PharmaSUG 2014 - Paper DS16

OpenCDISC Validator Implementation:

A Complex Multiple Stakeholder Process

Terek Peterson, MBA, PRA International, USA
Gareth Adams, PRA International, UK

ABSTRACT

The embracing of data standards by FDA following the vision of CDISC of reduced review time of drug applications opened the door for the creation of several tools to ensure conformance to standards. Tools like SAS PROC CDISC, WebSDM ™, and SAS Clinical Standards Toolkit helped industry ensure compliance to the CDISC standards defined as SDTM, ADaM and define.xml. With the introduction of a free conformance engine OpenCDISC Validator, the possibilities of less confusion and more synergies across Sponsors, CROs and FDA was possible. However, the authors would argue the use of this tool has not achieved that goal and has created complex processes between stakeholders that include clinical, data management, programming, sponsor, and FDA; each group having different understanding of the conformance reports. Confounding any implementation are multiple versions OpenCDISC, SDTM, ADaM, and sometime contradicting FDA documentation.

The way out of this confusion is with the implementation of good procedures, communication, and training. This paper will start with an example of waste where a clear process did not exist. It will provide examples of OpenCDISC checks that need to be managed early in the data lifecycle via edit checks to ensure fewer OpenCDISC warnings/errors. Communication and education needs to be in place for non-technical study team members so they can make informed decisions around the output. The paper provides processes to help control duplication of effort at different time points of a clinical trial. Budget considerations will be presented. Discussion and demonstration of example SAS® code will be provided.

INTRODUCTION

Due to the many stakeholders, managing the process of running the OpenCDISC report becomes tricky, dependent on when it is created. Creating the compliance/conformance report early in a study can create uncertainty on the quality of the SDTM conversion because of unresolved data issues and dirty data. Therefore, processes need to be put in place to control what is reviewed and actioned so that duplication of work can be avoided.

From recent experiences, using OpenCDISC validator as a tool for data error identification increased effort and risk to deliverables associated with its use. In general, the validator tool in many cases does not provide specific enough detail for a data reviewer to quickly identify the relevant record, locating this within the EDC system and associated documentation. In addition, as detailed in the guidance section of the rules document, certain conditions can pre-exist within a study design yet will be flagged as errors that 'can exist'. Depending on how the OpenCDISC process is deployed these will either be accepted as is or identified to be further reviewed by data management colleagues.

Regulatory agencies have recently introduced guidance encouraging the industry to adopt a more 'risk orientated' approach to data quality processes, focusing data cleaning and monitoring efforts on identifying errors in the data that matter. These errors can be defined as errors that affect either the integrity of the protocol or potentially endanger the safety of the patient population. This guidance effectively allows the existence of minor data anomalies and issues, which may conflict with the rigorous checks and rules defined by OpenCDISC validator.

By ensuring that the Edit Checks Specifications (ESPEC) document contains equivalent rules as defined in OpenCDISC validator for identifying potential data issues moves this process of checking much further upstream in the process, thus ensuring that when validator is run on clean data these potential issues have either been investigated, resolved or determined to be 'as is' or 'irresolvable'; per normal data cleaning and query resolution processes.

THE EMAIL THAT STARTED IT ALL

Dear Terek,

I'd like to raise the topic of OpenCDISC validator and its implementation into our standard processes. Currently we seem to have a mixed model approach with vastly different opinions of what exactly it should be used for – data cleaning or CDISC integrity and compliance checker. It reared its head this week when a few days prior to database freeze one of clinical data teams was suddenly presented with a sheet of over 2000 rows of OpenCDISC warnings. This spooked the project manager and we've now burnt well over 50 hours reviewing report then looking into the data and data clarification forms to raise the grand total of 1 query.

It's something I think we need to look at pretty urgently and address so that project teams and the project managers are well aware of where this tool fits in and what purpose is it going to serve in delivering high quality studies.

