDTM submission data standard reference documents consist of: the *Study Data Tabulation Model v1.1* - the underlying conceptual model behind the submission standards - and the *Study Data Tabulation Model Implementation Guide: Human Clinical Trials v3.1.1* - which include detailed domain descriptions, assumptions, and examples.

In accordance with the 1999 *Electronic Submission (eSub) guidance* and the *Electronic Common Technical Document (eCTD)* documents, all SDTM submissions must be accompanied by a data definition document. As a replacement for the traditional Define.pdf data definition document, a machine-readable Define.xml format may be submitted. The *Case Report Tabulation Data Definition Specification (Define.xml) Version 1.0* published by the CDISC team specifies the standards for providing SDTM Data Definitions in XML format.

All SDTM data and metadata submitted to the FDA will eventually be loaded into the FDA's Janus data warehouse. Janus is a standards-based data warehouse with reviewer-centric "smart tools". In order for the SDTM data to properly load into Janus, the metadata must be in XML format (i.e. the data must be accompanied by a Define.xml). The smart tools within Janus provide a means for validation checking of the data and Define.xml, reviewing the data, producing spontaneous reports, performing cross-study analyses, and facilitating communication of conclusions. A key smart tool used by Janus for these purposes is the WebSDM™ application. WebSDM™ is a stand-alone application that can be used outside Janus. While violation rules imposed via WebSDM™ within Janus almost entirely overlap with the WebSDM™ stand-alone application, they assign different severity levels to rule violations. This paper will focus only on the rules and severity levels imposed via WebSDM™ within Janus.

JANUS VALIDATION

Janus validation checking consists of checks of both the datasets and the Define.xml file. Janus does not merely check to ensure that the data conforms to the assumptions of the SDTMIG and the Define.xml conforms to the CDISC Define.xml schema. There are also checks for the existence of dirty data, that the Define.xml corresponds to the data in the domain datasets, that values of particular variables are among the expected controlled terminology, and that there is cross-domain consistency between and among the domain datasets.

A particular validation check may apply to all domains, a general class of domains, a particular domain, a set of variables in a given role (such as timing variables), or a particular variable. The list of anomalies produced by the validation checks are termed "rule violations". Janus assigns rule violations a severity of High, Medium, or Low. The severity is meant as an indicator of potential problems or anomalies in the data and the error's potential to affect the interpretation or use of the data for specific purposes. If any one of the rule violations that the FDA has identified as having a High severity is triggered, the study will fail to load into Janus. Violations flagged as Medium are considered to impact the reviewability of the submission. While violations flagged as Low are considered to only possibly impact the reviewability of the submission.

The actual list of checks used by the FDA, which now consist of approximately 105 checks, will evolve over time to reflect input received from the FDA reviewers on the basis of accumulating experience with SDTM data submissions.

The following are some examples of checks (The Janus severity appears in parentheses following the violation):

<u>Data Does Not Conform to the SDTMIG</u>

- Length of __TEST is greater than 40 characters (Low severity)
- __ORRES is null, but __STAT in not equal to "NOT DONE" (Low severity)
- Null value found where the core attribute is "Required" (Medium severity)
- Mandatory domains or fields are missing (High severity)
- A non-unique sequence number exists for a subject (High severity)

<u>Define.xml Does Not Conform to CDISC Define.xml Schema</u>

- Required XML field (such as the domain level 'Repeating' field or the variable level 'DataType' field) is missing (High severity)
- Value of required field is not valid (For example, valid values for the field DataType are: text, integer, float, date, time, or datetime) (High severity)

<u>Dirty Data</u>

- End date is provided but start date is missing (Low severity)
- AGE/CMDOSE/__DUR is a negative number (High severity)
- End date is prior to start date (High severity)

The Define.Xml Does Not Correspond to the Data

- Value for variable not found in codelist (Low severity)
- Variable in description file not in dataset (or the reciprocal of this) (High severity)
- Data type in description file does not match data type in dataset (High severity)

