Trends in eSub Evaluation Findings: A Twenty-Year Perspective on Data Traceability

Lisa Brooks, Iris Statistical Computing, Foster City, CA, USA
Piper Brooks, Iris Statistical Computing, Corvallis, OR, USA
Ann Shiba, Iris Statistical Computing, San Jose, CA, USA

ABSTRACT
Iris has evaluated hundreds of study data eSubmission packages over the last 20 years. These reviews utilized the publicly available validation software available at the time. There was a focus on checking the qualitative aspects of a compliant submission. These checks were informed by standards, guidance, industry publications, interactions with Food and Drug Administration (FDA) and expert debate in community forums. This paper will share trends in our findings over the years and provide insight into persistent issues that Biometrics Groups and Clinical Research Organizations are struggling to get right in the draft deliverables. We observed a consistent downward trend in traceability-related error rates over time, particularly in ADaM packages following FDA enforcement in 2021; however, persistent errors in origin and derivation documentation remain. This analysis has shown that those that create define.xml for submission would benefit from training as well as improved QC processes.

INTRODUCTION
This paper will provide a brief historical review of US FDA data standards, highlighting quality of data eSubmission packages and data traceability. Key FDA presentations in 2021 offered insight into evaluation of data quality traceability issues. Since then, information on this topic has been sparse. Recent FDA presentations have focused on the number of studies with technical rejection errors.

We ask the question: “Has the quality of electronic data submissions improved since FDA endorsed CDISC® standards?” Since 2007, Iris has reviewed over 356 CDISC data packages for issues with traceability, readability, and technical conformance. We utilize available validation software, but do not replicate their checks. In a review and analysis of comment logs, we identified prevalent errors and trends over time. We used string search and categorization of comments to identify our key focus of analysis. Here we will describe methodology, provide our findings, and summarize our conclusions. This evaluation of data issues is a granular, variable-level evaluation of eSubmission errors compared to FDA’s reporting on high-level technical rejections.

US FDA HISTORY OF DATA STANDARDS

The Advent of Electronic Clinical Trial Records
The publishing of FDA’s 21 CFR Part 11 Regulation in March 1997 paved the way for the transition from slow, paper-based clinical trial data collection to Electronic Data Capture (EDC) in the late 1990s. This regulation established criteria for trustworthy, reliable, and equivalent electronic records and signatures to paper ones, ensuring data integrity, authenticity, and security. FDA provided further instruction in the Part 11 Guidance published in August 2003 and attempted to ease the way by emphasizing the “predicate rules” for data reliability and focusing on validation, audit trails, security, and record integrity.

FDA Endorses CDISC Standards in 2003
In 2003, FDA endorsed the use of the Standard for Exchange of Nonclinical Data (SEND) by announcing a pilot project. This launched decades of collaboration between FDA, CDISC, PHUSE, and other organizations in the development and implementation of clinical and nonclinical standards. The pivotal “CDISC/FDA Integrated Data Pilot” launched in 2008 demonstrated that CDSIC data benefits and supports the requirements for FDA reviewers to conduct the needed evaluations across studies and compounds and aids in integration. In the case study published by Decker for PhUSE 2009, it is notable that traceability was identified as an area of process and terminology development. He wrote, “One gap identified within the first Pilot was the ability to provide traceability from the ADaM data sets to the source data. After a significant amount of discussion within the Pilot Team and in conjunction with the ADaM team no good answer was identified.”

Throughout the 2000s, CDISC developed the foundational standards. Validation software became available from organizations like Phase Forward, Pinnacle 21, and Formedix. In 2014, FDA published CDISC validation rules and the Study Data Technical Conformance Guide (TCG). Appendix 8 of the guide was dedicated to the discussion of support of validation rules and data traceability. The appendix established the requirement to document and explain

validation errors in the submission study data reviewer's guide. In the document, FDA identified data traceability as the most problematic issue and that they require that all submitted electronic data be traceable from its original source to the final analysis results.

FDA Mandates CDISC Standards in 2016
FDA mandated the use of data submissions in standardized and CDISC format in a federal register notice on September 17, 2016. In November of 2016, the FDA published the Technical Rejection Criteria (TRC) for study data. TRC are automated validations by CDER or CBER executed when electronic clinical trial data are submitted and processed by FDA inbound systems. TRC included requirements for specific datasets (trial summary dataset, demographics and subject level analysis data) and data definitions in the form of a define.xml for both tabulation data and analysis data.

