PharmaSUG 2023 - Paper SS-266

Industry Metrics for Standards Utilization and Validation Rules

Sergiy Sirichenko, Pinnacle 21 LLC

ABSTRACT

In recent years the industry has completed the adoption of standards for study data. It opens new opportunities. One of them is the availability of industry-wide analytics which allows us to understand common trends and reveal potential issues.

At this presentation, we will share metrics for standards utilization and compliance with regulatory requirements. We will discuss the potential use of industry-wide analytics for the enhancement of existing standards, improvement of validation rules, and refinement of current processes.

INTRODUCTION

Metrics are an important tool. It provides an objective evaluation of the implementation strategy, measures progress towards the goals set, and allows making correct adjustments if needed. Also, metrics can be utilized as a data source for business research to better understand your users' behavior and needs.

Pinnacle 21 Enterprise has collected metrics since the early versions. Pinnacle 21 Community introduced this functionality in 2019. The original scope of metrics collection was the enhancement of existing validation rules. Soon, clients realized the value of operational metrics and started using them for analysis, enhancements, and monitoring of their existing business processes in data standardization and preparation for regulatory submissions [1]. Now Analytics is an out-of-box module in Pinnacle 21 Enterprise. While users already have real-time access to Pinnacle 21 (P21) metrics within their own company, they also want to compare themselves to the industry and to use the industry-wide metrics as a benchmark for their own performance.

However, utilization of any metrics should be done with a clear and good understanding of targeted goals as well as content, applicability, and limitations of collected data. Like clinical trials, the collection of any metrics should be well-designed and cleaned. Like statistical analysis in clinical trials, analysis of metrics should be done carefully to avoid misinterpretations and misuse.

As a first step, we would like to provide some examples of what metrics are already available and their potential usage, show some issues in collected metrics, and challenges in metrics interpretation. We hope that these can start a discussion about the industry's interest in global metrics.

METHODOLOGY

Metrics from both Pinnacle 21 Enterprise and Community were utilized. They are similar to each other, but not the same. Enterprise metrics are in general more robust and include additional attributes which are simply not applicable in Community. For example, additional validation engines and therapeutic areas info. Most metrics show quite similar results in both Enterprise and Community.

We performed minor data cleaning of collected metrics to filter out obviously invalid data. For example, partially failed validations and records which required fixing due to special characters. This procedure should not have a significant impact on results due to the huge volume of input data, but it simplified the analysis process.

While metrics are available in Enterprise since 2011 and in Community since 2019, we focused on the 2022 as the most recent and relevant period to us today. In some cases, we will review historical changes in recent years.

We used 2 main evaluation metrics for validation: the number of studies and the number of validations. We will use both if needed or just select one according to our "best fit" interpretation.

A unique study was defined as a combination of Company and Study ID. There is an obvious overlap between some studies which were done by CRO and additionally validated by sponsor. However, we kept them as separate entities in our analysis to not over-complicate this research. We believe that those can be considered as two different processes: data preparation at CRO and work quality confirmation at the sponsor side. Therefore, they represent two different use cases.

We have already done research on the industry metrics utilization and presented them at PharmaSUG and PhUSE conferences [2, 3, 4]. One important finding was that there are many challenges in interpretation of metrics. In this exercise, we also would like to explore the practical applicability of metrics and their limitations.

VALIDATION METRICS

P21 validations were executed in 40 different countries with an additional 1.5% in unknown locations. USA represents most data validation activity (56%) in the global market, with China (9.5%) and India (8.5%) as runners-up. Combined EU countries and UK represent 14.1% of global validations.

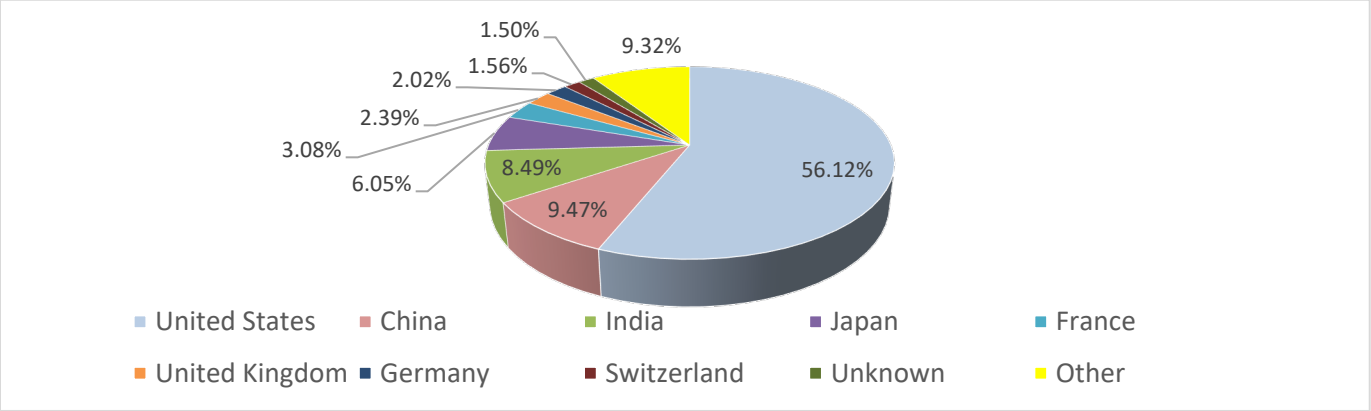

Figure 1. Validations performed by location

About 20% of studies are validated by several different companies. This usually happens when the same study is validated by both a sponsor and one or more CROs.

