PharmaSUG China 2026 - Paper AI-174

Automating ADaM Specification with AI: Intelligent Generation, Risk Assessment, and Human-in-the-Loop Review

Shuang Gao, Yaohui Zhu, Tingting Zeng, Jinling Li, Wei Wang, BeOne Medicines

ABSTRACT

Authoring Analysis Data Model (ADaM) specifications is a knowledge-intensive activity in clinical statistical programming. A complete specification must align CDISC standards, study design, source data, statistical analysis plan requirements, protocol definitions, historical study practices, and internal programming conventions. Adapting existing materials to a new study still requires substantial manual effort, because many decisions depend on current-study evidence and expert interpretation.

This paper presents a controlled AI approach to ADaM specification authoring. The system generates structured metadata candidates for human review, drawing on company standards, historical study references, and current-study documents. A read-only Ask Agent supports day-to-day ADaM knowledge queries. AI is used to assist drafting and surface evidence; final decisions remain with expert reviewers. The paper describes what the system produces, how reviewers interact with it, how candidates are evaluated, and what governance controls are in place.

1. INTRODUCTION

ADaM specifications define the structure and derivation logic of analysis datasets used in clinical trial reporting and regulatory submission. They include dataset metadata, variable definitions, value-level metadata, source information, derivation methods, comments, programming notes, controlled terminology, and analysis parameter definitions. Because ADaM specifications sit between SDTM source data and downstream statistical outputs, errors or inconsistencies can propagate into programming, analysis, review, and submission deliverables.

In practice, ADaM specification development is repetitive but highly study-specific. Many datasets and variables follow established standards, yet the final derivation often depends on protocol design, SAP text, source data structure, compound conventions, analysis windows, and decisions made in similar studies. Programmers and data standards experts therefore spend significant time copying, adapting, checking, and reconciling specification content.

The manual effort is concentrated in three areas. First, adapting templates and historical specifications to a new study requires cross-referencing multiple documents: the SAP, the protocol, the SDTM metadata, and prior-study decisions. For each variable, a programmer may need to read across several sources to confirm that the assumed derivation is valid for the current study. Second, managing differences across studies adds overhead. When a dataset has been used in several prior studies with slightly different conventions, deciding which version to start from and what to carry forward requires judgment that templates alone cannot provide. Third, onboarding new team members is slow because ADaM knowledge is distributed across specifications, documents, and team experience rather than being centrally accessible.

ADaM teams also repeatedly answer operational knowledge questions: how a variable is defined in a given study, whether historical precedent exists, and how definitions differ across studies. These questions often require manual searches across multiple documents. The consultation workload is less visible than drafting, but it also consumes expert time and slows new-study startup and onboarding.

Traditional automation helps with template management and standard reuse, but it is less effective when tasks require interpretation across documents or adaptation of logic from one study context to another. Free-form generative AI can produce fluent text, but without grounding in study-specific evidence and without structured review, it is unsuitable for regulated metadata where traceability and expert accountability are required.

The central design principle of this work is that ADaM metadata automation requires a controlled AI approach rather than free-form autonomous generation. AI assists with drafting and evidence retrieval, while human experts remain responsible for review, interpretation, and final acceptance.

2. OBJECTIVES AND CONTRIBUTIONS

The system has two linked objectives. The first is authoring assistance: to reduce the routine effort of generating, adapting, and reviewing ADaM metadata candidates for new studies. The second is knowledge consultation: to let users ask natural-language questions about variable definitions, historical precedent, and cross-study comparisons, rather than manually searching documents.

Both objectives share the same principle: AI supports drafting and consultation while preserving traceability and expert control. The system reduces routine effort but does not bypass human accountability or update official specifications without review.

Specifically, the system is designed to:

- generate candidate variable-level and value-level ADaM metadata;
- reuse company standard templates and historical reference studies;
- adapt candidates to current-study context;
- identify fields likely to need expert attention;
- provide transparent comparison between AI candidates and official specifications;
- preserve human decision authority before official metadata is updated;
- support continuous improvement through structured difference analysis.

The goal is not to replace ADaM experts, but to shift their effort from routine drafting toward high-value review, study-specific interpretation, and quality decisions.