Thanks,

Gareth

REFINEMENT OF THE PROCESS NEEDED

Many processes have been put into place to ensure potential data issues do not surface late in the conduct of a study, particularly around database freeze or lock. When working with multiple groups across both CRO and sponsor, accountability and responsibility of a data point is difficult to pin down and can create confusion when none should actually exist. For example, the CRO may have one group that creates specifications and converts collected data into SDTM then creates the ADaM specifications and datasets. The sponsor has a standards group that approves the SDTM specifications, a traditional data management function that owns the SDTM dataset domains, and a statistical programming team that owns the ADaM. Who should be talking with whom? Establishment of a clear process with a clear communication plan will reduce the interference and noise plus wasted energies. If an organization does not have a clear process internally then it is difficult to impose one on the other.

CHECKS EARLY IN DATA LIFECYCLE

Dear Terek,

As you know we're currently looking to adopt a different approach internally within PRA to best control the back and forth of data issues discovered by programming and our clients when they review the SDTM datasets. By ensuring that the Edit Specifications (ESPEC) document contains equivalent rules as defined in OpenCDISC validator for identifying potential data issues we are moving this process of checking much further downstream in the process, thus ensuring that when validator is run on clean data these potential issues have either been investigated, resolved or determined to be 'as is' or 'irresolvable'. It is this process that we are proposing should be adopted as our standard process.

Here is one of our most recent examples of using OpenCDISC as a data cleaning tool on a final, clean database of 280 patients. The tool created a spread sheet produced by the clinical programming team with 2032 lines of issues to be investigated by the data managers. We spent over 70 hours working through this, identified 2 queries that needed to be sent, neither of which resulted in a data change. Applying the old 'paper sample' error rate formula to this it gives us a potential error of 0.01%. In addition this was for a study in which we hadn't amended the ESPEC document to include the OpenCDISC rules.

Thanks,

Gareth

DOWNSTREAM POINTS OF DATA CHECKING

Waiting until OpenCDISC validator is run on a study's SDTM datasets for the purposes of identifying 'data errors' creates unneeded, duplicated effort and confusion for team members that are not informed or trained on the purpose of the checks. During the conduct of a clinical trial there are multiple data review processes that either operate in parallel or in sequence that are designed to identify data issues that may impact the integrity of the protocol or endanger patient safety, see Figure 1 – Typical Data Review Processes on the next page. During the data review cycles these issues are queried with the investigational sites and either corrected if the data is erroneous or confirmed as correct with the resulting data points left unchanged. The design of these processes and the nature of the data being organic in structure will invariably result in anomalies within the resulting datasets. Of course even with these multiple data review processes there is still a level of risk that a significant erroneous data point is still present upon a clean data transfer or deliverable.