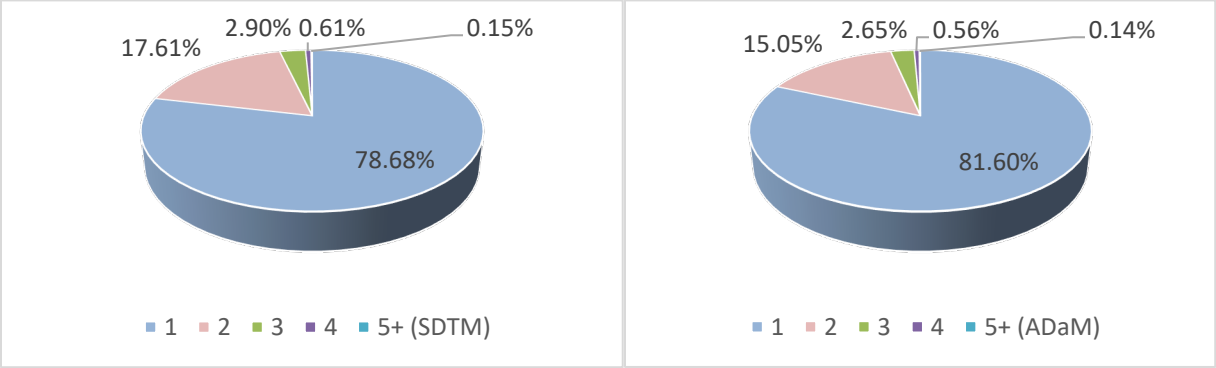

Figure 2. Number of different companies validating the same study

Most data validations are performed for SDTM standard. A total number of studies with ADaM validations is only 62% of studies with performed SDTM validation. It's not clear if this is due to the differences between SDTM and ADaM processes (e.g., SDTM is created during study conduct for all studies, while ADaM is created only for studies going to regulatory submissions) or some issue in collected metrics or misinterpretation. Another explanation may be due to sponsors common practice to outsource SDTM

creation to CROs but handle ADaM internally. In this way some studies will have ADaM validation done only by sponsors. The number of validations for SDTM data within the same study is usually more than for ADaM data.

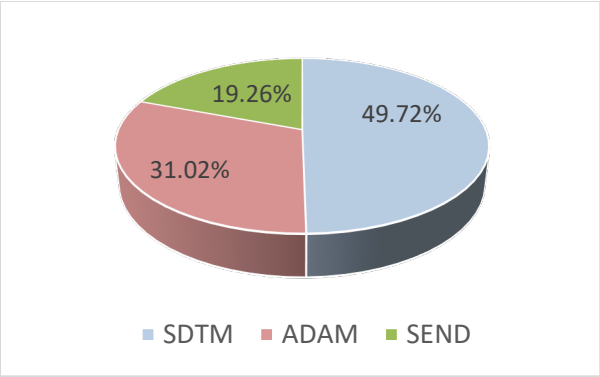

Figure 3. Number of studies by standard

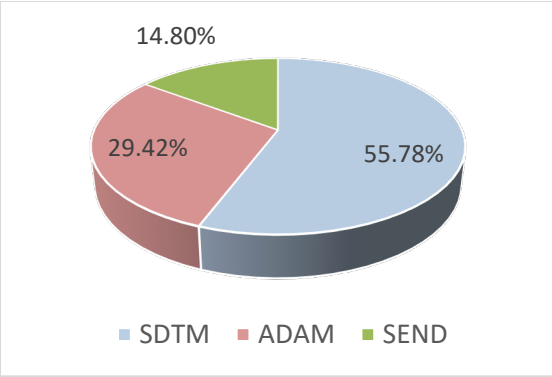

Figure 4. Number of validations by standard

In 2022, most studies still used SDTM-IG 3.2. Adaption of new SDTM-IG versions is low. Enterprise users look more conservative and keep older versions of SDTM-IG for longer time despite early availability of support for new CDISC standards in Enterprise. Another interesting finding is that in previous years Community users selected outdated SDTM-IG 3.1.2, ignoring SDTM-IG 3.1.3, and then jumped directly to SDTM-IG 3.2.

