



CDISC India Day

Beyond Compliance, What Really Makes a Submission Review-Ready: Lessons Learnt from Regulatory Deliverables.

Nilakshi Srivastava | Priyadharshini Sankar



Speakers



Nilakshi Srivastava, who brings around three years of clinical statistical programming experience to her current role of Statistical Data Scientist at Pfizer, where she primarily supports Inflammation & Immunology (I&I) studies. She is passionate about leveraging statistical programming and machine learning techniques to generate data-driven insights and enhance clinical research outcomes. Her core professional goal is to continually grow as a data scientist by taking on complex, challenging roles that drive strategic organizational impact.



Priyadharshini Sankar is a Senior Statistical Data Scientist with over 5 years of experience in the clinical research industry. She specializes in statistical programming, clinical data analysis, and end-to-end clinical study delivery across multiple phases of drug development. With expertise spanning diverse therapeutic areas, she has successfully supported global clinical trials by delivering high-quality analyses and regulatory-compliant reporting. She is passionate about leveraging data-driven insights to advance clinical research and collaborates closely with cross-functional teams to drive efficient and impactful study outcomes.

Agenda

1. Why Compliance is not enough?
2. Reviewer's Lens
3. Friction Points
4. Practical Case Studies
5. Summary and Closing Remarks





Why Compliance is not enough?



Navigating the Compliance Gap: True Submission Readiness

Achieving compliance with CDISC standards is a critical milestone, but compliance alone does not guarantee a successful regulatory review. True submission readiness requires a holistic combination of standard data structures and human-focused reporting context.

Validation Blind Spot:

Organizations often focus heavily on automated validation outcomes (such as Pinnacle 21), while completely overlooking the descriptive factors that directly influence human reviewer understanding and efficiency.

High-Quality Datasets

Pristine SDTM & ADaM

Robust Traceability

Unbroken data lineage

Well-Defined Metadata

Roadmap via Define.xml

Clear Documentation

Strategic Reviewer Guides

Clinical Programming: The Mindset Shift

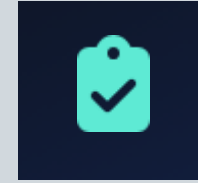
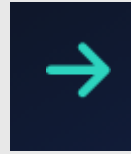
From Analysis-Ready to Submission-Ready Early in the Study Lifecycle



Analysis-Ready

Optimized for Statistical Modeling

- ✓ Drives rapid internal insights
- ✓ Flexible and optimized for efficiency
- ✓ Simplifies data exploration



Submission-Ready

Meets Strict Regulatory Standards

- ✓ Strict structure and documentation
- ✓ Ensures end-to-end data traceability
- ✓ Reproducibility required by regulators globally



Reduce Regulatory Risk



Improve Submission Efficiency



Support Parallel Global Filings

Compliance ≠ Review Readiness

A structural comparison: Passing the rules vs. enabling the human reviewer

Aspect	Technical Compliance	True Review Readiness
Validation Goal	Passes automated structural checks (e.g., Pinnacle 21) with findings resolved or documented as appropriate.	Goes beyond algorithms; the data structure logically supports the clinical narrative and analysis.
Traceability	Mechanical variable links between source data and target datasets technically exist.	End-to-end data lineage is crystal clear, transparent, and easy for an auditor to follow.
Documentation	Provides basic structural metadata (e.g., standard Define.xml) to satisfy upload requirements.	Comprehensive Reviewer Guides (cSDRG/ADRG) proactively explain context, anomalies, and decisions.
Target Audience	Software validation engines and automated compliance algorithms.	Human regulatory reviewers at global health authorities evaluating the submission.



Reviewer's Lens

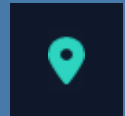


Think Like a Reviewer: 5 Questions for Review Efficiency



WHAT

What does this data represent? Clear structural predictability and alignment with the clinical narrative (CSR).



WHERE

Where did this data come from? Seamless traceability back to raw source data and CDISC SDTM domains.



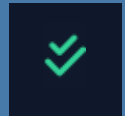
HOW

How was this derived? Transparent methodologies and derivations properly documented in metadata.