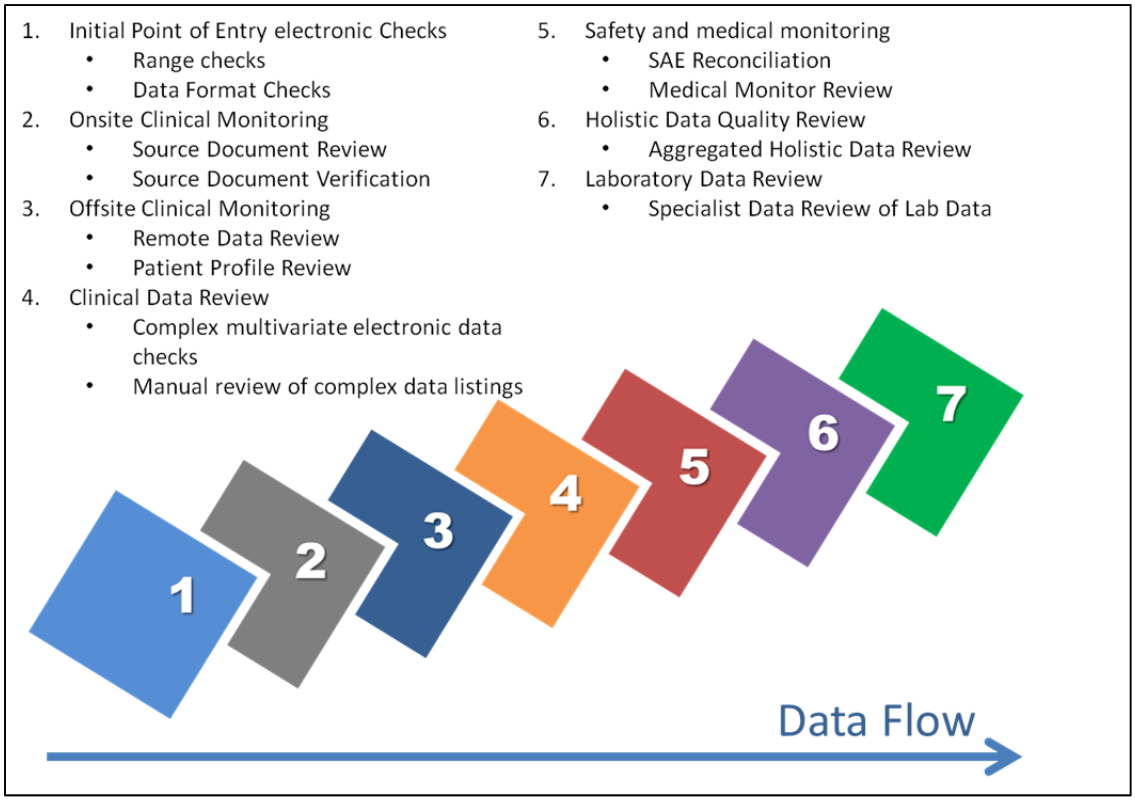


Figure 1. Typical Data Review Processes

Aligning and implementing up front ESPECs based around validator rules, coupled with the multi-level review approaches detailed above considerably reduces the level of risk of significant numbers of resolvable data errors making their way through a clean transfer. This level of risk needs to be offset against the additional effort required to review multiple lines of a validator