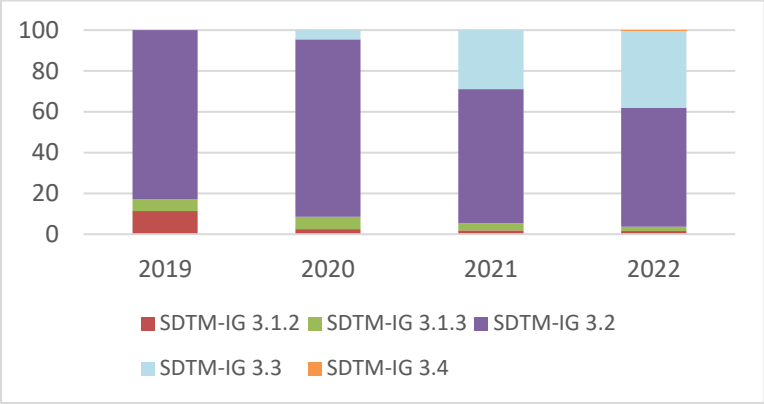

Figure 5. Validations by SDTM version in Community

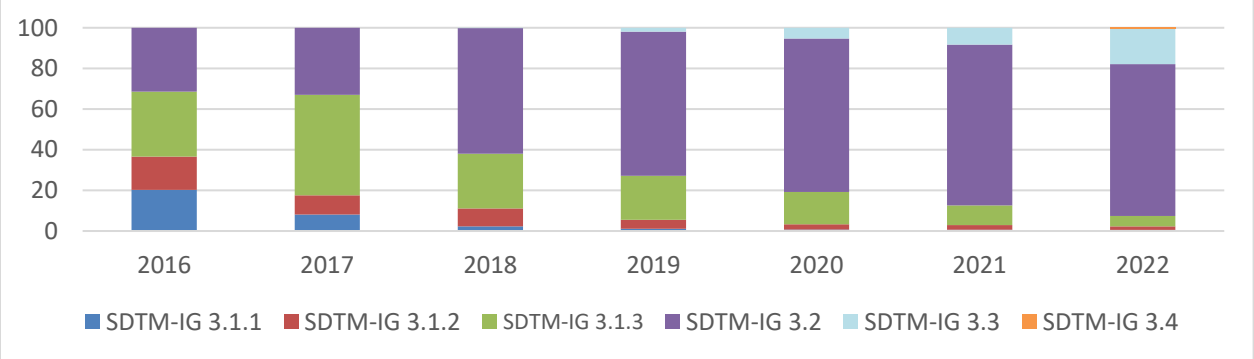

Figure 6. Validations by SDTM version in Enterprise

A similar pattern is in the adaption of new ADaM versions. Despite their early availability in Enterprise, its users prefer to stay longer on older versions of ADaM standard. However, an alternative explanation may be due to submissions to PMDA agency. Many studies are submitted to PMDA which accepts slightly outdated versions of the standards. For example, SDTM-IG 3.3 can be used only in 2023. ADaM-IG 1.2 is still not an option for PMDA. It means that when preparing PMDA submissions, some sponsors may need to use older standards or just formal validation for older standards to comply with PMDA requirements. And those studies have already been submitted to FDA using later versions of ADaM standard.

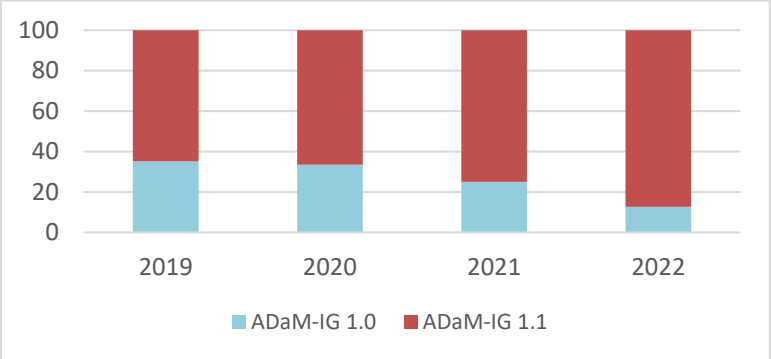

Figure 7. Validations by ADaM version in Community

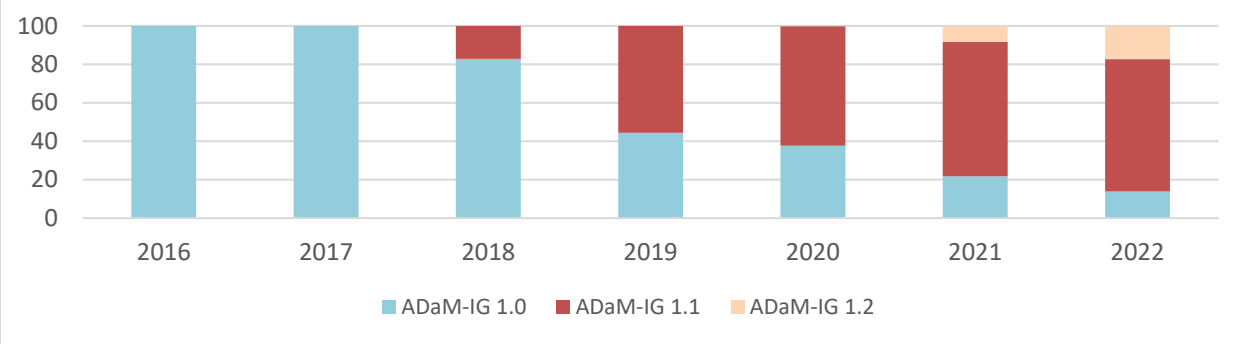

Figure 8. Validations by ADaM version in Enterprise

Looking at validation engine utilization, 86% of validations were performed with the FDA engine. There are 3 potential explanations:

- Most studies are used only for FDA submission
- Most data preparation and cleaning are initially done for submission to FDA and later the study data is only adjusted for PMDA submissions requiring less validations
- FDA engine includes the most up-to-date validation rules and is thus preferred by users for ongoing studies

NMPA engine accounted for 4% of validations in 2022 and is growing. While validations with CDISC CORE engine only accounted for less than 0.1%.