During this period, there was an expansion of the idea that records should be "authentic and reliable" to a broader concept that the "data traceability" from source to analysis endpoints (and *vis-a-versa*) be well understood. In the 2014 TCG, FDA laid out its desire to understand the construction of analysis data, including derivation of variables and statistical calculations. The TRC required that define.xml be submitted with electronic data to this end. That same year, Pinnacle 21 launched its version 2.0 software that not only validated define.xml but generated it. It was during this mandated period that industry saw progression from home grown software solutions to digitalization, interoperability, and automation. During this time the existing standards and terminology were being revised and expanded.

FDA Enforces CDISC Standards in 2021
FDA enforced the use of electronic data submissions in CDISC data and accompanied by define.xml in a federal register notice July 29, 2021, with an effective date of September 15, 2021. With this enforcement, FDA would reject submissions upon transmittal that did not meet the high validation errors outlined in the TRC.

It was anticipated that, with this new era of enforced standardized and validated clinical trial data, electronic data submissions would be enhanced. With technological advancement and robust standards, workflow has been improved, bringing faster delivery of standardized data and analysis output. There is the expectation that there will be less variability in submission quality with fewer errors. In an FDA presentation "Study Data Rejections in eCTD Submissions", Crandall reported that only 1% of submissions are rejected with the majority of those being from nonclinical data. From a review of online presentations and papers submitted in the past few years, there has been a transition from focusing on what is going wrong with submissions to updates regarding modernizing transport formats, real world data, and artificial intelligence. Important presentations in 2021 were published by FDA on the JumpStart data assessment program, one for Traceability, and one for Quality findings. Their conclusion was that there was a wide range of traceability findings and specific frequently occurring CDISC conformance issues. Based on our findings, industry would benefit from regular updates from the team at FDA responsible for evaluating standard data submissions similar to what was provided in 2021.

METHODOLOGY

Comment Logs and Sponsor Permission
Beginning in 2007, Iris reviewed clinical standardized submission packages (SDTM and ADaM) based on criteria in a proprietary checklist. This checklist was developed to identify quality issues in data submissions that would not otherwise be picked up in existing validation software checks. The reviews resulted in Microsoft Excel comment logs with a relatively persistent format over the years. Of the 356 studies Iris evaluated since 2007, 289 comment logs were available for review. Iris's authors contacted sponsors to gain permission to use the comment logs. This effort was constrained by corporate acquisitions, personnel changes, and data privacy initiatives. Six out of fourteen sponsors agreed to the use of 179 comment logs for this analysis. Although the comment logs originate from a single organization, the findings reflect recurring issues observed across sponsors, phases, and submission eras.

Compilation and Deidentification
The comment logs were compiled and underwent a deidentification process. All references to sponsor, protocol, subject, endpoint, indication, and vendor name were redacted. Metadata information from the Excel file was also removed. Review of the deidentification was conducted by a second author for accuracy and remediation. Information for each comment log was collected to accompany each log:

- Type of data package (SDTM, ADaM, Legacy Tabulation, Legacy Analysis, SEND, Integrated)
- Clinical Trial Phase (Phase 1, 2, 3, 4, Integrated or nonclinical)
- Period of Review Date (Endorsed, Mandated and Enforced, based on date of first comment)
- Period of Study Start (Endorsed, Mandated and Enforced, based on study start date from ts.xpt - actual study start date not retained)
- Number of Records (observations in the data package – from Pinnacle 21 report)
- Number of Variables (counted from define.xml based on a search for "<ItemDef")

Formatting and Importing
The comment logs were formatted with standardized headers and terminology of package element for ease of import. The excel files were then imported with RStudio. The data was then transformed into a list of data frames.

