PharmaSUG 2015 - Paper SS06

The Most Common Issues in Submission Data

Sergiy Sirichenko, Pinnacle 21, Plymouth Meeting, Pennsylvania
Max Kanevsky, Pinnacle 21, Plymouth Meeting, Pennsylvania

ABSTRACT

On December 17th, the FDA made its long-awaited announcement that future submissions will be required in standardized format. FDA published the technical requirements for standardized submission data in a new binding guidance and supporting documents including the Data Standards Catalog and Study Data Technical Conformance Guide. They also encouraged sponsors to communicate with the review divisions on study specific data questions. Even though the guidance is new, most sponsors have already migrated to standardized submissions with the level of compliance rising rapidly. This has enabled FDA to improve the efficiency of the review process by developing automated review and analysis tools, which have been operationalized by the FDA JumpStart service. This presentation will share our experience of the most common data quality issues we observed during JumpStart across many regulatory submissions. We also provide recommendations on how to ensure high quality submission data by evaluating the risk or potential impact of each issue and how each can be corrected.

INTRODUCTION

It was a long time coming. But, finally, on December 17, 2014, the FDA announced that applications must be submitted electronically, and that submissions will be required to contain study data in conformance with CDISC standards. The industry has been given 24 months from the publication of the final guidance documents to comply, at which point the FDA may refuse to file (RTF) any submission that isn't received in electronic form and/or doesn't conform to the required FDA study data standards, formats, and terminologies.

It all started back in 2002, when FDA launched its 21st Century Review Initiative, which sought to establish a set of performance standards to be followed during drug review. One of FDA's goals was to make the submission process more efficient and standardized. The primary driver was time. Just a few years ago, when the FDA received a submission, reviewers spent up to 45 days just assessing the quality of that submission. They had to first churn through enormous files — some of which are gigabytes in size — to determine if the data quality and content are strong enough to support the review. Between the time spent by the FDA to conduct this "pre-review," and the time spent by the submitting organization to fix these issues and resubmit, valuable months are often wasted.

FDA also needed a way to leverage automated analytics and data-driven tools to assess data from clinical trials more efficiently. This meant that data had to be standardized to enable development of a new generation of review, analysis, and data visualization tools that could work with data from any submission. This would enable reviewers to spend less time trying to analyze data and more time ensuring that safe and effective drugs are approved quickly and seamlessly for public use.

High quality standardized data became the key ingredient to the FDA meeting its goals.

Figure 1. High quality data, the key ingredient to FDA's 21st century review initiative [2]

FDA'S DEFINITION OF HIGH QUALITY DATA

The FDA defines "High Data Quality" as that which is both *compliant* and *useful* [3].

- *Compliant* means the data confirms to applicable data standards
- *Useful* means the ability of data to support the intended use

The need for standardized, compliant data was the impetus for the establishment of CDISC standards. But the usefulness of data is an issue that still needs to be addressed. What good is a set of standards if the data being submitted doesn't support its intended use?

In the context of an FDA submission, useful could mean any of the following:

- Does the data support the use of FDA's standards-based review tools?
- Can reviewers reuse a common analysis, like Liver function or Hy's Law plot, across submissions?
- Is data well documented, so reviewer can quickly orient themselves to the contents of the submission?
- Are there data quality issues that could impact the review process and results?

An important take-away is that usefulness is depended on intended use, which could differ between medical and statistical reviewers, review divisions, or even the reviewer's preferred review and analysis tools. So understanding the intended use is very important in ensuring you data is useful.

FDA DATAFIT AND JUMPSTART PROJECTS

To ensure "High Quality Data", FDA launched the DataFit project [9]. The project's goal, quite simply, is to enable FDA reviewers to rapidly assess whether submitted standard data is suitable for analysis and for loading into analytical review tools. Through DataFit — specifically, through a combination of OpenCDISC Enterprise software and implementation — a detailed assessment of submitted data is performed very early in the review process, based on intended use requirements and identified review activities. This helps reviewers understand, immediately, if there are any data-quality issues that could prevent them from doing their job.