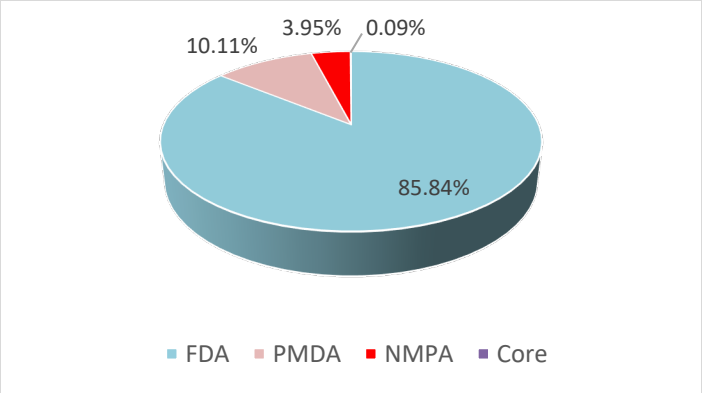

Figure 9. Validations by agency-specific engine

The next 2 graphs show monthly utilization of validation engines for PMDA and FDA agencies. The rapid switch to 2010.2 engine compared to previous transition to 1810.3 is due to added PMDA support for ADaM-IG 1.1.

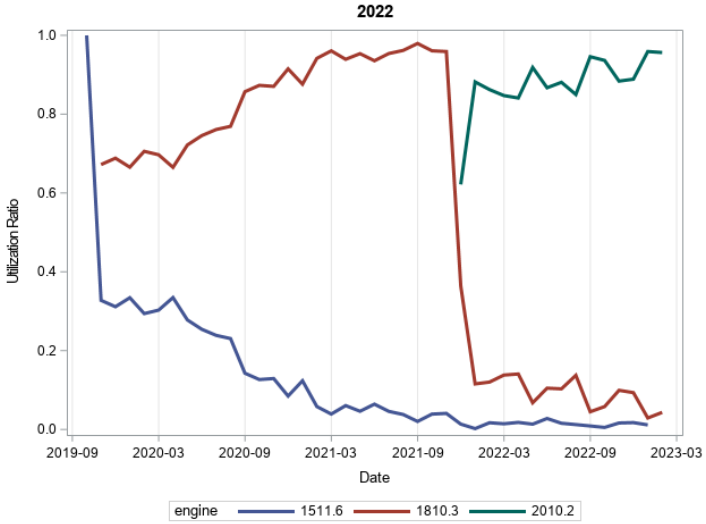

Figure 10. PMDA engine adaption

Adaptions of new FDA engines are in general quicker and smoother compared to PMDA due to different requirements and recommendations from these agencies. While for PMDA submissions sponsors must use pre-specified versions of validation rules based on the date of submission, for FDA submissions the latest and greatest of validation rules are recommended.

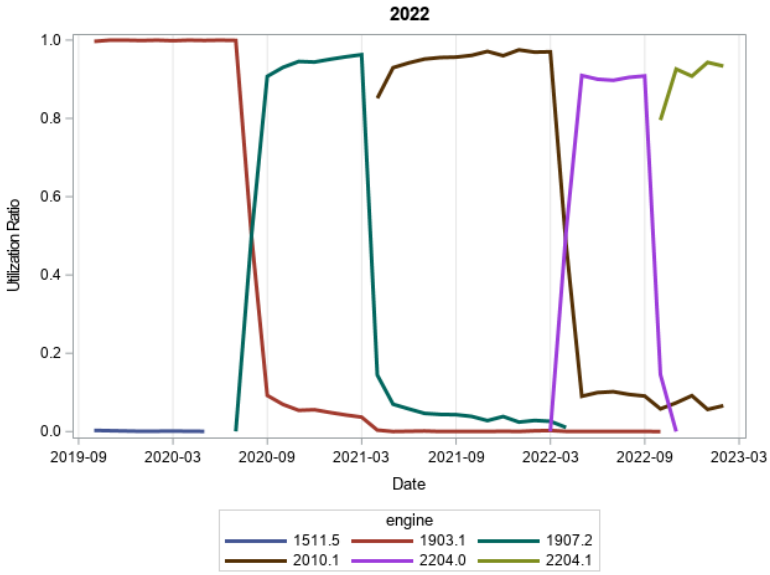

Figure 11. FDA engines adaption

In 2022, the three the most common MedDRA versions utilized in SDTM study data were 24.1 (22.1%), 24.0 (20.3%) and 25.0 (19.0%). 6% of studies still used the old MedDRA 8.0–19.1 versions.