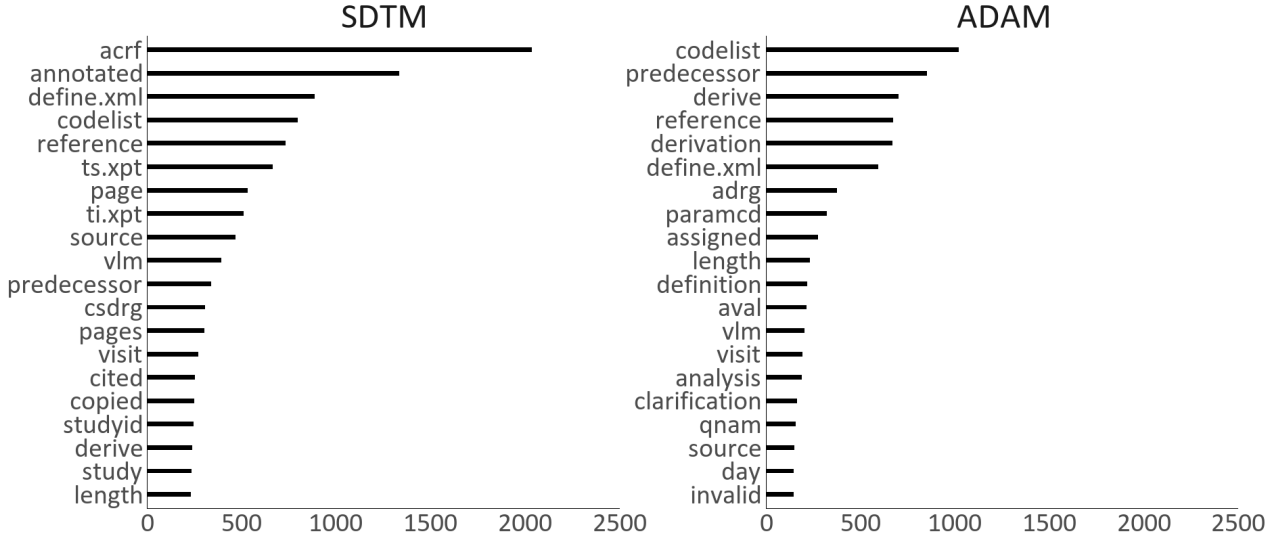
A category variable (Reference Category: REFCAT) was created based on the values of the package element (Reference: REF) to more generally identify what type of dataset package the comment referenced:

- SDTM for REF=TDM, SDTM, SDTM define, acrf, csdrg, or icsdrg
- ADaM for REF=ADaM, ADaM define, adrg, iadrg, or programs
- SEND for REF=SEND, SEND define, or nsdrg
- Legacy Tabulation for REF=Legacy Tabulation acrf, Legacy Tabulation define, or Legacy Tabulation data
- Legacy Analysis for REF=Legacy Analysis define or Legacy Analysis data
- MISC for REF=MISC or miscrg

Identifying Prevalent Terms and Selecting Focus of Analysis
The words in the comment variable were tokenized individually to search and identify the top 50. A list of all unique words was created and sorted by prevalence. Common words were filtered out using the RStudy tidytext package and the data frame stop_words. The top 50 instances of unique words based on count were flagged. Similar words were harmonized, for example, “ts” and “ts.xpt” were set to “ts.xpt” and “derive” and “derived” were set to “derived”. This list was used to create a word cloud to identify prevalent words used in the comment logs:

To create Figure 1, the top 20 words were counted for each of the reference categories, SDTM and ADaM. The following plots were created using R ggplot.

Figure 1: Top 20 words in SDTM and ADaM Package Comments



The predominant terms were clearly associated with concepts of data traceability (I.e., acrf, annotated, reference, page, source, vlm, predecessor, pages, derive, derivation, assigned, definition, clarification, and source). Data traceability comments included issues where define sources were insufficient to identify origin. Based on the preponderance of these types of terms, we focused our attention on analyzing these comments.

