SDTM-ETL 4.2: Validating SDTM/SEND datasets
Using CORE (CDISC Open Rules Engine)

Author: Jozef Aerts, XML4Pharma

Last update: 2023-04-10

Table of Contents

Introduction ......................................................................................................................... 1
Features of CORE .............................................................................................................. 2
CORE Implementation in SDTM-ETL .................................................................................. 2
Running CORE in SDTM-ETL ............................................................................................ 3
Conclusion .......................................................................................................................... 9

Introduction

CDISC CORE (CDISC Open Rules Engine) is currently revolutionizing the submission validation world.
Until recently, sponsors were dependent on software from a vendor who had a quasi-monopoly, but delivered validation software of pretty bad quality that was not very user-friendly and that produced large amounts of "false positives", for which it was notorious. CORE has been developed by CDISC by CDISC coworkers and a large number (>100) of volunteers, especially from the standards developments teams, i.e. persons who know the standards very well, and do know the validation rules very well, and understand how to implement them. This in contrast to the usual vendor, who mostly implemented their own interpretation of the CDISC and FDA rules.
The great advantage of CORE is that all the rules are completely transparent, and can be inspected by the users. So in contrast to the old validation software from the mentioned vendor, it is not "black box" at all.
CORE also is "open source", i.e. everyone can download it from the project's GitHub site, and can implement it or integrate it in their own tools. We therefore expect that CORE will be implemented in many applications and by a good number of vendors: these will then compete on software quality, user friendliness, cost of ownership etc.. However, all these vendors will use the same basis validation engine, so that the validation results will essentially be independent of from which vendor one uses the validation software.

As one of the first vendors, CORE has now also be integrated in our SDTM-ETL software. At the moment of writing (April 2023), CORE is already in an advanced state for the CDISC SDTM rules, less for SEND, but also still in development, and new versions released and new rules added almost every week.
Therefore, we have implemented CORE in SDTM-ETL in such a way that new versions of CORE can be installed without needing an SDTM-ETL software update: new versions of CORE that have been tested by us, will be made available to our customers through a website for plugging in into their existing SDTM-ETL installation.
Features of CORE

Besides that all CORE rules are fully transparent, and the CORE engine is free and open source, some features of CORE that are of enormous importance, and a huge improvement over existing validation software are:

- One can select individual rules (one, or a set of) for execution of the validation. With the existing software, it was an "all or nothing", i.e. one could not exclude specific rules for execution. This was especially nerve-wracking during development of the SDTM/SEND datasets: when some datasets were failing, one would get thousands of error messages. With CORE, each individual rule can be switched off or on for execution.
- One can select individual datasets to be involved in the validation, even when one has generated more than one dataset. This allows to concentrate on a single or a few datasets only. With the old software, it was again an "all or nothing". When for example the TS dataset or DM dataset was failing (maybe as it was not developed yet), one would even get "rejection messages", usually causing a lot of panic and/or confusion.
- Some of the rules developed in the past are more like "guidance" or "best practice". So, in some cases, only part of the rule can be made executable. When applicable, CORE always prefers to be rather "under-reporting" than "over-reporting". This means one should not get "false positives" (over-reporting), but that it can happen that an issue is not reported (false negative - under-reporting). CORE has the principle that false positives are worse than false negatives.
- In future, sponsors will be able to extend CORE with their own, internal SDTM/SEND/ADaM rules. This is especially important for quality assurance during the development of the datasets. Exending the validation with own, internal rules was not possible with the old software that is being used by many sponsors, at least not at a reasonable cost.

With all these advantages, it is expected that regulatory authorities will soon switch to CORE as their major validation mechanism. CDISC is now already implementing the FDA rules in CORE, and is in close contact with the FDA and discussing shared governance over the FDA validation rules.

CORE Implementation in SDTM-ETL

As of version 4.2 of SDTM-ETL, the CORE engine software and files are located in the directory "CDISCCORE". There is also a directory "CDISCCOREFiles" which is used during validation to store a copy of the generated XPT files that need to be validated.