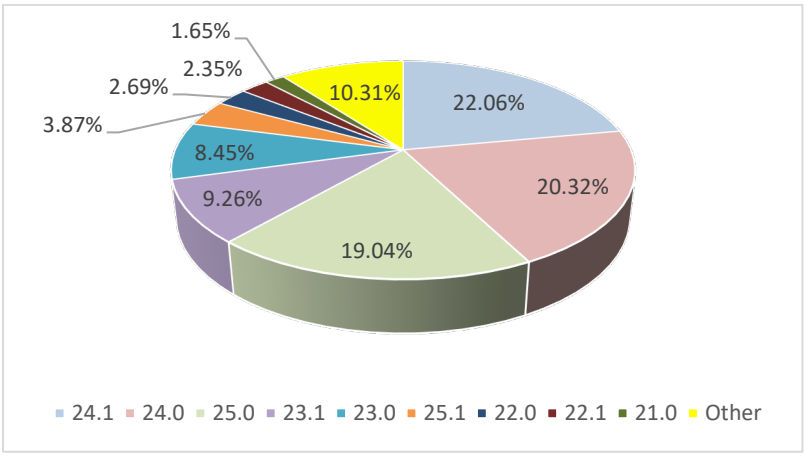

Figure 12. MedDRA versions usage for SDTM

The next Table 1 shows operational metrics for validation. Keep in mind that there are always outliers in terms of processes like testing and user training. For example, the study with maximum subjects in ADaM includes exactly 1,000,000 subjects. It does not look like data validation of a real study. Therefore, Min/Max statistics may be misleading, while 5th and 95th percentile are better options for interpretation.

Metrics All / SDTM / ADaM	Mean	Median	95th Pctl	Max
Subjects	319 / 362 / 406	57 / 60 / 70	834 / 872 / 1.2K	2.3M / 2.3M / 1M
Records	234K / 309K / 227K	29K / 43K / 26K	920K / 1.3M / 827K	119M / 94M / 119M
Records per subject	/ 1.6K / 818	/ 732 / 394	/ 4.8K / 2.4K	/ 146K / 309K
Datasets	25.4 / 34.4 / 12.6	24 / 35 / 11	53 / 57 / 26	133 / 129 / 133
Custom datasets *	- / 1.3 / 9.3	- / 1 / 8	- / 5 / 22	- / 57 / 128
Suppqual datasets	- / 10.8 / -	- / 10 / -	- / 22 / -	- / 64 / -
Issues	53 / 88 / 19	26 / 71 / 9	186 / 219 / 61	17K / 17K / 16K

* "Custom datasets" for ADaM in P21 metrics are any ones not recognized as standard datasets. For example, some OCCDS datasets in ADaM-IG 1.0.

Table 1. Validation metrics for all data

Note that distributions of validation statistics are usually quite skewed. Therefore, depending on the targeted application either Mean or Median can be a better option.

There are expected differences between small and large studies. Table 2 shows SDTM validation metrics for studies with different numbers of participating subjects.

The "No subjects" group is responsible for dedicated validations of either Define.xml file, Trial Summary domains, or empty datasets. The number of records per subject is going down slightly with increase of study size and drops significantly when study includes more than 10,000 subjects. Complexity of study data represented by number of datasets and custom domains is increasing with study size until study enrolment hits about 10K. The average number of issues is increasing because in larger studies there are always more unique types of data collection errors. 100K+ studies look like either greatly simplified or they are just test data.

Metrics Mean (Median)	No subjects	1-10	11-100	101-1000	1001-10,000	10,000-100,000	100,000-1,000,000
Records per subject	0	1.8K(628)	1.6K(746)	1.6K(757)	1.4K(560)	225(154)	13.5(3.3)
Datasets/Custom/SUPP	15.8/1.7/5.4	32/1.2/9.1	32/1.1/10	38/1.7/13	38/2.2/13	31/1.5/10	4.5/0.8/0.2
Issues	55(16)	82(67)	71(54)	114(101)	143(133)	155(145)	57(45)

Table 2. Validation metrics for SDTM data

Statistics for the number of validations within the same study for all SDTM data has Mean = 10.5 and Median = 3. Of course, there are always some extreme cases like studies with 1,000+ validations. 29% of studies have only a single validation for SDTM data.