WHY

Why was this decision made? Contextual explanations of data anomalies or imputation rules in the Reviewer Guides.



CAN I VERIFY?

Can I independently confirm this? Effortless reproducibility of primary efficacy and safety claims using provided programs.

One Chain: From data standards to regulatory evidence

Raw Data

Clinical study data • endpoints • safety • exposure

SDTM

Tabulation-ready standardized representation of collected data

ADaM

Analysis-ready datasets with analysis variables and derivations

TLFs

Tables • Listings • Figures • statistical outputs

**Submission
Package**

Define-XML • aCRF • cSDRG • ADRG • supporting documents



Friction Points



The Annotated CRF (aCRF) Roadmap

The critical visual bridge between data collection and SDTM tabulation.



1. Blank CRF

The original data collection instrument shown to site investigators.



2. SDTM Annotation

Mapping each field to SDTM domains, variables, and controlled terms.



3. Define.xml Linkage

Origin metadata points directly to specific aCRF pages.



4. Reviewer Navigation

Reviewers click from Define.xml straight to the exact aCRF page.

aCRF Pitfalls & Reviewer Friction

Common annotation errors that break the traceability chain and frustrate reviewers.



Broken Bookmarks & Links

When Define.xml links point to the wrong page or the aCRF lacks a functioning table of contents, reviewers cannot verify data origins efficiently.



Sloppy Annotations

Text boxes that overlap, obscure the original CRF text, or use inconsistent colors. The original data collection context must remain fully legible.



Missing RELREC Annotation References

RELREC references are missing, limiting traceability and documentation of relationships across SDTM domains.



Leaving Blank "Ghost" Fields

Leaving operational/test fields blank without explicitly annotating them as "[NOT SUBMITTED]".



Omitted Amended Form Versions

When study protocols amend the eCRF, failing to annotate all unique versions of a form leaves reviewers guessing how mid-study data was collected.



Omitted Coded Values

Mapping the variable name but omitting specific checkbox or radio button values (e.g., VSTESTCD="SYSBP").



Inconsistent Annotation Color

Inconsistent colors are used for domain and variables annotations, reducing aCRF readability and review efficiency



Random Color Coding Schema

Mixing annotation box colors for domains and variables randomly without a standardized color structure.

Define-XML

The Metadata Roadmap for Clinical Trial Data Review

Define > Datasets > Variables > VLM > Codelists > Dictionaries > Method > Comments > Documents

Define

What study, standards, and metadata are being described?

Datasets

Which datasets exist and what is the purpose of each dataset?

VLM

Where do variable definitions vary by parameter, test, or condition?

Variables

What does each variable in each dataset represent and where did it originate?

Codelists

Which controlled terminology and allowable values used?

Methods

How were derived values calculated?

Comments

What additional context is needed to interpret the data?

Documents/Dictionaries

Which supporting documents provide traceability and explanation? Which standards and CT versions are referenced?

The Metadata Blueprint: Define.xml Friction Zones

Methodology & Linking Gaps

-  **Wrong / Stale Version References** Dataset uses one standard/CT version while Define.xml points to another, triggering technical validation findings, reviewer queries, or submission-quality concerns.
-  **Missing Methods** A derived variable exists in the dataset, but the critical derivation method is absent, vague, or incomplete in the metadata.
-  **Broken Codelist Links** Values are present in the dataset but the codelist or code system relationship is entirely unclear or missing in Define.
-  **Missing Define.xml Stylesheet** Omitting the standard stylesheet (XSL) makes the XML file difficult for human reviewers to navigate in a browser-like view.