DataFit assessments are performed as part of the JumpStart service that provides FDA review teams with additional exploratory data analyses [10]. These give reviewers a better understanding of their data and provide important information for conducting an effective evaluation of the submission. The JumpStart service has transformed the way reviewers approach their review by helping them be prepared and proactive early in the process. The positive impact of JumpStart was recognized when it received the HHS Innovates Secretary's Choice Award [4].

THE MOST COMMON ISSUES IN SUBMISSION DATA

OpenCDISC is the industry gold standard for evaluating data compliance with CDISC standards and FDA requirements. The open source and freely available Community version is widely used as a personal desktop application by programmers involved in preparation of standardized data [7]. The Enterprise version extends the open source software with focus on data composition analysis and data fitness assessment [8]. OpenCDISC Enterprise forms the core of FDA's DataFit project.

As the creators of OpenCDISC, we at Pinnacle 21 had the opportunity to collaborate with many biopharmaceutical companies, clinical research organization, and FDA on ensuring "High Quality Data". In the following sections we are sharing our experience and observations of the most common issues in submission data with the hope of helping you better prepare for your next submission.

METADATA ISSUES

Metadata for submission data is described by three primary documents: Define.xml, Annotated Case Report Form (aCRF), and Reviewer's Guide. These are also the most commonly overlooked deliverables and are responsible for the majority of common issues.

DEFINE.XML

The data definition file (Define.xml) is the most important part of the electronic dataset submission for regulatory review [3]. It's also one that is most often noted by reviewers to be deficient. A sufficiently documented Define file offers significant benefits. It provides detailed specification for datasets, variables, codelists, data origins, and derivations, which allow reviewers to interpret submission data faster and move through the process more quickly. Define.xml is also critical to some FDA systems, such as the Clinical Trial Repository (JANUS), which requires a compliant Define file before study data can be loaded.

Yet many have struggled to create FDA compliant Define files due primarily to limitations of Define.xml v1.0 standard. It was released a decade ago and has become outdated for today's submission needs.

Issues caused by Define.xml v1.0 limitations

For example, Define.xml v1.0 cannot adequately handle Value Level metadata. CDISC models are normalized and store different test observations (height, weight, etc.) in same variables. Thus, Value Level metadata is needed to provide sufficient detail for each observation (like allowed units or controlled terminology) to support data review and analysis. This is especially important for analysis data, which is why FDA requests that sponsors utilize Define.xml v2.0 instead that corrects the Value Level limitations.

Another major limitation of Define.xml v1.0 that causes issues for reviewers is the lack of specific requirements for the capture of data origins. Was the data collected on CRF, derived, or received from laboratory? If collected on CRF, then on what pages? If derived, then with what method? Define.xml v1.0 only provided optional fields called "Origin" and "ComputationMethod", but no clear requirements or controlled terminology on how use them. This has resulted in the following common issues:

- Missing Origin
- Origin="CRF", but no reference to particular page(s)
- Inconsistency between origin and derivation (ex: Origin="CRF Page" and ComputationMethod populated)
- Origin="Derived" without detailed derivation algorithm

Define.xml v2.0 has been released in 2013, which has resolved most of the prior version's limitations. It's more robust and is better suited to support current reviewer's needs. However, the industry has been very slow to implement Define.xml v2.0, which is surprising considering that Define.xml v1.0 is almost as old as SDTM IG 3.1.1. Do you know many companies that are still using SDTM IG 3.1.1? We highly recommend that industry upgrade to Define.xml v2.0 to take advantage of new functionality that improves description and reviewability of submission data.

Regardless of what version of Define.xml sponsors use, there are a few other common deficiencies that we typically observe with Define content.