Values are Not Among the Expected Controlled Terminology

- Value for AGEU is not YEARS, MONTHS, DAYS, HOURS, or WEEKS (High severity)
- Value for SEX is not M, F, or U (High severity)
- Value for __STREF is not BEFORE, DURING, AFTER (High severity)
- Value for most YES/NO type variables is not Y or N (High severity)

Cross-Domain Inconsistency

- No DS/EX domain record found for a subject with data in the DM domain (Low severity)
- Visit number in SV not found in TV (Low severity)
- A value for ARM or ARMCD existing in a domain is not found in the TA domain (Medium severity)
- A value for IETESTCD is not found in the TI domain (High severity)

BACKGROUND OF SANOFI-AVENTIS SDTM SUBMISSION

In 2005 sanofi-aventis made the decision to submit SDTM datasets as part of a Phase III study submission and notified the FDA of its intention. The FDA accepted this offer and approximately two months prior to the submission date, requested that all the Phase II studies conducted for the drug also be submitted. The Phase II studies were quite old, the data management had been done externally, and their formats differed greatly from each other. Clearly the FDA would only be able to properly review the Phase II data if it were standardized across studies. We decided to take advantage of the SDTM format of the Phase III study and map all the Phase II studies to SDTM as well.

While the Phase III SDTM datasets were prepared in-house, there was insufficient time to undertake the mapping of all the Phase II studies to SDTM. Phase Forward's Lincoln Technologies safety group (the developers of WebSDM™) was contracted to map the Phase II studies and prepare the Define.xmls.

Approximately one year after the dossier package was submitted and approved, while requesting feedback on the submitted SDTM datasets, sanofi-aventis was notified by the FDA that the SDTM datasets failed to load into the Janus pilot. We learned at this time that no studies submitted to the FDA at that point had successfully loaded into the Janus pilot environment. The FDA requested that sanofi-aventis explore the reasons why the datasets would not load and revise the data in order to enable it to load. To this end, they provided the rule violations produced by WebSDM™ in support of the Janus loading process.

The approach we took was to revise the data so that not only would it load into Janus, but so that it would no longer trigger any lower level violations - provided that revision was a reasonable recourse.

SANOFI-AVENTIS RESUBMISSION EFFORTS
We consider the rule violations incurred for the sanofi-aventis submission as being able to be grouped into the following categories:

- Caused by Janus bugs
- Caused by Janus strict violation rules
- Caused by the SDTMIG
- Caused by the data
- Caused by nonconformity with the standards

For each of these categories, a general description of the types of violations which fell into that category will be provided and the rule violations incurred will be listed with details about the violation. The details will include:

- the reason the violation was triggered,
- the Janus severity classification (in parentheses), and
- the fix/workaround used or reason why no fix/workaround was considered reasonable.

Caused by Janus Bugs
The Janus rules imposed via WebSDM™ are in their first release; there are minor bugs which will be rectified as the rules evolve. Below is a violation which was caused by a Janus bug and the workaround by which the bug was circumvented:

- Value of “U” (Unknown) was flagged as unacceptable for Medical History Occurrence (MHOCCUR) (High severity)

  The STDMIG states that all Yes/No controlled terminology can be extended to “U” if it was captured on the CRF.

  Workaround :
  - Delete the value of “U”
  - Fill Medical History Status (MHSTAT) with "NOT DONE“
  - Fill Reason Medical History Not Collected (MHREASND) with "CRF CAPTURED RESPONSE = 'U' (UNKNOWN)"

Caused by Janus Strict Violation Rules
Not all of the violations triggered by the tool are intended to signify that there is necessarily a problem with the data. There are many strict violation rules in place in order to alert to the possibility of a problem. However, when these rules trigger frequently when there is no problem with the data, their usefulness must come into question.