Structural Mismatches

-  **Incorrect Origins** The stated origin (e.g., Assigned vs. Derived) is not aligned with how the value was actually obtained or calculated.
-  **Value-Level Gaps** Parameter-specific definitions (VLM) are omitted for transposed domains, making generic variable labels highly misleading.
-  **Dataset / Metadata Mismatch** Variable names, labels, lengths, types, or assigned values in the metadata simply do not match the actual physical XPT file.
-  **Hyperlink Verification Failures** Failing to ensure all table of contents, bookmarks, and links to appendices or the aCRF function correctly prior to submission.

cSDRG and ADRG Framework: Navigating the Lineage

SDTM Tabulation

Raw clinical database structured into predictable domains.



cSDRG Key Contents

- ✓ **Standards Versioning:** Explicitly states SDTMIG, MedDRA, and WHODrug versions utilized.
- ✓ **Trial Design Mapping:** Explains how complex arms and epoch were mapped to TS, TA, TE, TV.
- ✓ **Protocol Deviations:** Links specific study events to protocol deviations and their handling.
- ✓ **Validation Findings:** Provides clear, study-specific rationale for unresolved or expected validation findings.
- ✓ **Custom Domains:** Justifies the creation of non-standard domains or SUPPQUAL usage.

ADaM Analysis

Derived datasets designed directly to support statistical claims.



ADRG Key Contents

- ✓ **SAP Alignment:** Explicitly links analysis derivations to the final Statistical Analysis Plan.
- ✓ **ADSL Dependencies:** Documents the core subject-level variables that drive downstream ADaM.
- ✓ **Imputation Rules:** Fully explains complex derivations (e.g., LOCF, BOCF, multiple imputation).
- ✓ **Analysis Windowing:** Details exactly how visits were windowed (AWLO/AWHI) if deviating from standard.
- ✓ **Program Execution:** Provides the exact execution sequence to reproduce the primary efficacy TFLs.

The best guide is not the longest guide — it is the guide that removes the most uncertainty.

cSDRG & ADRG Pitfalls: Where Reviewers Face Friction

cSDRG (SDTM) Critical Faults

- ✗ **Vague Validation Explanations** Using generic excuses without detailing the clinical context behind the P21 violation.
- ✗ **Missing Trial Design Context** Failing to explain complex, protocol-specific study phases or arms in the Trial Design datasets.
- ✗ **Unexplained Data Exclusions** Failing to explain why certain collected data fields (e.g., from the eCRF) were not tabulated.
- ✗ **Poor Conformance Formatting** Forgetting to use **landscape orientation** for the "Data Conformance Summary", making long P21 error tables unreadable.
- ✗ **Unsynchronized Relock Dates** Failing to update the Database lock and relock information should be harmonised across the Reviewer Guide, TS domain and submission documentation.

ADRG (ADaM) Critical Faults

- ✗ **SAP Mismatch & Discrepancies** ADaM derivations (like baseline definitions) that silently alter the methods originally approved in the SAP.
- ✗ **Inconsistent Terminology** Failing to standardize terms across the document (e.g., mixing "subjects" with "**Participants**").
- ✗ **Code Dumping over Clinical Narrative** Copy-pasting complex programming code or arithmetic operators into the text instead of using simplified words to describe derivations.
- ✗ **Unresolved Draft Artifacts** Leaving messy internal team comments, draft notes, or vendor-specific jargon in the final submission package.
- ✗ **Redundant Basics & Implicit Issues** Over-explaining obvious, entry-level concepts while completely glossing over the actual, complex data issues.

Practical Case Studies



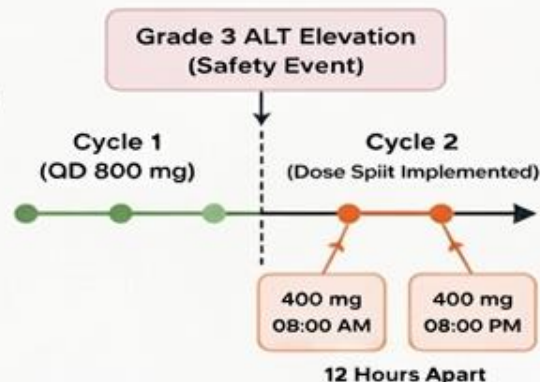
Case Study: Safety-Driven Dose Splitting and Review Readiness

1 The Scenario



A patient experiences a severe hepatic safety event mid-cycle, forcing the investigator to split a single daily 800 mg dose into two 400 mg doses taken 12 hours apart.