listing, particularly at points in the study where timeline pressures are typically increased. Data issues resulting from inadequate management of issues at the appropriate point can result in timeline pressures, increase in resources and associated costs for this task. Sponsor and CRO must work together to consider where to include in study budgets and planning tools.

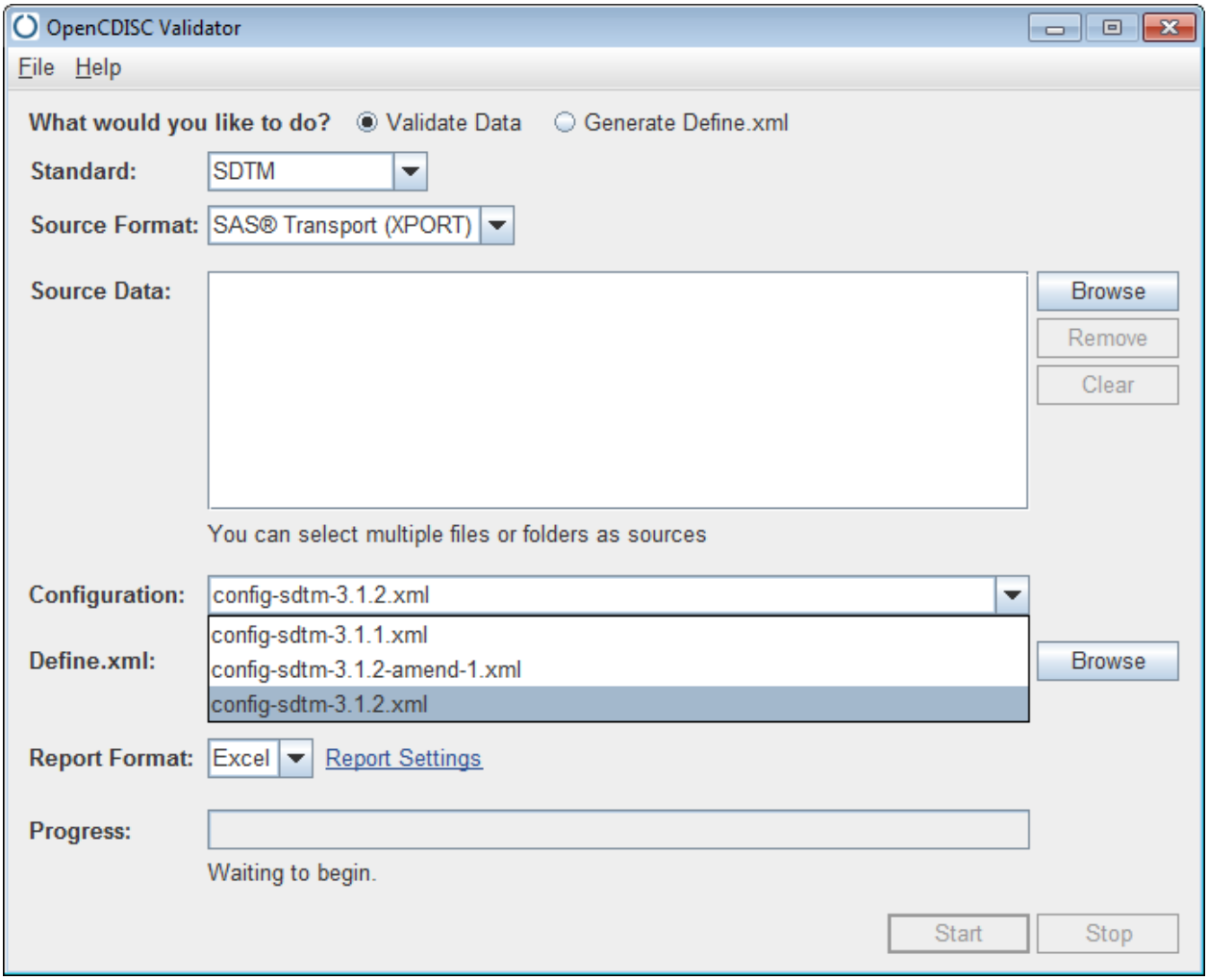
WHAT VERSION OF OPENCDISC VALIDATOR ARE YOU USING?