Incorrect or missing codelists

- *Missing codelists for study specific data elements* – sponsors populate codelists only for variables that have standard CDISC Control Terminology (AEACN), but do not create study specific codelists. For example, for Category (--CAT), Subcategory (--SCAT), or EPOCH variables.
- *Missing codelists for Value Level metadata* – SUPPQUAL domains are typically described using value level metadata, but sponsors often leave out codelists for supplemental qualifiers that have controlled terminology.
- *Codelists created for variables collected as a free text* – Codelists in define.xml should describe data collection process. We recommend creating codelists only for variables where data was collected, derived or assigned based on a list of pre-specified terms. For example, if CMDOSU is collected using values from a drop-down menu in EDC system, it should reference a codelist in Define.xml file. However, if CMDOSU was collected as free text, a codelist is not necessary as it will result in presence of several hundred unique terms. We believe that in most cases study data codelists with more than 30-40 terms are impractical and are never used.
- *Collapsed codelists for multiple variables across domains* – for example, a single UNIT codelists for all --ORRESU, --STRESU and --DOSU variables within a study. In some studies, such collapsed UNIT codelists can result in >500 terms assigned to EXDOSU variable, while in reality EXDOSU variable only used one term "mg". We strongly recommend creating a separate codelist for each variable.

Missing, unclear or invalid Computational Algorithms

All "Derived" variables must have clear and detailed description of computational algorithms so reviewers can understand how values were derived and can independently reproduce them if needed. However, majority of submissions still have missing or poorly documented computational algorithms. Quite often sponsors provide "generic" algorithms for Study Day and Baseline Flag variables, but do not provide any information for important study specific derivations like EPOCH, SESTDTC, etc.

Sometimes in computational algorithms sponsors refer to non-available information like raw data from EDC system or external look-up conversion tables, additional documentation which is not included in submission data package. Please ensure that all Derived variables and Value Level have clear, correct and detailed computational algorithms, which only use data elements and information included in the data package.

Missing descriptions for study and sponsor specific variables

One of the most severe issues in Define content are the missing descriptions for study and sponsor specific variables, like --SPID (Sponsor ID), --GRPID (Group ID), etc. Very often these sponsor-specific variables are one of the Key Variables in datasets, responsible for "duplicate" records and play other important roles. However, if sponsor did not fully describe these variables (e.g., meaning, source, computational algorithms, etc.), then there is no way to understand the submitted data. The biggest value of Define file is to provide descriptions for study specific data elements. But unfortunately many sponsor just copy CDISC notes from SDTM IG in place of providing the important study specific metadata.

ANNOTATED CRF

An Annotated Case Report Form (aCRF) documents how the data was collected and how it was mapped to SDTM datasets. When Origin attributes in Define.xml are properly populated, a reviewer can simply click on a hyperlink and be taken directly to the CRF page where the dataset variable was original collected. This greatly improves reviewer's ability to understand the source of data. It also shows traceability to ensure that all collected data has been submitted (except the data that is clearly marked as NOT SUBMITTED).

We have observed the following common issues with annotated CRFs:

- Missing or incorrect annotations
- Annotations that were not created with PDF Annotation feature, but instead are represented by highlighted text or PDF form fields
- Annotations reference EDC database fields instead of variables in SDTM

REVIEWER'S GUIDE

Study Data Reviewer's Guide (SDRG) was introduced in 2013 to provide FDA reviewers a high-level summary and additional context for the submission data package. It purposefully duplicates information found in other submission documentation (protocol, clinical study report, annotated CRFs, define.xml, etc.) in order to provide FDA reviewers with a single point of orientation to the submission data [6]. Reviewer's Guide communicates additional information about mapping decisions, sponsor-defined domains, and sponsor extensions to CDISC controlled terminology. It also captures sponsor's explanations of data validation issues, specifically the reason why those issues were not addressed during study conduct, mapping, and submission preparation.

We have observed a rapid adoption of Reviewer's Guide by the industry, primarily due to its popularity with FDA reviewers but also for its usability. On average, a Reviewer's Guide has only about 20 pages, which is a lot less than hundreds of pages across protocol, define.xml, and other documents.

Overall, they quality of Reviewer's Guides has been improving, however a number of common issues are still observed.

Not following the recommended structure

The structure of the Reviewer's Guide must follow the recommended template provided by PhUSE (http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide). However, some sponsors only fill out parts of the template, which significantly reduces the document's value for FDA reviewers.