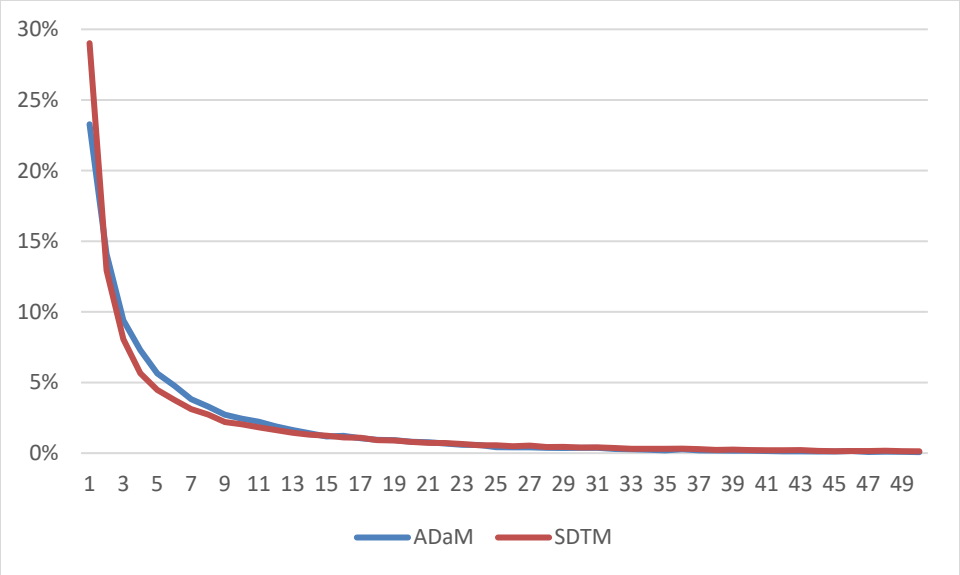

Figure 13. Distribution (%) of the number of validations performed per study

To explore changes in study data over a development period, we will exclude studies with a single validation and summarize the differences between study's first and last validation.

For studies with more than 1 validation, the average time between the first and the last validation has Mean = 81 days and Median = 34 days. 5% of studies were under SDTM development for more than 328 days. There are some expected increases in the number of records (mean = 47K, median = 220) and few datasets added (mean = 2.4, median = 0). On average the number of fixed issues is not significant (mean = 6.7, median = 2). However, 5% of studies had more than 83 fixed issues since their first validation.

Statistics for ADaM are consistent with SDTM, but all numbers are smaller. The number of validations within the same study for all ADaM data have Mean = 8.3 and Median = 4. 23% of studies have only a single validation for ADaM data.

Metrics **SDTM / ADaM**	**Mean**	**Median**	**95th Pctl**	**Max**
Number of validations	14.4 / 10.5	7 / 6	51 / 34	1053 / 369
Validation period (Days)	81 / 67	34 / 23	328 / 284	1147 / 1010
Records Difference	47K / 32K	220 / 6	267K / 148K	65M / 36M
Datasets Difference	2.4 / 1.1	0 / 0	30 / 12	107 / 75
Issues Difference	-6.7 / -3.5	-2 / -1	-83* / -31*	-1.4K* / -873*

* For Issue Difference 5th Pctl is used instead of 95th Pctl and Min instead of Max

Table 3. Changes between the first and the last validation in a study

RULES AND ISSUES METRICS

Analysis of metrics for validation issues helps identify bugs and rules which may need enhancements. However, based on our previous experience, these metrics are the most complicated for interpretation and usually require additional research. There is huge diversity in the scope of validation rules and their implementation algorithms. Almost every case is unique.

Rule ID	Message	Affected Studies	Issue Rate
CT2002	Variable value not found in extensible codelist	99.4%	38.4%
SD1076	Model permissible variable added into standard domain	98.9%	10.0%
SD1149	Expected variable with missing value for all records	92.2%	38.0%
SD1117	Duplicate records	85.1%	11.4%
SD1078	Permissible variable with missing value for all records	84.5%	33.8%
SD0021	Missing End Time-Point value	80.0%	21.4%
SD0022	Missing Start Time-Point value	79.9%	14.4%
SD1339	Missing EPOCH value when a start or observation date is provided	75.5%	20.3%
SD0026	Missing value for --ORRESU, when --ORRES is provided	68.3%	29.9%
SD0029	Missing value for --STRESU, when --STRESC is provided	67.9%	31.3%
SD0031	Missing values for --STDTC, --STRF and --STRTPT, when --ENDTC, --ENRF or --ENRTPT is provided	64.0%	21.4%
SD0080	AE start date is after the latest Disposition date	61.1%	28.7%
SD1097	No Treatment Emergent info for Adverse Event	58.6%	81.4%
SD1201	Duplicate records in domain	58.2%	8.7%
SD0002	NULL value in variable marked as Required	56.8%	25.6%
SD1124	Missing value for --REASND, when --STAT is 'NOT DONE'	55.7%	51.7%
SD1320	Missing value for --STRESC, when --STAT is null	54.6%	19.6%
SD2239	Inconsistent value for --TPT	53.8%	8.5%
SD0006	No baseline flag record in Domain for subject	50.8%	40.4%
SD0065	USUBJID/VISIT/VISITNUM values do not match SV domain data	49.5%	12.2%

Table 4. 20 most common issues in SDTM data

As expected, CT2002 issue is present in almost every study. Additional diagnostics metrics for CT2002 issue can focus on specific domains and variables.