To categorize issues within the comment logs, data mining was conducted using string search terms. Three variables were created issue category (IssueCat), issue code (IssueCd), and issue text (IssueText) based on the term search. The results of the data mining were reviewed by two programmers to ensure there was not a pattern of misclassifications. For brevity, we include in Table 1 below the IssueCat codes and text our analysis ultimately pinpointed: “Define Sources Insufficient” along with an example of the search term logic:

Table 1: Subcategories and Example Search terms for ‘Define Sources Insufficient’ Category

IssueCD	IssueText	Example Search Term Logic
Defacrf	aCRF Origin Errors	"not found on acrf" "not found in define" "not found p"
Defder	Derivation Errors	REF=”SDTM define” or “ADaM define” and “unclear” “incomplete” “doesn’t match data” “need definition”
DefRaw	Source not in submission	"non sdtm source data" "variable not found in study" “raw”
DefVLMO	VLM multiple origin	“multiple origin”
DefVLMP	Define VLM not sufficient	“vlm”
OrInc	Origin Errors	“predecessor” “derived” “assigned” “need domain prefix” “need definition/derivation” “unresolved source”

The data frame was converted into a JSON file and then imported into a SAS dataset. The primary programming was done in R, whereas the validation programming was conducted in SAS.

Planned Analysis Periods Using Study Start Date

Analysis periods were selected based on key milestones in FDA data standards policies (Endorsed, Mandated, Enforced). In choosing between study start or evaluation date to determine the analysis period, study start was selected. The rationale was that packages were created at the time of study conduct and according to the FDA data standards policies based on the time of study start. Whereas evaluation date was often delayed to the time of New Drug Application filing. Below are the Analysis Periods:

- Endorsed: Jan 2003 – Nov 2016
- Mandated: Dec 2016 – Aug 2021
- Enforced: Sep 2021 – Dec 2025

Error Rate Calculation

In clinical research, particularly data management, the denominator for Error Rate would be the total number of populated fields. Here is the basic calculation:

$$\mathbf{ER} = \frac{\textbf{Number of Errors Found}}{\textbf{Total Number of Fields}} \textbf{ X 100}$$

For our purposes, the total number of variables is more meaningful as errors in traceability are on a per-variable basis. Based on this concept we modified the error rate calculation to be:

$$\mathbf{ER} = \frac{\textbf{Number of Traceability Errors}}{\textbf{Total Number of Variables in Define.xml}} \textbf{ X 100}$$

Error rates were calculated for each comment log. One for each reference category (REFCAT) of SDTM and ADaM, as well as SDTM and ADaM combined. Error rates were summarized by period based on study start date (Endorsed,

Mandated, and Enforced). Tables and figures were programmed in R. The reference categories and periods were converted into factors so the error rates could be displayed chronologically and by reference category. The maximum values were calculated to create the upper limit for the graph. The mean, median, minimum and maximum, interquartile range, and 95% confidence interval were calculated inside ggplot to generate the resulting box and whisker plots and tables. The plots are faceted to show the error rates for the periods and reference categories. The plots were formatted for readability by jittering the outliers, adjusting box and font sizes, and ensuring that the y axis was the same for all plots while retaining all outliers. Summary tables were generated to compare the validation output generated by proc glm in SAS.

The criteria for inclusion in the analysis of "Define Sources Insufficient" was:

- data package had to be SDTM or ADaM (not SEND or Legacy data)
- define.xml was in the package to count the number of variables and calculate an error rate

One hundred and one studies were included in this analysis. Table 2 shows the number of studies and dataset packages included in the analysis.

Table 2: Number of Studies and Data packages for each Period Based on Study Start Date

Period	# of Studies	#of SDTM Packages	# of ADaM Packages
Endorsed	14	14	3
Mandated	54	54	53
Enforced	33	28	31

FINDINGS

Figure 2 and Table 3 show the mean error rate for 'Define Sources Insufficient'. In general, the mean trended downwards from Endorsed to the Enforced period. The trend was more pronounced for the ADaM packages. There were a high level of variability and several outliers. Of note, there were less outliers in the Enforced period, indicating that packages with a high number of errors were less likely earlier in industry experience. This can be interpreted as a maturation of submission quality control processes.