Below are violations which were caused by strict violation rules:

- No baseline result (Low severity)
  - No fix possible – For some domains, even though baseline readings were scheduled, for some subjects a reading prior to study drug administration was never done. In the case of Efficacy Findings (EF) and some Questionnaire Findings (QS) domains, by study design, all findings were intended to be captured after the start of study drug, so it is not possible for any of these findings to have a baseline result.

- AE is serious but none of the qualifiers are set to “yes” (Low severity)

    - No fix possible - Due to the study design, none of the qualifiers on which the Serious flag (AESER) is based were captured in the study. For example, none of the Phase II studies captured qualifier data such as: Is Life Threatening, Involves Cancer, Requires or Prolongs Hospitalization, etc. It would be unacceptable to remove a serious flag as judged by the investigator.

- Numerically coded qualitative findings values in the original result (__ORRES) which decode to text values in the standardized result (__STRESC) were flagged as missing a unit (__ORRESU). For example, when VSTEST = “KILLIP CLASSIFICATION”, a VSORRES of 3 is a code for the VSSTRESC of “KC3“ (a Killip Classification of III) and, therefore, a unit of measure is not applicable. (Medium severity)

    - Workaround - Added a unit of “.” for such values with an indication in the Define.xml codelist that “.” decodes to “Not applicable”.

- No exposure record (Low severity)

    - No fix possible - Demographics data was captured, but the subject was never dosed.

Similar types of data issues, which are not indicative of problems with the data or noncompliance with SDTM, are typical and unavoidable in large trials. This is an area where the utility of such violation rules needs to be evaluated.

Caused by the SDTMIG

Violations caused by the SDTMIG are due to either imprecise requirements or the fact that too much variability is allowed. The SDTMIG does not cover all possible scenarios of study design or data situations. Also, the SDTMIG does not require specific controlled terminology for some of the variables for which Janus does require specific controlled terminology, and it does not require the use of particular coding dictionaries while Janus has specific requirements. Below are violations caused by the STDMIG:

- A Description of Planned Arm (ARM) of “NOT TREATED” was used for subjects who did not fail screening but were not treated for other reasons, however, Janus expects controlled terminology of “SCREEN FAILURE” for these subjects. (High severity)

    The STDMIG specifies that data for screen failure subjects should be submitted with ARM = “SCREEN FAILURE”. It does not specify what ARM should be for subjects who did not fail screening but were not treated for other reasons.

    - Workaround – Even though these subjects did not fail screening, values of “NOT TREATED” were changed to “SCREEN FAILURE” and the comment in the Define.xml for the variable explains that these subjects did not fail screening.

- AE Outcome (AEOUT) of “DIED” (as captured on CRF) was used, however, Janus expects controlled terminology of “FATAL” for these AEs. (Low severity)

    The STDMIG lists “FATAL” as an example of possible values, not as required controlled terminology.

- Fix - Changed values of “DIED” to “FATAL”

- The Phase II study AEs were coded to COSTART, however, Janus expects MedDRA coding of AEs. (High severity)

  The STDMIG specifys that AEs must be “coded using a standard dictionary such as MedDRA”, but does not indicate that MedDRA coding is required.

  - Fix - Recoded to MedDRA

Caused by the Data

Violations caused by the data are triggered by dirty data. Dirty data has two causes. The first is typically due to the investigator making an incomplete or incorrect entry on the CRF – or failing to make any entry. The second cause is due to incorrect data entry. The existence of at least some dirty data persisting after database lock, especially for large studies, is virtually unavoidable. While it is typically unacceptable to programmatically alter such data (sometimes referred to as “hard coding”), we considered it acceptable to drop dirty data from the submission if, due to its dirtiness, it is not useful.

Below are the violations which were caused by dirty data for which it was necessary to programmatically alter the data or it would not be able to be loaded into Janus:

- End date-time preceded start date-time (High severity)

  There is no way to determine for sure which date is incorrect, so the dates are not considered useful.