Rule ID	Message	Dataset	Variable	Affected Studies	Issue Rate
CT2002	LBTEST value not found in Laboratory Test Name extensible codelist	LB	LBTEST	84.2%	8.9%
CT2002	LBTESTCD value not found in Laboratory Test Code extensible codelist	LB	LBTESTCD	83.8%	8.8%
CT2002	LBORRESU value not found in Unit extensible codelist	LB	LBORRESU	67.9%	14.2%
CT2002	CMDOSU value not found in Unit extensible codelist	CM	CMDOSU	65.1%	9.7%
CT2002	CMDOSFRQ value not found in Frequency extensible codelist	CM	CMDOSFRQ	63.8%	9.5%
CT2002	LBSTRESU value not found in Unit extensible codelist	LB	LBSTRESU	58.9%	11.4%
CT2002	RACE value not found in Race extensible codelist	DM	RACE	57.8%	5.5%

Rule ID	Message	Dataset	Variable	Affected Studies	Issue Rate
CT2002	QSCAT value not found in Category of Questionnaire extensible codelist	QS	QSCAT	56.4%	82.9%
CT2002	EGTEST value not found in ECG Test Name extensible codelist	EG	EGTEST	51.6%	38.2%
CT2002	EGTESTCD value not found in ECG Test Code extensible codelist	EG	EGTESTCD	51.5%	37.4%

Table 5. 10 most common variables with CT2002 issue

When analyzing each reported case, a major potential source of reported issue may be different. For example, for CMDOSU a major source of non-standard terms is due to common practice to collect this info as free text format or “Other, Specify”. In the case of QSCAT variable it looks like potential deficiency in CDISC CT or ignoring this standard codelist by users.

A good source of potential candidates for improvement of rule algorithms is metrics available in Enterprise Issue Management system. Here are metrics for validation issues explained as False Positive (FP) finding.

Rule ID	Message	Studies
CT2002	Variable value not found in extensible codelist	11.48%
SD1076	Model permissible variable added into standard domain	5.19%
SD1082	Variable length is too long for actual data	4.86%
SD0058	Variable appears in dataset, but is not in SDTM model	2.43%
SD1320	Missing value for --STRESC, when --STAT is null	1.76%
SD1201	Duplicate records in domain	1.67%
SD1117	Duplicate records	1.48%
SD0047	Missing value for --ORRES, when --STAT or --DRVFL is not populated	1.33%
SD1078	Permissible variable with missing value for all records	0.86%
SD0002	NULL value in variable marked as Required	0.81%

Table 6. 10 most common issues explained as False Positive findings

In 11.5% of studies users explained CT2002 issue for extensible CT as False Positive. While our research show that in many cases validation results represent actual violation and invalid extension of CDISC CT, improvements of CT2002 rule algorithm and diagnostics for reported issues are still needed.

Note that there is common misinterpretation for scope of validation rules. Almost all issues in Table 6 are classified by Pinnacle 21 as Warnings which means they represent potential issues which require manual review and confirmation. Some part of validation cannot be fully automated. Some rules are expected to both false positive and false negative results like clinical diagnostics tests. For example, SD0029 (*Missing value for --STRESU, when --STRESC is provided*) cannot be fully automated because Lab Test Terminology is extensible, and we cannot identify all lab tests whose results have units in advance. So, manual review of SD0029 issues is needed to confirm correct implementation. While such work and explanations for SD0029 issues can be annoying, ignoring or removing SD0029 rule may introduce high risk to study data review process. If lab test results that should have units are submitted without units to FDA/PMDA, the data’s reviewability is likely to be significantly impacted and might result in information requests and/or delays in approval. See [5] for more information about false positive findings.

Here are some metrics for ADaM most common issues.

Rule ID	Message	Affected Studies	Issue Rate
CT2002	Variable value not found in extensible codelist	76.7%	8.9%
AD1012	Secondary custom variable is present, but its primary variable is not present	70.7%	33.6%
AD0018	Variable label mismatch between dataset and ADaM standard	48.1%	9.4%
DD0101	Missing define.xml file	33.6%	100.0%
DD0084	Referenced File is missing	30.6%	100.0%
AD1026	Traceability rules not executed due to missing EX dataset	26.2%	100.0%
AD0196	Required Variable value is null	24.9%	21.6%
AD1024	Traceability rules not executed due to missing DM dataset	24.5%	100.0%
AD1025	Traceability rules not executed due to missing AE dataset	21.9%	100.0%
SD0037	Value for variable not found in user-defined codelist	20.3%	40.3%
AD0149B	Inconsistent value for AVALC	17.3%	2.6%
AD0154	Multiple baseline records exist for a unique USUBJID, PARAMCD, BASETYPE	16.8%	34.6%
AD0124	Inconsistent value for PARCATy within a unique PARAMCD	13.9%	19.7%
AD0253	Record key from SDTM AE is not traceable to ADaM ADAE (not enough ADAE recs)	13.4%	11.8%
AD0320	Non-standard dataset label	12.2%	100.0%
AD0221	Inconsistent value for *CATy	11.2%	6.5%
AD1011	Secondary variable is populated but its primary variable * is not populated	11.1%	25.0%
AD0225	Calculation issue: PCHG != (AVAL - BASE)/BASE * 100	10.9%	8.8%
SD1231	Variable value is longer than defined max length when value-level condition occurs	10.7%	49.8%
DD0085	Missing Define XSL	10.5%	100.0%

Table 7. 20 most common issues in ADaM data

In general, ADaM data is cleaner compared to SDTM because most rules are about ADaM compliance and do not include many data quality checks. Many reported issues are about study metadata and may indicate that data package has not yet been finalized. Metrics for some rules from ADaM Top 20 Issues show that their refinement is needed.