Missing or meaningless explanations for data conformance issues

The Data Conformance Summary section of the Reviewer's Guide provides an opportunity for sponsors to identify and explain in detail why some of the data issues were not fixed. This helps reviewers navigate around the data issues during analysis and preempts the need for additional question and clarifications.

Many sponsors still use outdated versions of OpenCDISC, which results in some issues not being identified or explained. Since FDA always uses the latest version of OpenCDISC (a.k.a. DataFit) and validation rules, this leads to missing explanations in Reviewer's Guide.

Even if sponsors use current or recent versions of OpenCDISC, in many cases the explanations provided are not sufficient or seems to be more of an excuse than an explanation. Here is a list of our favorite explanations that you should NEVER use in a Reviewer's Guide:

- *"Expected result"*
- *"This is a common practice"*
- *"As received from our vendor"*

- *“Sponsor decided not to fix”*
- *“We did not collect nor derive this data element”*
- *“We do it differently than the standard”*

Issues explanations that show incorrect interpretation of CDISC standards and FDA requirements

Very often data conformance issue explanation show that a sponsor does not understand or has an incorrect interpretation of CDISC standards and FDA requirements. For example

- Issue: *“Date is after RFPENDTC”* (in most domains, 10 – 60% of records)
- Sponsor explanation: “*RFPENDTC is the last date of participation for a subject for data included in a submission. RFPENDTC is set to the latest DSSTDTC in DS domain where DSCAT=’DISPOSITION EVENT’*”

As you can see, a Sponsor derivation algorithm does deviates from a CDISC definition of RFPENDTC variable. Also in this particular study Sponsor used DSDTC variable instead of DSSTDTC for most Disposition Events including “Study Completion” records.

We recommend paying special attention to Reviewer’s Guide. Consider this as an additional opportunity to communicate with your FDA reviewers. Remember that Reviewer’s Guide is a product of collaborative work and input is expected from your entire team, including statisticians, programmers, data managers, clinical team and data vendors.

NONCOMPLIANCE WITH FDA BUSINESS RULES

FDA started publishing business rules for submission data in May 2011 with the introduction of CDER Common Data Standards Issue Document. The document was designed as a supplement for CDISC implementation guides providing additional requirements for review specific data elements, the use of controlled terminology, and other general considerations like file sizes. The goal was to clarify FDA expectations and reduce variability in submission data. However, despite the long availability of these requirements, the industry has been very slow to comply.

In November 2014, FDA re-published their business rules as two validation rule specifications, one for SDTM and one for SEND data [5]. These validation rules specifications were formalized just a month later in the Study Data Technical Conformance Guide, a part of FDA’s final guidance to industry for “Providing Regulatory Submissions In Electronic Format — Standardized Study Data” [1]. OpenCDISC has followed suit and published an executable version of FDA validation rules in OpenCDISC Community v2.0, release in December 2014.

Now that FDA business rules have been raised to guidance level, let’s review the most common issues that could impact the reviewability of submission data.

Missing EPOCH values

FDA asks sponsors to populate EPOCH variable for clinical subject-level observations (adverse events, laboratory, exposure, vital signs, etc.). This greatly helps in analysis by allowing reviewers to easily select records related to a particular phase of the trial.

Missing AE Seriousness Criteria

ICH guidance documents (E2A, E6) ask that sponsors use a special set of AE Seriousness Criteria to ensure correct classification of Seriousness Adverse Events to avoid confusion between “serious” and “severe”. FDA validation rules includes a special rule, FDAC206, which reminds sponsors to collect and submit AE Seriousness Criteria data. Some sponsors do not include this info in submission data package and make it limited to a separate Pharmacovigilance reporting only. There are many potential problems with such non-compliant approach. There is no evidence that a classification of AE Seriousness was done correctly, reducing trust in sponsor data. There is also an increased risk of data management errors. For example, non-serious AEs with AEOUT=”FATAL”, AEOUTOTH=”Hospitalization” or AETERM=”Myocardial Infarction”.