  - Fix 1 - If the date portions were conflicting, both dates were deleted as well as their associated timing variable values

  - Fix 2 - If only the time portions were conflicting, both time portions were deleted

- Times with no dates (High severity)

  - Fix - these “date-times” were deleted as they were not useful

For violations of lower severity, for which dropping of data would result in the loss of information that might be of interest to a reviewer, we did not consider it acceptable to drop the data. Below is the violation which was caused by dirty data for which it was not considered reasonable to revise the data:

- Start date-time was missing when and end date-time was provided (Low severity)

  - Not Fixed - There is no reason to assume the end date-time is not correct. The only fix would be to drop this date. It was not dropped because the date may be of interest to a reviewer.

Caused by Nonconformity with the Standards

Violations caused by nonconformity with the standards are caused by overlooked compliance with the SDTMIG or failure of the Define.xml to agree with the data. It should be noted that the Define.xml must correspond to the data exactly. Even capitalizations and spacing of variable labels and values in codelists must match the data exactly. The rule violations in this category are very useful in pointing to the need for revision by the sponsor. Below are the violations which

were caused by nonconformity with the standards and the method by which the by nonconformity was fixed:

- Unscheduled lab dates from Lab domain were not in Subject Visit domain (Low severity)
    - Fix - added unscheduled dates to Subject Visit domain
- Lab results for Phase II studies were never standardized. The standardized result variables contained original result variable data (with varying units per parameter) (High severity)
    - Fix - derived standardized lab results
- "Other, Specified" values retained in domains (Low severity)

  This was identified as a rule violation because the codelist in the Define.xml had only "Other" but the domain had actual values.
    - Fix – moved these free-text values to supplemental qualifier datasets
- Codelist in Define.xml was discrepant with values in the domain (Low severity)

  This was almost entirely a problem of case sensitivity and/or spaces.
    - Fix - corrected the define.xml
- Both the end relative to reference date (__ENRF) and end date (__ENDTC) were null (Low severity)

  This occurred when no end date was provided and there was no variable capturing if the finding/event/intervention was ongoing.
    - Fix - Since there is no way to determine whether the end date was before, during, or during/after, __ENRF was populated with "U" (Unknown).

Since the initial programmatic mapping to SDTM for most of the studies was done by Phase Forward's Lincoln Technologies safety group, the only method by which we could revise the data for those studies was to revise the SDTM datasets already submitted. The programming approach used was driven by the fact that often the same violation occurred across several (or all) studies and sometimes across several datasets within a study. A single SAS® program was written which looped through the studies, conditionally making the needed changes to select datasets.

Arriving at datasets which triggered no High violations, nor any Medium or Low violations for which it was reasonable to revise the data, required a couple iterations of loading the datasets into the sanofi-aventis trial version of WebSDM™.

The Define.xml for each study was then revised to have it correspond to the revised data. Detailed documentation of all revisions to the data and explanations for violations which will still trigger were added to each respective study's Define.xml as variable comments. For example, since the rule violation "Start date expected when end date provided" will still be triggered for one of the studies for the Subject Elements (SE) domain, the variables SESTDTC and SEENDTC have the following in the comment field associated with those variables in the define.xml:

For one subject, the start date of onset of symptoms which precipitate randomization (the date of the MI) is missing, but the date of randomization is available. Therefore, in the Subject Element domain, for the element "Randomization", since the start date (SESTDTC) is the date of MI and the end date (SEENDTC) is the date of randomization, this subject will have an end date, but no start date

While such explanations are not required, we feel that it is important in order to address all possible questions a reviewer may have upon reviewing the rule violation output.

COLLABORATION WITH PHASE FORWARD'S LINCOLN TECHNOLOGIES SAFETY GROUP AND THE FDA

A key component to being able to effectively revise the datasets and Define.xmls was having access to a trial version of WebSDM™. This allowed us to see similar rule violation output to that which the reviewers at the FDA will see. We highly recommend having copy of WebSDM™, since it would seem that this is the best way to mirror what the FDA will be seeing.