This also means that CORE can be run from CLI (command line) starting from the "CDISCCORE" directory. For information how to run CORE from the command line, see the CDISC-CORE GitHub website explaining this (under "Running a validation").

In SDTM-ETL, CORE can be run immediately after the generation of the SDTM/SEND SAS-XPT files. In near future, also an interface with the generation of SDTM or SEND files in Dataset-JSON format will be implemented.
Running CORE in SDTM-ETL

When generating SDTM or SEND files in SDTM-ETL, the following dialog is displayed:

Remark the new checkbox "Perform CDISC CORE validation on generated SAS XPORT files". When it checked, the user will get the opportunity to perform CORE validation on the generated SAS-XPT files.

After generating the SDTM/SEND XPT files, the following dialog will then be shown:
allowing to select which of the generated XPT files must be validated (default: all), and which of the rules must be executed (default: all available).

One can also choose whether the report needs to be generated in Excel format (default) or in JSON format. The latter will become very useful in future for use of the report in other applications, such as "smart" SDTM/SEND viewers.

When one moves the mouse over one of the CORE rules, a tooltip is displayed with some information about the rule:
Even better is to use the button "Explore Validation Rules". This leads to a new window allowing to search for and select rules:

For example, if one only wants to have the rules executed that apply on "Findings" datasets, one first selects the checkbox "Classes Included" and then select "Findings" in the list on the left side (multiple selection allowed):
Like this, one can also include/exclude domains, or filter on a word in the rule description or message. For example, to only select rules about the \texttt{--SCAT} variable:

and then clicking the "Filter" button leads to:
One can then select the rules about --SCAT that one want to be executed by checking the checkbox on the left side.

Clicking the "OK" button then confirms the selection. This e.g. leads to:

![Image of rule selection](image1)

But now let us do a validation with all rules on all generated files.

When then clicking the button "Execute CORE Validation", after about 20 seconds, the user is informed that the validation is ready. For example:

![Image of validation result](image2)

with the files in the D:\temp directory in our case:
and then opening the Excel report with your favorite spreadsheet program, and navigating to the tab "Issue Summary":

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CORE-ID</th>
<th>Message</th>
<th>CORE-ID</th>
<th>Dataset</th>
<th>CORE-ID</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>CORE-000022</td>
<td>At least one of the Sensitive criteria (AESCAN, AESCONS, AESDISAB, AESDTH, AESHOSP, AESLIFE, AESGO, or AESMUE) = &quot;Y&quot;, but AESER = &quot;N&quot; or empty</td>
<td>AE</td>
<td>CORE-0000264</td>
<td>Primary analysis used but AESDGOYS and AESOC are not equal</td>
<td>DM</td>
</tr>
<tr>
<td>PE</td>
<td>CORE-000021</td>
<td>PESTREC should not be blank when either PEORRES is provided or PEEPREL = &quot;Y&quot;</td>
<td>PE</td>
<td>CORE-0000999</td>
<td>Value for PESTAT is populated, when PEORRES is populated</td>
<td>PE</td>
</tr>
</tbody>
</table>

or the tab "Issue Details":

Very important is that, under "Rules Report", the report also displays which rules have been executed, and which have been skipped, for example as the rule is only applicable to a domain for which there is no dataset in the set of generated XPT files (or selection of it):
For example, the rule CORE-000028 (CDISC CG0008) was skipped, as none of our datasets contains the variable --TPTREF or --ELTM. Rules however that are applicable to at least one of our datasets and that ran successfully (not meaning whether a violation was found) are marked "Success".

One can now start using the results of this CORE validation to improve the mappings for the generation of SDTM or SEND datasets.

**Conclusion**

CDISC CORE is a revolution in the area of validation of CDISC datasets for submissions of datasets to the regulatory authorities (and beyond that).

As one of the first software vendors (if not the first), we have implemented CORE in our software, enabling to use CORE from within SDTM-ETL v.4.2. The implementation allows to select on datasets generated, as well as on rules to be executed during the validation process.