The dataset now shows two rows for the same day in EX. It is a violation because the system identifies duplicate active dosing dates for a protocol that strictly mandates once-daily dosing.



2 What the Data Shows (EX Domain)

USUBJID	EXSEQ	EXSTDTC	EXSTTM	EXDOSE	EXDOSU	DOSEFRQ
SUBJ-101	1	2025-05-15	08:00	400	MG	BID
SUBJ-101	2	2025-05-15	20:00	400	MG	BID



System flags a protocol deviation:

Duplicate active dosing dates detected (once-daily dosing mandated by protocol).

3 How to Explain it in the cSDRG



“ There are duplicate exposure dates for Subject X. This is intentional. Following a Grade 3 ALT elevation, the investigator initiated a split-dose safety mitigation strategy. The daily total dose remained compliant with the 800 mg protocol limit but is captured across two rows to accurately model exposure intervals and safety compliance. ”

4 How to Document it in define.xml

Under the Method Description for DOSEFRQ:

```
<ItemDef OID="DOSEFRQ" Name="DOSEFRQ" DataType="text">
  <Question>
    <TranslatedText xml:lang="en">Dose frequency per day</TranslatedText>
  </Question>
  <Definition>
    <TranslatedText xml:lang="en">
      If safety_split_flag = 'Y' then DOSEFRQ = 'BID';
      Else DOSEFRQ = 'QD';
    </TranslatedText>
  </Definition>
</ItemDef>
```



Key Takeaway:

Not all protocol deviations are data errors.
Clear explanation in cSDRG and proper metadata documentation help reviewers understand the clinical intent and maintain data integrity.



Outcome for Reviewers:

Transparent explanation + proper documentation =
Faster understanding, informed review, and better confidence in the data.

Case Study: Lab Sample Recollection and Review Readiness

From Hemolyzed Specimen to cSDRG Explanation and define.xml Documentation

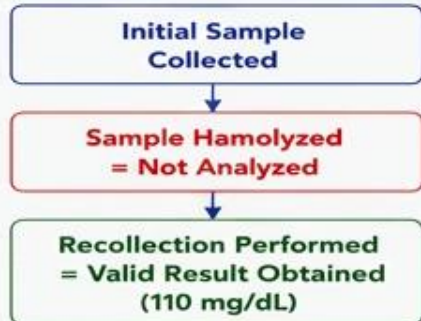
1 The Scenario



1. Initial laboratory sample collected for **SUBJID 1001**.
2. During processing, the specimen was identified as a **HEMOLYZED SAMPLE** and could not be analyzed.
3. A replacement specimen was collected on the same collection date to obtain a valid result.
4. The dataset therefore contains multiple LB records for the same subject and test date.
5. Automated validation checks may identify these records as potential duplicates, although they represent separate collection events.



Example: Same-Day Recollection



Same Date
(2025-05-15)

2 What the Data Shows (LB Domain)

SUBJID	LBSEQ	LBDC	LBTEST	LBSTRESN	LBSTRESU	LBREASND	LBSTAT
1001	1	2025-05-15	Glucose	NOT REPORTED	Null	HEMOLYZED SAMPLE	NOT DONE
1001	2	2025-05-15	Glucose	110	mg/dL	Null	Null



System flags a potential issue:

- Multiple laboratory records detected for SUBJID 1001.
- Same subject, same test, same date.
- Validation checks may identify potential duplicates.
- First record represents an unusable specimen, not a duplicate result.

3 How to Explain it in the cSDRG

“

Multiple laboratory records exist for SUBJID 1001 on 15-May-2025 due to sample recollection following identification of a hemolyzed specimen. The original sample could not be analysed (LBSTAT = "NOT NONE"; LBREASND = "HEMOLYZED SAMPLE") and was replaced by a subsequent collection. Data Management confirmed that these records represent separate collection events associated with the same scheduled visit and collection date and are not duplicate laboratory records. Therefore, these records should not be interpreted as duplicate observations.