This seemingly simple question usually comes about when a clear agreement is not made between sponsor and CRO. Expectation should clearly be set early before SDTM programming begins so that there are no surprises when the data and the OpenCDISC report are delivered. It is more then what version of validator is used but what is the version of the corresponding standard, controlled terminology, coding dictionary, configuration file, and validator version.

At the writing of this paper, there have been 7 releases of OpenCDISC Validator. The most stable versions are OpenCDISC Validator v1.3 and v1.4.1. In addition, bewilderingly over the past year and why this discussion is so important is because many companies stabilized their version of SDTM on version 3.1.2 with amendment 1. This seems to be a minor point however the configuration file for this version was deprecated in OpenCDISC tool with release v1.4 and v1.4.1. The version 1.4 flavors of OpenCDISC validator do have an important feature that checks compliance with CDER Common Data Standards Issues Document [1]. However, v1.3 of the tool has the appropriate configuration file for checking SDTM v3.1.2 amendment 1. This is an addition place were complexity exists and misunderstanding of the reports can occur.

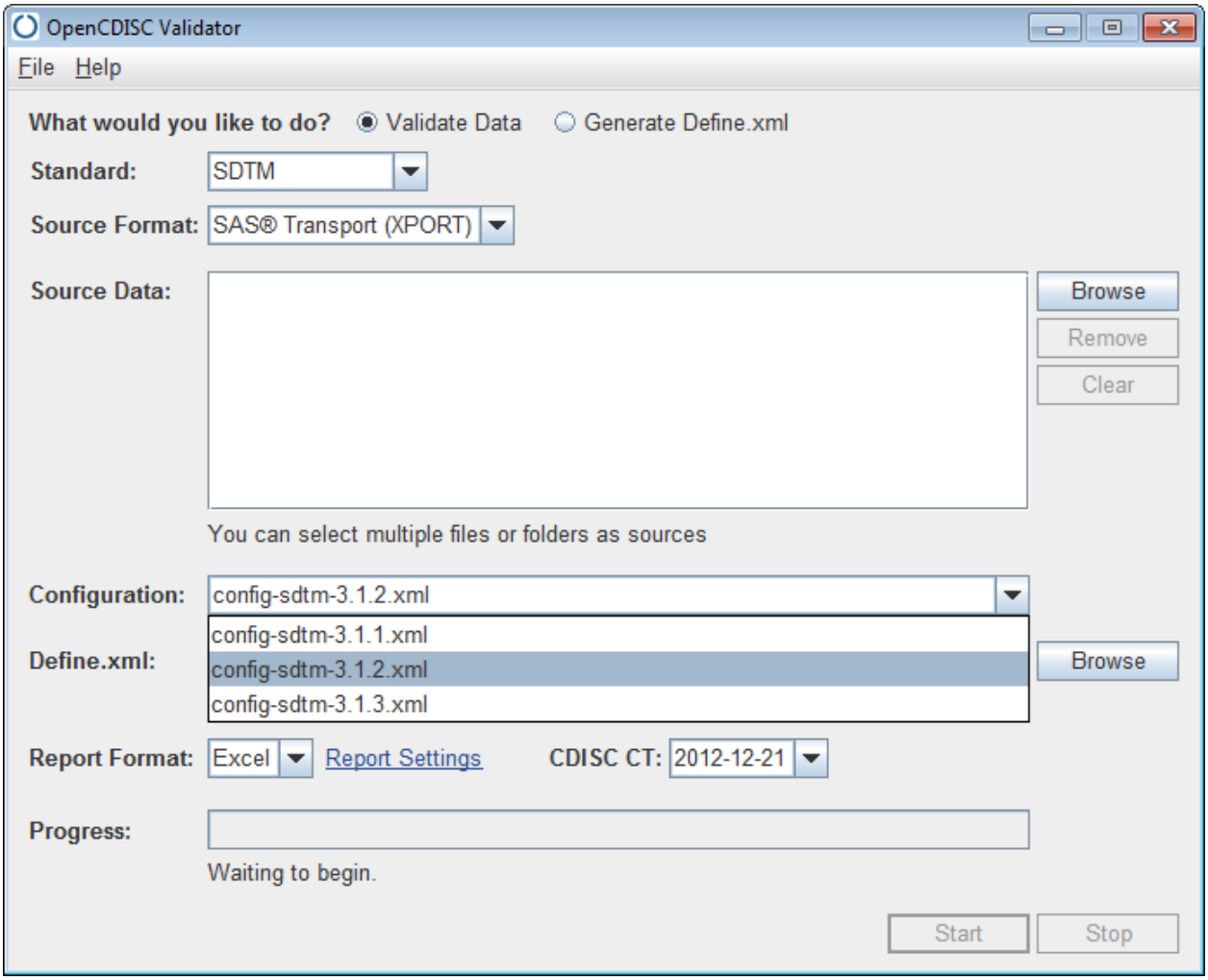
VERSIONS OF OPENCDISC VALIDATOR AND CONFIGURATION FILE

As can be seen in Display 1 below, configuration files for SDTM v3.1.1, v3.1.2 amendment 1, and v3.1.2 all exist in OpenCDISC Validator v1.3. The ability to check CDISC controlled terminology, however, does not exist. The lack of the ability to check controlled terminology and requirements of lengths set by FDA CDER should be considered.



Display 1. Validator Version 1.3

As can be seen in Display 2 on the next page, configuration files for SDTM v3.1.1, v3.1.2, and v3.1.3 exist in OpenCDISC Validator v1.4.1, however, SDTM v3.1.2 amendment 1 does not. The ability to check controlled terminology (see CDISC CT) and requirements of lengths set by FDA CDER are also present within the report created. This creates the specific situation where CRO and Sponsor or Sponsor and FDA could be speaking apples to oranges. Waste is clearly present between stakeholders both internal to a company and between stakeholder companies. Budgets can quickly erode if this simply decision is not discussed and agreed upon.



Display 2. Validator Version 1.4.1

So, what is the right answer to this situation? Unfortunately, there is not one answer that fits all situations. The answer is it depends. It depends how embedded a particular standard is within a company or program of studies. Standard programming and macros can limit a company from being agile and just up-versioning. A study may have been running for several years and it is cost or resource prohibitive to change.

WHAT ARE THE CURRENT SUPPORTED DATA AND DEFINE STANDARDS?