Also very helpful was ongoing support provided by Phase Forward's Lincoln Technologies safety group. For example, at one point we were unable to determine why we were still receiving "Invalid Preferred Term" rule violations for values of AEDECOD even though we had recoded to MedDRA and specified the version of MedDRA we used in the Comments field for the variable AEDECOD. Phase Forward's Lincoln Technologies safety group reminded us that we must also update the Define.xml code specifing the external code list dictionary.

Below is an excerpt of this code:

```
<CodeList OID="CL.MEDDRA" Name="MedDRA 10.0" DataType="text">
        <ExternalCodeList Dictionary="MedDRA" Version="10.1"/>
</CodeList>

<ItemDef OID="AE.AEDECOD" Name="AEDECOD" DataType="text"
        Length="200" SASFieldName="AEDECOD" SDSVarName="AEDECOD"
        Origin="Derived" def:Label="Dictionary-Derived Term" def:DisplayFormat="$200."
      <CodeListRef CodeListOID="CL.MEDDRA"/>
</ItemDef>
```

When we informed Phase Forward's Lincoln Technologies safety group of the version of MedDRA we used (version 10.1), they informed us that Janus had not yet been loaded with this version. They offered to work with the FDA to provide them with this version.

While we gained much from being able to rely on Phase Forward's Lincoln Technologies safety group for quick answers and other assistance, they gained from the collaboration by our feedback about the tool and how we used it. Such real life client experiences should prove valuable to them in their ongoing work with the FDA.

The FDA has assigned two reviewers to analyze the newly resubmitted SDTM datasets. We will be working with these reviewers to answer their questions, bring to their attention the types of issues we encountered, and offer our evaluation of these issues. This process is ongoing at the time of this paper.

CONCLUSIONS

The most valuable information we learned from our experience submitting SDTM data to the FDA was that compliance with the SDTMIG is not sufficient to ensure the data will be successfully loaded by or useful to the FDA tools within Janus. For all planned submissions, the industry needs to address both SDTMIG rules and Janus rules. Beyond validation that the data conforms

to the assumptions of the SDTMIG, the rule violations triggered by Janus smart tools also alert to Define.xml that does not conform to the CDISC Define.xml schema, the existence of dirty data, that the Define.xml corresponds to the data in the domain datasets, and that values of particular variables are among the expected controlled terminology.

While all rule violations that prevent a study from being able to be loaded into Janus must be rectified, we feel that it is important to reduce as many of all the other types of errors from triggering as is reasonable. When data is manipulated so as not to trigger a violation, for example to deal with dirty data, we think that it is important to document this. We also think that it is very important to provide explanations for violations that will trigger but for which it is not possible or not reasonable to revise the data.

It is important to keep in mind that the violation rules will evolve over time as the FDA accumulates experiences with SDTM data submissions. Since our submission will be the first to load into Janus, this may well be the inception of that evolution.

ACKNOWLEDGEMENTS
We would like to thank Phase Forward's Lincoln Technologies safety group for all their assistance. Their insights into the rule violation checking and the Janus loading process greatly helped us to understand the process better and facilitated us in revising the SDTM data for re-submission.

CONTACT INFORMATION
Your comments and questions are welcome. Contact the authors at:

Author # 1 Name	Carol Vaughn
Company	Sanofi-Aventis Group
Address	200 Bridgewater Crossings
City, State ZIP	Bridgewater, NJ 08807
Work Phone	908-304-6298
Email	Carol.Vaughn@sanofi-aventis.com

Author # 2 Name	Gregory Ridge
Company	Sanofi-Aventis Group
Address	Building #9, Great Valley Way
City, State ZIP	Malvern, PA 19355
Work Phone	610-889-6321
Email	Gregory.Ridge@sanofi-aventis.com

Author # 3 Name	William Friggle
Company	Sanofi-Aventis Group
Address	Building #9, Great Valley Way
City, State ZIP	Malvern, PA 19355
Work Phone	610-767-3801
Email	William.Friggle@sanofi-aventis.com