Figure 2. Mean Error Rate for 'Define Sources Insufficient' Category

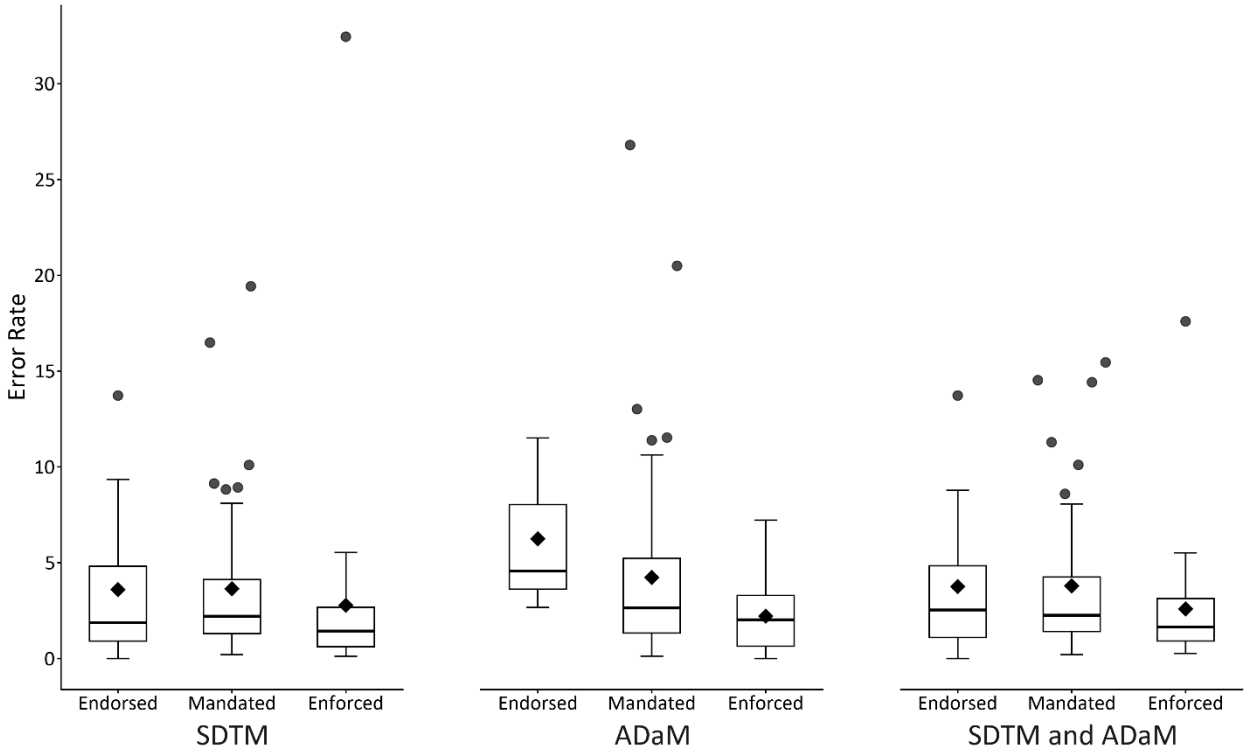


Table 3: Mean Error Rate for 'Define Sources Insufficient' Category

	SDTM			ADaM			SDTM and ADaM		
	Endorsed	Mandated	Enforced	Endorsed	Mandated	Enforced	Endorsed	Mandated	Enforced
N	14	54	28	3	53	31	14	54	33
Mean	3.6	3.6	2.8	6.2	4.2	2.2	3.8	3.8	2.6
95% CI	(1.3, 5.9)	(2.6, 4.7)	(0.5, 5.1)	(-5.3, 17.8)	(2.8, 5.6)	(1.5, 2.9)	(1.6, 5.9)	(2.8, 4.8)	(1.5, 3.7)

We reviewed the comment logs to identify the predominant issues contributing to the outliers. The 13 studies represented in the outliers for SDTM or ADaM are described in Table 4. We have provided a description of the errors that occurred most. Based on the review of the errors, we can see that there were:

a. Gross errors in listing aCRF page numbers in the origin.
b. A fundamental misunderstanding of the definition of origin in SDTM and ADaM.
c. ADaM computational methods not reviewed by a statistician.

Regardless of the reason, there were 13 in 101 data packages with an extreme number of errors.