”

4 How to Document it in define.xml



Method
Description
for LBSEQ

```
1 <ItemDef OID="LBSEQ" Name="LBSEQ" DataType="integer">
2   <Description>
3     <TranslatedText xml:lang="en">
4       LBSEQ uniquely identifies each laboratory
5       collection event for a subject. Multiple
6       records collected on the same date may
7       represent distinct collection events and
8       should not be interpreted as duplicate
9       records.
10    </TranslatedText>
11  </Description>
12 </ItemDef>
```

Show more lines

Additional Metadata

- SUBJID identifies the study participant.
- LBSEQ uniquely identifies individual laboratory collection events.
- Multiple records for the same subject and test date may occur when a specimen requires recollection.
- Data Management confirmed that records with different LBSEQ values represent distinct collection events and are not duplicate laboratory records.
- LBSTAT indicates the status of the laboratory sample.
- LBREASND documents the reason a sample could not be analyzed, such as HEMOLYZED SAMPLE.



Initial Sample
Collected



Hemolyzed Sample
(Not Analyzed)



Data Management
Confirmed
Not Duplicate



Replacement Sample
Collected



Explained
in cSDRG



Documented
in define.xml

Case Study: Documenting Protocol-Specific Safety Window Exceptions

From Protocol Exception to *define.xml* Implementation



1. SCENARIO

Protocol-Specific Safety Window Exception

- Protocol defines a standard 4-week (28-day) post-treatment safety window.
- Subjects who crossed over early to an optional extension study signed a new consent form before completing the full safety follow-up period.
- For these subjects, the safety window was truncated on the re-consent date.
- Automated validation tools flag the resulting differences in safety window duration as inconsistencies across subjects.



2. DATA CONFORMANCE

“

This is expected that variable AWTDIFF is inconsistent for 14 subjects. Per Protocol Amendment 3, subjects entering the long-term extension study early had their safety follow-up window truncated on the day of re-consent. The specific derivation logic for these subjects is detailed in define.xml under the AWTDIFF variable metadata.”

”



Explains why the validation finding is expected.



Links the exception directly to Protocol Amendment 3.



Provides traceability to define.xml metadata.



3. DEFINE.XML DOCUMENTATION

Method Description for AWTDIFF

```
If crossover_flag = 'Y' then  
    AWTDIFF = extension_consent_date -  
              trt_end_date;  
  
Else  
    AWTDIFF = 28;
```

Purpose

- ✓ Captures subject-specific safety-window derivation.
- ✓ Documents protocol-driven exceptions.
- ✓ Ensures transparency and auditability for reviewers.



Key Takeaway:

This approach ensures that protocol-specific exceptions are clearly explained in define.xml, supporting data conformance, transparency, and regulatory review.



Case Study: Documenting Imputed PK Values for AUC Calculations

From Missing Data Scenario to ADRG Explanation and define.xml Implementation

1



SCENARIO

Missing PK Sample Requiring Manual Imputation



A site forgot to collect blood for a scheduled PK time point, or the laboratory lost/spilled the sample.



The Statistical Analysis Plan (SAP) requires complete concentration data to support Area Under the Curve (AUC) calculations.



The missing value prevents standard derivation of AUC.

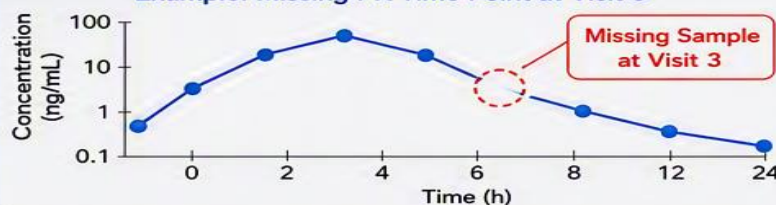


The programmer imputes the missing concentration using Last Observation Carried Forward (LOCF) or another approved imputation approach.



Automated validation tools may flag these records as "Unverifiable Derivations" because the value was not directly collected.

Example: Missing PK Time Point at Visit 3



2



ADRG DOCUMENTATION



“For ADPC, there exist records where concentration values (AVAL) were imputed. Due to documented sample loss, values were imputed using the LOCF method specified in the SAP to enable complete AUC calculations. Records are flagged using DTYPE = 'LOCF' for transparency and traceability.”