Per the FDA Study Data Standards Resources [2] website under the heading Data Standards Catalogs there is a link to a spreadsheet that clearly defines what standards the FDA supports. Table 1 on the next page is an excerpt of that spreadsheet indicating that the FDA centers of CDER and CBER do recognize STDMIG v3.1.2 amendment 1 as a supported SDTM version. So why doesn't the current version of OpenCDISC validator have this functionality? Some say it is "easy" enough just to up version to v3.1.3 or just use or modify an older configuration file; but these types of activities continue to support the argument that this seemingly simply process is really complex. Communication between sponsor and CRO become more complicated due to the fact that this technical minutia has to be explained to create clarity between multiple stakeholders then to the FDA during submission.

Use	Standard	Exchange Format	Standards Development Organization	Supported Version	Implementation Guide Reference	FDA Center	Date Support Begins (yyyy-mm-dd)	Date Support Ends (yyyy-mm-dd)
Clinical & Non-Clinical Study Datasets								
Clinical study datasets	SDTM	XPT	CDISC	1.3	3.1.3	CBER, CDER	2012-12-01	
Clinical study datasets	SDTM	XPT	CDISC	1.2	3.1.2 Amendment 1	CBER, CDER	2013-08-07	
Clinical study datasets	SDTM	XPT	CDISC	1.2	3.1.2	CBER, CDER	2009-10-30	
Clinical study datasets	SDTM	XPT	CDISC	1.1	3.1.1	CBER, CDER	Ongoing	2015-01-28
Clinical study datasets	ADaM	XPT	CDISC	2.1	1.0	CBER, CDER	Ongoing	
Animal study datasets	SEND	XPT	CDISC	1.2	3.1.2	CDER	2011-06-13	
Clinical study data definition	Define	XML	CDISC	2.0	N/A	CBER, CDER, CDRH	2013-08-07	
Clinical study data definition	Define	XML	CDISC	1.0	N/A	CBER, CDER, CDRH	Ongoing	

Table 1. Excerpt from FDA Data Standards Catalogs

VALIDATION RULES PROVIDE FALSE POSITIVES

Even with the best ESPECs there are checks that will continue to fire because they have not been written in a way that reflects the real world. For example, pH does not have a unit so the OpenCDISC check SD0026 will fire a warning that is a clear false positive. Unfortunately, without detailed knowledge that this exists, it can created misunderstanding around the quality of the data.

The PhUSE organization has documented 20 of these checks as part a working group to improve the OpenCDISC reports. http://www.phusewiki.org/wiki/index.php?title=Top_20_Validation_Rule_Failures_(CBER)

SD0026

Message: Missing value for --ORRESU, when --ORRES is provided

Description: Original Units (--ORRESU) should not be NULL, when Result or Finding in Original Units (--ORRES) is provided

Category: Consistency

Severity: Warning

Dear Project Manager,

The client's concern with the SDTM delivery is due to a misunderstanding and outstanding Data Clarification Forms that have not been processed. The errors from the validator output occurred because the client used OpenCDISC Validator v1.4.1 to check our STDM v3.1.2 amendment 1 domains using the SDTM v3.1.3 configuration file and the incorrect CDISC CT dated 2012-12-21. On review of the missing values for LBORRES, this exists on the data for the LBORRES where the value is pH. Unfortunately this is a known false positive with OpenCDISC Validator. The length reduction errors will be present until the database is frozen because our process is to apply the length reduction on variables nearer to the end of the study so that the conversion programming stays consistent between draft deliveries. We better have a meeting to discuss expectations and the process soon.

Thank you,

Terek

SAMPLE SAS CODE TO HELP

Now that the process has been ironed out between groups, presented here are a couple simple bits of SAS code to help with this process. First is a simple set of code to convert your datasets to SAS v5 transport files. FDA continues to only except this version of "datasets". They are looking into other ways to submit data, like dataset XML, but that is currently under discussion.

```
proc datasets lib=work memtype=data kill; run;

/** Macro to put out one xpt per dataset rather than all combined **/

%macro senddat (selfile =,inlib = sdtm);

   proc datasets library = &inlib
                 memtype = data;
      copy out = work;
      %if (&selfile ^= ) %then
         %do;
            select &selfile;
         %end;
   run;

   libname outdat xport "\\&Server\&Client\&Project\Data\xpt\&selfile..xpt" ;

   proc copy in      = work
             out     = outdat
             memtype = data;
      %if (&selfile ^= ) %then
         %do;
            select &selfile;
         %end;
   run;

   proc datasets library = work
                 memtype = data
                 kill;
   run;
%mend senddat;

libname sdtm "\\&Server\&Client\&Project\Data\SDTM";      ** Converted data **;

%senddat(selfile = ae,     inlib = sdtm);
%senddat(selfile = suppae, inlib = sdtm);
```