Death reporting inconsistencies

Death of a study subject always gets very close scrutiny by FDA reviewers. New versions of SDTM IG have special variables called DTHFL (Death Flag) and DTHDTC (Death Date) in Demographics (DM) domain to simplify the reporting of death data. FDA also asks that sponsors create a special DEATH record in Disposition (DS) domain for all subjects who died and ensure that it’s the last record for the subject. The compliance with these FDA business rules is still very low, making reconciliation of subject death information across submission data difficult and slows down the review process. Here are a few common death reporting inconsistencies:

- Inconsistency between DM and DS death information
- Subject death information is listed in DM, but not in DS
- Missing death dates in DM or DS domains. Ensure date are collected and captured in DTHDTC and DSSTDTC variables
- Invalid coding of DS terms. For example, DSTERM=”Death” is coded to DSDECOD=”ADVERSE EVENT” or “OTHER” instead of “DEATH”
- Death information is in domains other than DM and DS. For example, subject death is not listed in DM and DS, but
  - Subject has a FATAL Adverse Event
  - Subject has a Comment record like “After subject death …”
  - Subject has a Protocol Derivations record like “Due to subject death …”
  - Subject has a Date of Autopsy record in SUPPQUAL domain

To improve reviewability of data, please ensure that data quality of subject death data is compromised.

Missing Disposition dates

Analysis of Disposition data is very important in regulatory review. It can provide indirect estimation of drug safety and efficacy. Unfortunately current standards and regulatory documents do not explicitly emphasize the need to collect timing info in DS domain. Missing dates in Disposition is still one of the most common issues. Data managers should ensure that subject study completion and follow-up contact dates are included in CRF design and collected data is complete and cleaned. This info should definitely be included in a checklist of Risk Based Data Verification plan.

Duplicate records

Duplicate records can complicate analysis by causing issues with FDA review tools. These can occur for many different reasons, here are the most common examples:

- Records with all variables having the exact same value (except --SEQ)
- Different results for the same laboratory test and collection time-point (e.g. one Normal and one Abnormal)
- Records with same Original Result, but different Original Units
- One record with actual result and another record on the same time-point as “NOT DONE”
- Records with same Results and Collection Dates/Times, but different Visits
- Records that are only different in LBSPID variable, which is not described in define.xml or SDRG

Data management teams should investigate and eliminate duplicate records before data lock, preferably during data collection and while cleaning data from external data transfers. Special effort should be dedicated to safety and efficacy data.

Incorrect collection and mapping of Race

Another issue that is often observed in submission data is the incorrect collection and mapping of subject Race. Common practice is to use terms MULTIPLE, OTHER and UNKNOWN in addition to FDA suggested minimal set of standard terms AMERICAN INDIAN OR ALASKA NATIVE, ASIAN, BLACK OR AFRICAN AMERICAN, NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER, and WHITE. When subject RACE=”OTHER”, details are expected to be provided in SUPPDM domain. Unfortunately, quite often such information was collected using free text for “Other, specify” field on CRF. If data was not cleaned or data mapping was done on variable rather than value level, it resulted in invalid presentation of subject Race information. For example, subject with RACE=”OTHER” may have “Race Other” in SUPPDM as

- “Caucasian” (should be mapped to “WHITE”)
- “Hispanic” (it’s Ethnicity, not Race)
- “United Kingdom” (it’s Nationality, not Race)

- “Not Reported” (use “UNKNOWN” term for Race in DM domain instead)

Another common example of invalid mapping is RACE=”MULTIPLE” and SUPPDM term as “White and Hispanic”. Correct mapping would be to split collected term into RACE=”WHITE” and ETHNIC=” HISPANIC OR LATINO”.

Analysis by Race subgroups is a standard procedure during regulatory review, so please ensure that this information is accurately collected, cleaned, and correctly mapped.

Missing values for SDTM required variables

Structural data consistency issues may result in failure of FDA review and analysis tools and cause significant data management challenges. Missing values for “SDTM Required” variables still exist in submission data. Typical examples are PRTRT, HOTERM, EXTRT variables. There is a record that a study treatment was administered, but no information to which one. Since most studies are now run using EDC systems, data management should ensure that collection of all “Required” information is enforced by edit checks. If the issue is present in a locked database, then a recommended approach is to at least use some term like “UNKNOWN” rather than keeping “Required” variable value as missing.