Table 4: Studies With Outlier Error Rates and Their Predominate Errors

Reference Category	Phase	Period Based on Start Date	# of Issues	Error Rate	Predominant Errors
SDTM	P1	Mandated	70	8.9	31 pointed to wrong CRF page. 17 incorrect origin. 14 unclear computational methods.
SDTM	P1	Mandated	99	9.1	52 unclear computational methods. 39 pointed to wrong CRF page.
SDTM	P2	Endorsed	66	9.3	33 pointed to wrong CRF page. 21 unclear computational methods.
SDTM	P1	Mandated	102	10.1	96 were pointing to wrong CRF page.
SDTM	P1	Endorsed	7	13.7	Low number of variables compared to other packages. Possible partial submission.
SDTM	P1	Mandated	107	16.5	91 missing computational methods.
SDTM	P3	Mandated	245	19.4	97 pointed to wrong CRF page. 87 origins were to external data. 48 unclear computational methods.
SDTM	P2	Enforced	517	32.5	327 define.xml was based on SDTM spec referencing raw EDC variables. 127 pointing to wrong CRF page.
ADaM	P1	Endorsed	180	11.5	88 unclear computational methods. 87 source not set to predecessor, predecessor doesn't match data, or incorrect source.
ADaM	P1	Mandated	49	11.5	31 origins fields not set to predecessor or assigned correctly. 18 unclear computational methods.
ADaM	P2	Mandated	227	13.0	169 unclear computational methods. 50 origins not set to predecessor or assigned correctly.
ADaM	P3	Mandated	199	20.5	115 origins not set to predecessor. 62 unclear computational methods.
ADaM	P2	Mandated	320	26.8	268 origins not set to predecessor. 44 unclear computational methods.

The root cause of these outliers reveals that there are gaps, not only in knowledge, but also in quality control processes. There is a clear misunderstanding of the origins of predecessor, assigned, and derived. Lack of statistical review results in unclear computational methods. The high number of errors with linking variables to the aCRF indicate a lack of review.

To gain a better understanding of the issues behind the category of 'Insufficient Data Sources', whisker plots were generated for each of the sub-categories. Figure 2 shows three whisker plots that were picked to display the errors driving the trends. The three types of errors in the comment logs that contributed to the error rate were SDTM aCRF Origin Errors, ADaM Origin Errors, and ADaM Derivation Errors.