Second, simple PUT statements, like below, or PROC FREQ statements can indicate issues prior to ever creating the OpenCDISC validator output. If known false positives exist, catch those as you are working with the data.

```
data _null_;
     set lb;
     if upcase(lborres) = "PH" and lborresu = "" then
         put "SDTM NOTE SD0026 "usubjid= visitnum= lborres= lborresu=;
run;

SAS LOG:
SDTM NOTE SD0026 usubjid=111-111-111 visitnum=1 lborres=pH lborresu=
```

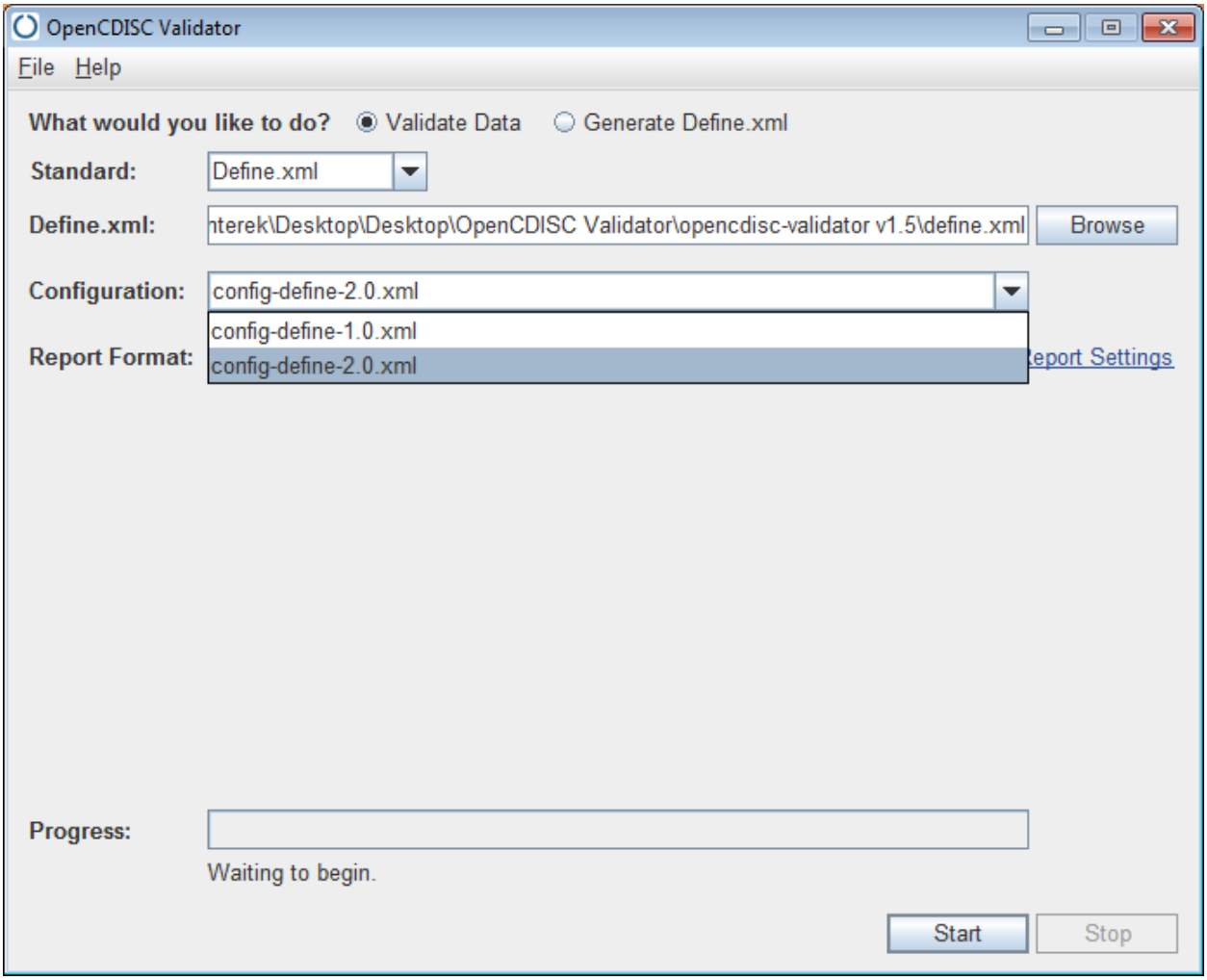
CONCLUSION

In summary, aligning and implementing up front ESPECs based around validator rules, coupled with the multi-level review approaches detailed above reduces the level of risk of significant numbers of resolvable data errors making their way through a clean transfer considerably. This level of risk needs to be offset against the additional effort

required to review multiple lines of a validator listing, particularly at points in the study where timeline pressures are typically increased.

It may seem like the authors of this paper are not advocates of the OpenCDISC Validator product, on the contrary they do see the purpose of the product and applaud the idea that a free tool is being made to industry to better meet FDA data requirements. The tool should be used as a CDISC integrity and compliance checker to confirm the conformance of a set of data that they meet the requirements set out by the FDA and CDISC organization. The point being made is where should data checks reside in the process of collecting, cleaning, and reporting of data. The process seems to have been ignored when some checks were created and the goal of reducing effort and cost to industry has done the opposite. This is related to two primary causes (1) the data cleaning checks in the OpenCDISC Validator tool and (2) the numerous versions of standards and versions of OpenCDISC validator that are not completely aligned with the supported standards in the FDA Data Standards Catalogs and other FDA guidances. The implementation within a CRO and sponsor partnership therefore can be a complex multiple stakeholder process.