PROGRAMMING AND MAPPING ERRORS

In addition to data collection issues, there are many common issues due to programming errors:

- Inconsistency between Trial Visits (TV) and Subject Visits (SV) domains versus other domains is still quite common issue. It’s usually due to different spelling of VISIT values, inconsistent usage of VISITNUM values across domains, or missing Visit records in TV or SV domains.
- RELREC or SUPPQUAL domains have references to non-existing records. It’s a severe violation of structural data integrity, which may prevent execution of review and analysis tools.
- Submission data should be cleaned to remove leading spaces and special characters like line breaks and carriage returns. This is especially important for variable labels.
- A concept of Standard Unit means that its values should be the same for a particular test defined by --TESTCD, --SPEC, --METHOD and other Record Qualifier variables. A violation of this rule leads to unpredictable results in data analysis.
- SDTM Model has special paired variables with expected one-to-one consistency in their values. Examples include --TESTCD/--TEST, --PARMCD/--PARM, ARMCD/ARM, VISITNUM/VISIT, --TPTNUM/--TPT, etc. These variable pairs can be used interchangeably in analysis, thus submission data should have consistency between their values.
- In addition to consistency in values, a programmer should ensure consistency in usage of Control Terminology. For example, terms in Lab Test Code (LBTESTCD) and Lab Test Name (LBTEST) codelists are linked by the same NCI Code value, while NCI Codelist Code values are different. If a standard term is used for one variable, a paired variable value should also be populated by standard term as defined in Control Terminology. For example, if EGTESTCD=”QTC” and EGTEST=”QT Uncorrected”, one of these two values is definitely incorrect and we do not know which one? Some programmers use CDISC Synonym(s) column to populate values in “paired” variable. It’s not a valid approach. Remember to utilize the first “Code” column to find corresponding term.
- Some programmers still use a --TPT variable for Actual time instead of Planned time.
- --STRF and --ENRF variables are utilized for Screen Failures and Not Treated subjects, which per study data definition do not have RFSTDTC, RFENDTC values. Remember, that --STRF variable provides reference to RFSTDTC value. If for some reason a subject does not have RFSTDTC info, then a usage of --STRF variable is not applicable. In such cases a programmer should utilize another set of variables --STRTPT and --STTPT.
- SDTM standard has a special purpose domain Comments (CO). However some sponsors still prefer to use SUPPQUAL domains to keep comments. While sometimes such approach may seem more logical for a programmer, it could be confusing for a reviewer because information was put into an unexpected location.
- SDTM data should only have collected or in some cases derived data. No imputation is allowed. However there are still cases when Study Days are imputed for partially missing dates.

CONTROL TERMINOLOGY ISSUES

Standard data has standard structure and standardized content. A usage of standard Control Terminology is required for regulatory submissions, because automated tools rely on it. For example, FDA DataFit has many validation checks where algorithms include filters for screen failure or non-treated subjects. If a sponsor utilized non-standard terms for ARMCD, ACTARMCD variables, it results in high rate of false-positive messages in data validation reports. Liver Function Analysis tools look for standard LBTESTCD names like ALP, ALT, AST, BILI and will not produce reports if study data is not compliant with standard Terminology.

Here are examples of common Terminology issues in submission data

- Ignoring existing terms in extensible codelists. New terms can only be added if they are not already represented in standard codelist.
- Modification of standard terms by conversion into Upper Case or misspelling.
- Not following new CDISC Controlled Terminology (CT) codelists. CDISC SDTM and Terminology are separate standards and are published separately, however terminology is assigned to variables in SDTM Implementation Guides (IG). So when CDISC introduces new terminology codelists there are unfortunately no assignment updates in already published versions of SDTM IG for new control terminology. Sponsors should monitor new releases of CDISC CT for new codelists, which may be applicable for their study data.
- When data collection as free text it leads to huge problems with implementation of standard terminology. A mapping of original values into standard term is resource consuming. Also, entering data as a free text does not provide any control to ensure correctness of information. For example, very often collected Dose Units for Concomitant Medication include invalid data like “000”, “1 Patch every four days”, “Tablet”, etc.
- Invalid data collection design. The most notorious example is a common misunderstanding the meaning of AEACN variable. While it’s “Action Taken with Study Drug”, some sponsor see it as any action taken for Adverse Event including Hospitalization, additional medication, etc. Also, some sponsor still follow outdated practice to collect AEACN as “Drug dose was modify” without specifying if the dose was increased or decreased? Such data collection design is not compliant with CDISC CT or FDA requirements.