Figure 3: Mean Error Rates for Categories Driving Trend

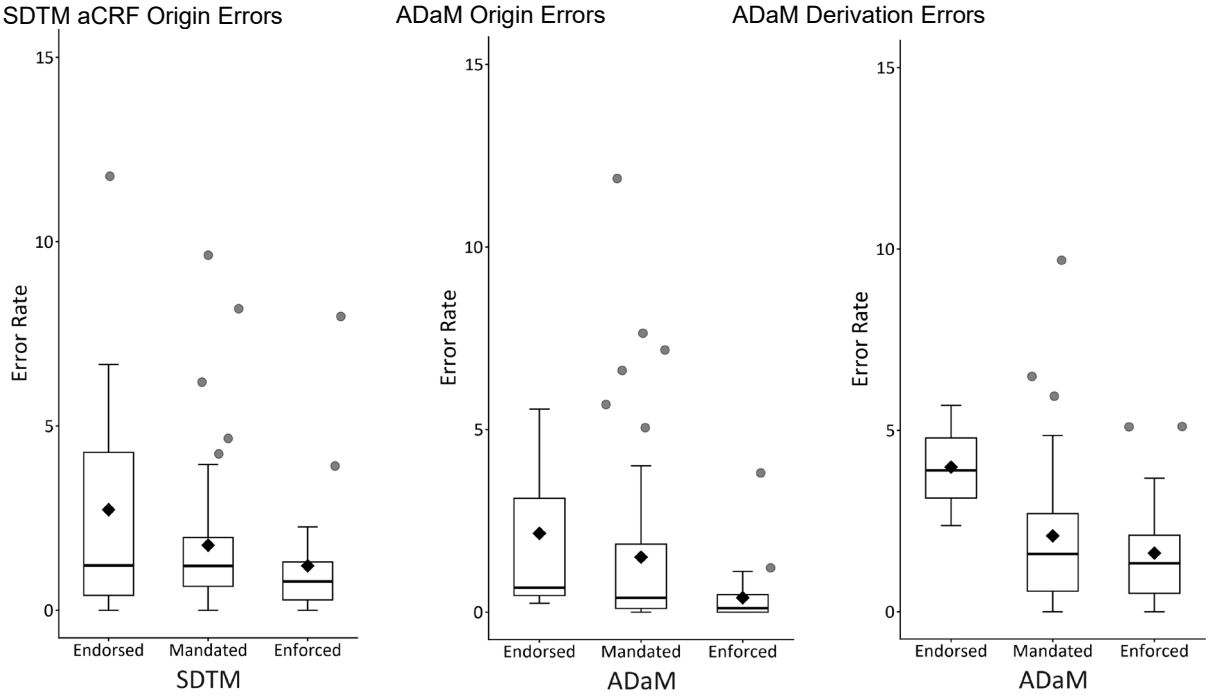


Table 5: Mean Error Rates for Categories Driving Trend

	SDTM aCRF Origin Errors			ADaM Origin Errors			ADaM Derivation Errors		
	Endorsed	Mandated	Enforced	Endorsed	Mandated	Enforced	Endorsed	Mandated	Enforced
N	14	54	28	3	53	31	3	53	31
Mean	2.7	1.8	1.2	2.2	1.9	0.4	4.0	2.1	1.6
95% CI	(0.8, 4.7)	(1.2, 2.3)	(0.6, 1.8)	(-5.2, 9.5)	(0.7, 2.9)	(0.1, 0.7)	(-0.1, 8.1)	(1.5, 2.6)	(1.1, 2.1)

SUMMARY

The mean error rates in CDISC data packages in this analysis have trended down from the Endorsed to the Enforced period for both SDTM and ADaM packages as shown in Figure 2 and Table 3. That trend is also evidenced when looking at the highest contributing sub-categories to the error rates shown in Figure 3 and Table 5. Errors occurring in ADaM dataset packages during the mandated period were the highest in the analysis (Mean: 4.2), but saw a reduction in the enforced period (2.2).

Limitations of Analysis

This analysis is intended to inform discussion and improvement, not to benchmark or rank industry performance.

There was a limited number of comment logs that had ADaM in the Endorsed period. An explanation is that sponsors were more likely to submit legacy data converted to SDTM when CDISC wasn't mandated, keeping their analysis data in legacy format to support the CSR.

Due to considerable variation in the error rates, the mean may not be a reliable representative value. We present an observational, exploratory analysis of error rates over time. It used a longitudinal, descriptive approach to identify patterns, trends, and anomalies. Instead of testing a specific hypothesis, this analysis aimed to generate hypotheses about why error rates change by visualizing and summarizing historical error rates.

Error rates in this analysis are not typical of those found in clinical research. Garza et al. (2025) reports in her meta-analysis that the pooled error rates for optical scanning, single-data entry, and double-data entry methods were 0.74% (95% CI: 0.21, 1.60), 0.29% (0.24, 0.35) and 0.14% (0.08, 0.20), respectively. Medical records abstraction (MRA) had higher and more variable error rates, with the pooled error rate of 6.57% (95% CI: 5.51, 7.72). In this

analysis, the error rates are much higher but that may be a result of our differing calculation of error rate by using a count of variables and not a count of populated fields. Regardless, the error rates reported here are sufficient for comparing within this analysis, but not to error rates used in Garza's paper.

CONCLUSION

Has the quality of electronic data submissions improved since FDA endorsed CDISC standards? The answer is yes. We observed a downward trend in error rates related to data traceability in ADaM packages over time. Over the periods of Endorsed, Mandated, and Enforced, our analysis shows that the quality of clinical trial data submissions have improved. Following the enforcement of standardized data submission requirements in 2021, the trend is more pronounced. This suggests that regulatory enforcement of Technical Rejection Criteria, maturing standards, and increased understanding of standards over time created a measurable positive impact on submission quality.

Despite the improvements over time, it is apparent that there are still issues with traceability in eSubmission packages. Errors such as SDTM aCRF Origin Errors, ADaM Origin Errors, and ADaM Derivation Errors continue. Although the outliers became less frequent during the enforced period, their continued existence underscores the issue of traceability errors. The errors in pointing to the wrong aCRF page are easily corrected with a review. Computational method errors indicate a need to increase statistical oversight. The use of incorrect origins (assigned, predecessor, derived) points to gaps in understanding of foundational define concepts.