Another useful metric is a list of validation rules which never failed.

Rule ID	Message
AD0039	*DTF variable is not in DATEFL codelist
AD0133C	Calculation issue: AyLO = 0 but R2AyLO is populated
AD0169	CNSR is not an integer >= 0
AD0382	Inconsistent value for BTOXGRN within a unique PARAMCD

Rule ID	Message
DD0008	Element in wrong position within Define.xml
DD0062	Duplicate xml:lang
OD0022	Duplicate Study OID
SD1235	Neither SPDEVID nor USUBJID values are populated
SD1275	--TESTCD equals 'MULTIPLE'
SD1355	Missing TM dataset, when variable MIDS is present in a dataset
SD1361	ARM is populated, but ARMCD is null, or vice versa

Table 8. Examples of never-failed rules (Enterprise)

Table 8 includes some examples of such rules which reported in ADaM (91 rules), SDTM (33), SEND (35) and Define-XML (11).

It helps to identify bugs in algorithms (False-Negative Error). However, there are some CDISC rules applied to rare exotic variables which are not yet utilized by the industry. This is a common case in ADaM. There are also many trivial CDISC validation rules.

CONCLUSION

Review of collected metrics identified several challenges for their practical utilization.

The most important is to remember that there are always different use cases represented in collected metrics due to different processes around study data processing and across different companies. We need to identify such different cases and find their unique attributes which can be used for filtering.

Here are some common examples of processes and suggested filters for metrics:

- Validation of empty datasets (# of records = 0, # of datasets != 0)
- Validation of define.xml file (# of datasets = 0)
- Validation of Trial Design (TD) domains (# of subjects = 0, # of records != 0, # of datasets != 0)
- Validation of ongoing studies (some specific data issues?)
- Validation of finalized studies (define.xml and TD domains are included, the last validation)
- System testing (too many records)
- Training (public data)

An example of Training can be represented by the 10 most common studies in Community which have 10K validations in total. Of course, ‘CDISC01’ and ‘CDISCPILOT01’ studies are on this list.

There are some unique business cases which may impact overall metrics. For example, some studies were validated 1.5K+ times. Other studies include 150+ empty ADaM datasets. Therefore, Max/Min statistics should be avoided for evaluation of the industry-wide metrics. We think that $95^{th}/5^{th}$ percentiles are a better option here. However, Max/Min statistics are completely valid for utilization within the same company with established and controlled processes.

Also, business processes at sponsor, vendor, technology, or educational organizations are expected to be different. We need to understand the potential practical impact to evaluation metrics and make decision if additional filters are required.

Data cleaning is expected to remove invalid or broken records like partially failed validations, special characters in study ID, or unusual business cases. A continuous metrics curation is required if we want to consider them as a reliable source of information for making informed decisions.

There are 2 major types of metrics: company and industry wide. A good example to illustrate the difference is metrics for change in reported validation issues observed over the complete life cycle of study data. For example, SDTM data preparation process includes study build, data collection period, SDTM mapping, ADaM programming, TFLs generation, data unblinding, different data-cuts, finalization, tuning for compliance with the specific regulatory agency. All these common stages may require only a subset of existing validation rules. Metrics can help develop rule filters to serve different study data processing stages. Ruiz [1] showed examples of utilization of company operational metrics measuring the number of validations to issue resolution and time to fix relative to study's major time points. Such metrics were used to define the company's good practices and enforce them. Integration with CTMS is expected, which is not possible for the industry-wide metrics. Similar industry-wide metrics may be based on comparison of first and last validation results for the study. The industry-wide metrics cover very diverse processes across the industry and should be used with good understanding and cautious interpretation of collected data.

Company data validation analytics provides better fit and accuracy for internal processes. The industry-wide metrics are more useful for improvement of industry-wide validation rules and can be used internally as reference benchmarks.

REFERENCES

1. Rodrigo (Ruy) Juarez y Ruiz. 2022. "P21E and Metrics Visual Analysis", *Proceedings of the P21 Live conference, Philadelphia, PA*
2. Sergiy Sirichenko. 2020. "SUPPQUAL Datasets: Good, Bad and Ugly". PharmaSUG 2020. Available at https://www.lexjansen.com/pharmasug/2020/DS/PharmaSUG-2020-DS-261.pdf
3. Sergiy Sirichenko "Industry Metrics for Extensions of CDISC Terminology" PhUSE US Connect 2021. Available at https://www.lexjansen.com/phuse-us/2021/ds/PRE_DS08.pdf
4. Sergiy Sirichenko, Michael Beers. 2021. "Metrics for Laboratory Test Panels Information in SDTM Data" PhUSE CSS 2021. Available at https://www.lexjansen.com/css-us/2021/POS_PP01.pdf
5. Kristin Kelly, Michael Beers. 2019 "The Truth about False Positives". PhUSE CSS 2019. Available at https://www.lexjansen.com/css-us/2019/PP19.pdf

CONTACT INFORMATION

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

Sergiy Sirichenko
Pinnacle 21 LLC
sergiy@pinnacle21.com
www.pinnacle21.com