Interestingly during the writing of the final version of this paper, OpenCDISC Validator v1.5 was released. The ability to check define.xml v2.0 files is now a configuration; see below Display 3 – OpenCDISC Validator Version 1.5 Define v2.0 Validator. Additionally added is the ability to check SDTM v3.2, however this is currently not listed as a supported FDA standard listed in FDA Data Standards Catalogs spreadsheet. SDTM v3.1.2 amendment 1 is still not a configuration in this newest version but is listed as a supported standard. For those of us that do not work with standards every day, the numerous versions and lack of alignment across FDA, OpenCDISC, and CDISC has created a good amount of misunderstand and miscommunications with stakeholders trying to get that breakthrough compound to market.



Display 3. OpenCDISC Validator Version 1.5 Define v2.0 Validator

REFERENCES

1. CDER Common Data Standards Issues Document (Version 1.1/December 2011) http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM254113.pdf
2. FDA's Study Data Standards: http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm
3. CDISC: http://www.cdisc.org
4. OpenCDISC: http://www.opencdisc.org

ACKNOWLEDGMENTS

The authors would like to thank PRA International for the time to make these issues known. We would also like to thank Karin LaPann, David Fielding, and Anja Koster for content, code, and material. These folks always have their ear to the ground and provide insight and solutions to the ever changing standard's landscape.

RECOMMENDED READING

- FDA Guidance Document - Guidance for Industry Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring: http://www.fda.gov/downloads/Drugs/.../Guidances/UCM269919.pdf
- EMEA Guidance Document: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2011/08/WC500110059.pdf
- Controlled Terminology, on NCI-EVS: http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/cdisc

CONTACT INFORMATION

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

Name:	Terek Peterson, MBA
Title:	Senior Director, Global Standards Strategies
Enterprise:	PRA International
Address:	630 Dresher Road
City, State ZIP:	Horsham, PA 19044
Work Phone:	+1 215.444.8613
E-mail:	PetersonTerek@PRAIntl.com
Web:	praintl.com

Name:	Gareth Adams
Title:	Senior Director, Data Services
Enterprise:	PRA International
Address:	Llys Tawe, Kings Road
City, State ZIP:	Swansea, UK SA1 8PG
Work Phone:	+44 1792 525611
E-mail:	AdamsGareth@PRAIntl.com
Web:	praintl.com