CONCLUSION

Since 2004, when FDA first requested sponsors submit data in SDTM format, there has been a slow progress toward standards implementation by the industry. FDA began developing standards-based review and analysis tools, such as the Clinical Trial Repository (JANUS), to take advantage of standardized data to improve the efficiency and effectiveness of the review process. However, it quickly became evident that standards compliance was not enough and more focus had to be placed on usefulness of data for the intended use. So beginning in May 2011, with the introduction of CDER Common Data Standards Issue Document, the FDA started providing additional requirements to help sponsors achieve higher quality submission data. This was later followed up with Study Data Technical Conformance Guide, the Data Standards Catalog, and validation rule specifications for SDTM and SEND data. And finally, on December 17, 2014, FDA requirements were officially formalized as part of the final guidance to industry for “Providing Regulatory Submissions In Electronic Format — Standardized Study Data” [1].

The industry’s pace of standards adoption has greatly accelerated in the last few years due largely to FDA intent and commitment to standards becoming clearer. Over the last year we have observed an obvious improvement in standards compliance as sponsors begin observing the business rules described by various FDA documents. However, we still see many data fitness issues with submission data, those issues that impact FDA’s ability to utilize standards-based automated review and analysis tools. This paper shared our experience with the most common data fitness issues we have observed across many sponsors and regulatory submissions. We also provided recommendations on how to ensure high quality submission data by evaluating the risk or potential impact of each issue and how each can be corrected. We encourage sponsors to communicate with the review divisions to better understand the intended use and create submission data that is compliant with latest FDA requirements. As a parting note, keep in mind that as more new tools are developed at FDA more requirements for higher quality data will emerge. Creating submissions data that is fit for use requires a continuous improvement process.

REFERENCES

[1] "Providing Regulatory Submissions In Electronic Format — Standardized Study Data". *CDER*. December 2014. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf

[2] Rosario, Lilliam, Ph.D.,"Office of Computational Science Symposium: Applying he Right Tools, at the Right Time to Regulatory Review." *PhUSE Computational Science Symposium.* March 2015. Available at http://www.phusewiki.org/docs/CSS2015Presentations/001%20PhUSE%20CSS%202015%20Rosario.pdf

[3] "Study Data Technical Conformance Guide". *CDER*. December 2014. Available at http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

[4] "HHS Innovates Celebrates 7th Round of Innovations!". *HHS*. July 2014. Available at http://www.hhs.gov/idealab/2014/07/21/hhs-innovates-celebrates-7th-round-of-innovations/

[5] "Study Data Standards Resources". *CDER*.
Available at http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm.

[6] "Study Data Reviewer's Guide Completion Guidelines v1.2". *PhUSE*. January 2015.
Available at http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide.

[7] OpenCDISC Community. Available at www.opencdisc.org

[8] OpenCDISC Enterprise. Available at www.pinnacle21.net

[9] "FDA DataFit: an introduction". *Pinnacle 21.* August 2014.
Available at http://www.pinnacle21.net/blog/fda-datafit-an-introduction

[10] "FDA JumpStart: an introduction". *Pinnacle 21.* July 2014.
Available at http://www.pinnacle21.net/blog/fda-jumpstart-an-introduction

CONTACT INFORMATION

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

Name: Sergiy Sirichenko
Company: Pinnacle 21 LLC
Work Phone: 908-781-2342
E-mail: ssirichenko@pinnacle21.net

Name: Max Kanevsky
Company: Pinnacle 21 LLC
Work Phone: 267-331-4431
E-mail: mkanevsky@pinnacle21.net