Reviewer Value



Explains why imputed values exist.



References the approved methodology documented in the SAP.



Provides transparency through the use of DTYPE = 'LOCF'.

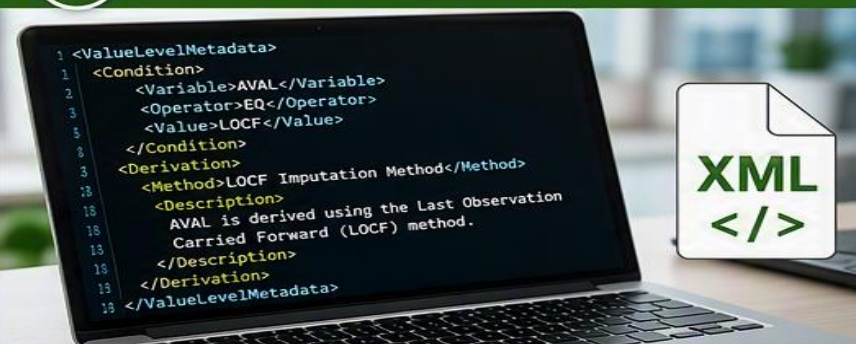


Supports traceability during regulatory review.

3



DEFINE.XML DOCUMENTATION



Value Level Metadata (VLM) for AVAL

Condition	Derivation Method	Method Description
DTYPE = 'LOCF'	LOCF Imputation Method	AVAL is derived using the Last Observation Carried Forward (LOCF) method.
All Other Values	Direct Laboratory Transfer	AVAL is obtained through direct electronic laboratory transfer.



Purpose

- Identifies when the derivation method changes.
- Distinguishes collected versus imputed values.
- Documents the exact imputation approach used.
- Improves transparency, reproducibility, and reviewer understanding.



Missing Sample at Visit 3



Imputed Using LOCF per SAP



Explained in ADRG



Documented in define.xml with VLM Condition (DTYPE = 'LOCF')