Although improvements have been made, traceability is still a persistent challenge for the industry. To ensure submission quality, technical conformance alone is not sufficient. While validation software can detect and identify many structural issues, it cannot fully assess data origins, derivations, and relationships that foster reviewer's readability, understanding, and confidence. To remedy this issue, industry can continue to invest in education, process rigor, and multidisciplinary collaboration. Doing so can help ensure that evaluation of traceability is built into the quality control process, rather than leaving traceability as a documentation afterthought.

The value of independent, longitudinal review of submission quality trends is demonstrated by this analysis. Critical industry insight into common deficiencies can be gained from periodic updates and public presentations like the FDA's 2021 presentations. This insight can benefit the industry as standards, technologies, and data sources continue to evolve. Transparency and dialogue between regulators and sponsors are essential to improving data quality therefore increasing efficiency and success with regulatory reviews.

What should we do when we return to work after PHUSE US Connect 2026? We should implement earlier traceability review into our processes. This includes more robust quality control around origin definitions and statistical review of computational methods.

REFERENCES

Federal Register (Vol. 62, No. 54, Thursday, March 20, 1997) 21 CFR Part 11 – Electronic Records; Electronic Signatures https://www.govinfo.gov/content/pkg/FR-1997-03-20/pdf/97-6833.pdf

Federal Register (Vol. 68, No. 17, Monday, January 27, 2003) Pilot Project Using CDISC SEND Announcement https://www.govinfo.gov/content/pkg/FR-2003-01-27/pdf/03-1742.pdf

Part-11--Electronic-Records--Electronic-Signatures---Scope-and-Application (August 2003) Part-11--Electronic-Records--Electronic-Signatures---Scope-and-Application-(PDF).pdf

CDISC History https://www.cdisc.org/history

The CDISC/FDA Integrated Data Pilot: A Case Study in Implementing CDISC Standards to Support an Integrated Review (PHUSE 2009) Chris Decker https://phuse.s3.eu-central-1.amazonaws.com/Archive/2009/Connect/EU/Basel/PAP_CD07.pdf

Federal Register (Vol. 80, No. 7, Monday, January 12, 2015) Study Data Technical Conformance Guide and Data Standards Catalog; Availability https://www.govinfo.gov/content/pkg/FR-2015-01-12/pdf/2015-00206.pdf

Federal Register (Vol. 81, No. 183, Wednesday, September 21, 2016) FDA Mandates Clinical Trial Data Submissions in Standardized Format https://www.govinfo.gov/content/pkg/FR-2016-09-21/pdf/2016-22129.pdf

JumpStart Traceability Assessment (PHUSE 2021) Jesse Anderson and Sarah Rodgers, FDA https://phuse.s3.eu-central-1.amazonaws.com/Archive/2021/Connect/US/Virtual/PRE_SA03.pdf

JumpStart Common Data Quality Issues in SDTM Data (PHUSE 2021) Alan Shapiro MD PhD and Michael Xue, FDA https://phuse.s3.eu-central-1.amazonaws.com/Archive/2021/Connect/US/Virtual/PRE_SA04.pdf

Federal Register (Vol. 86, No. 143, Thursday, July 29, 2021) Electronic Study Data Submission; Data Standards; Technical Rejection Criteria for Study Data Effective Date https://www.govinfo.gov/content/pkg/FR-2021-07-29/pdf/2021-16187.pdf

Study Data Rejections in eCTD Submissions (FDA April 24, 2024) Heather Crandall, FDA https://www.fda.gov/media/189617/download

Error Rates of Data Processing Methods in Clinical Research: A Systematic Review and Meta-Analysis of Manuscripts Identified Through PubMed (Int J Med Inform March 2025) Garza et al. https://pubmed.ncbi.nlm.nih.gov/39647291/

ACKNOWLEDGMENTS
The authors would like to acknowledge the anonymous approval by sponsor clients for Iris Statistical Computing to utilize comment logs created under their employment for this analysis. We are very grateful.

A special thank you to Spencer Hudson, who we brought out of retirement for free statistical advice.

RECOMMENDED READING
FDA Study Data Standards Resources: https://www.fda.gov/industry/fda-data-standards-advisory-board/study-data-standards-resources

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Lisa Brooks
Company: Iris Statistical Computing
Email: Lisa@IrisStatComp.com
Website: https://irisstatcomp.com/