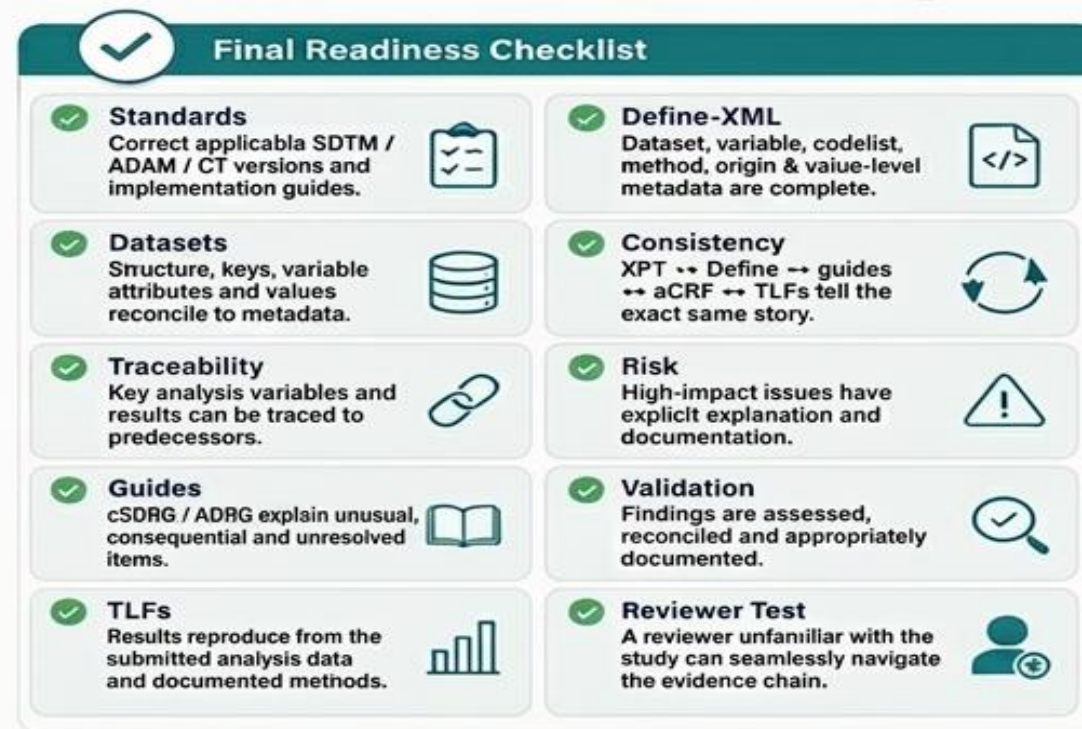
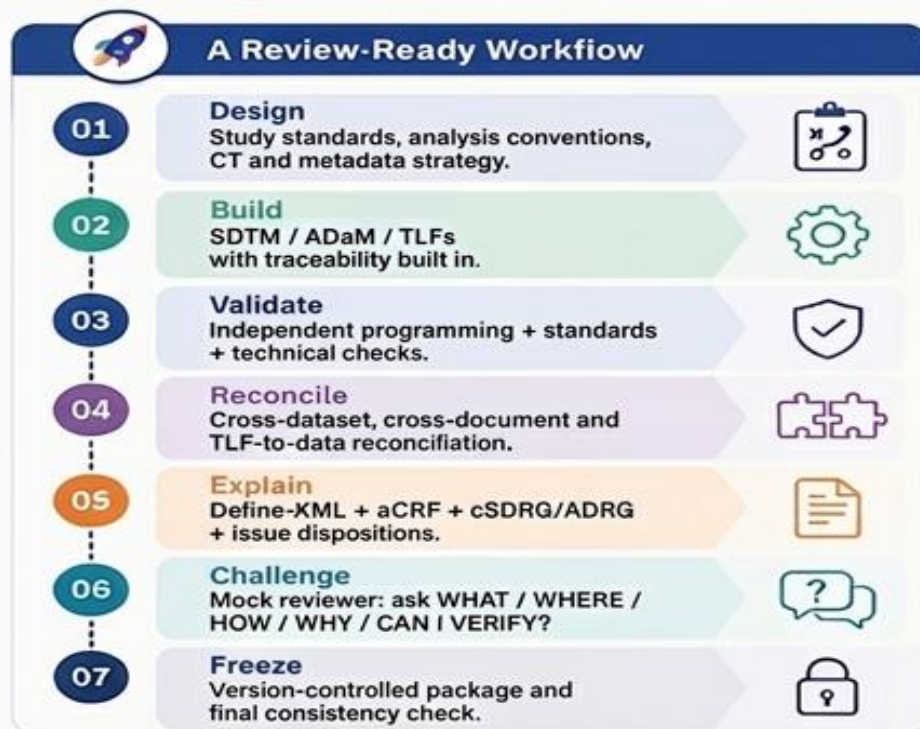


Summary and Closing Remarks



The Ultimate Review-Ready Framework

➤ Right Data. Right Story. Right Evidence. Right Outcomes. ➤





A review-ready submission is one where the reviewer can understand the data, follow the evidence chain, verify the derivations, and trust the conclusion.



References & Acknowledgement

FDA. Study Data Technical Conformance Guide — Technical Specifications Document, June 2026.

FDA. Study Data Standards Resources / Study Data Submission to CDER and CBER.

CDISC. Define-XML v2.1.11, published 6 April 2026.

CDISC. Define-XML v2.1 and Conformance Rules for Define-XML v2.1.

CDISC. ADaM Primer / Traceability resources and ADaM Examples of Traceability v1.0.

CDISC / NCI-EVS. Controlled Terminology; 27 March 2026 release and 2026 Package 62 public review.

CDISC. Standards in Development — 2026 status for SDTM / ADaM and related standards.

CDSIC. SDTM-MSG v2.0 and ADaM MSG v1.0

Acknowledgement:

Preeti Shailendra Ghatge

Krupali Prashantkumar Delvadia

Recommended practice: always verify the current FDA Data Standards Catalog and applicable agency guidance immediately before submission.





Thank You!





Questions?