3. CONTROLLED AI APPROACH

The system is designed as a controlled AI assistant, not an autonomous author. AI helps generate and adapt candidate content based on available evidence, but it does not make final decisions and cannot update official specifications independently. All candidates require explicit human review and acceptance before they are promoted to official ADaM metadata.

This design fits the requirements of regulated clinical metadata work. Expert reviewers remain accountable for every specification decision. The system prepares reviewable candidates, surfaces relevant evidence, and reduces the routine effort of searching, copying, and adapting across documents and studies. The final call always belongs to the reviewer.

From a reviewer's perspective, the experience is different from starting from a blank specification. Instead of drafting from scratch, the reviewer begins with a set of candidates that already reflect company standards, relevant historical experience, and current-study context. Each candidate comes with attached evidence and a confidence signal. The reviewer's task is to evaluate, compare, and decide — not to research and draft simultaneously. This shift makes review more focused and surfaces differences more clearly than a manual drafting process would.

From an organizational perspective, every decision taken in the system is recorded and traceable. Which evidence was used to generate a candidate, which reviewer accepted or rejected it, and whether the candidate matched the official specification are all available for audit. This traceability supports both compliance requirements and systematic quality improvement.

Figure 1 illustrates how AI assistance, evidence sourcing, candidate generation, and human review are organized in the system.

Figure 1. Controlled AI System for ADaM Specification Authoring

Evidence-grounded generation . expert review . human authority over all official metadata

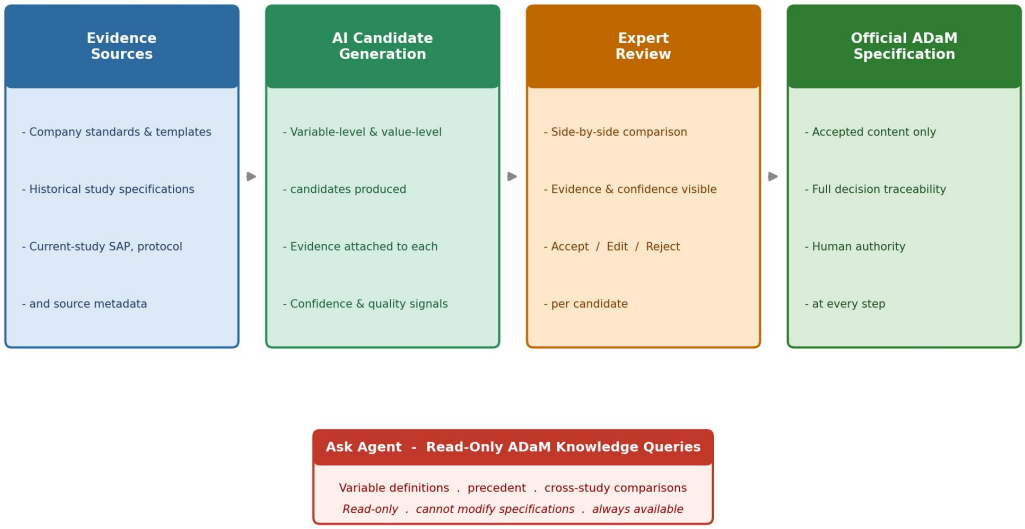

Figure 1. System overview: evidence sources, AI candidate generation, expert review, and official ADaM specification.

4. EVIDENCE SOURCES

The platform brings together the full set of information that an experienced ADaM programmer would consult when starting a new specification — company standards, historical study experience, and current-study documents — making it available at the point of candidate generation rather than requiring individual manual lookup. The result is that reviewers receive candidates already informed by established standards and prior practice, grounded in the current study's requirements, rather than starting from a generic template and then reading across multiple documents to determine what needs to change.

As reviewers work through candidates, their accept, edit, and reject decisions are recorded and feed back into the system. This creates a cycle where each completed study contributes to better starting points for future work, and where the accumulated knowledge of the team is captured rather than residing only in individual memory.

Table 1. Evidence source categories and their role in the system.

Evidence source	Role
Company standards and ADaM templates	Establish the controlled baseline for candidate generation.
Historical compound and study specifications	Provide implementation experience from prior similar work.
Current-study evidence (source metadata, SAP, protocol, supplementary documents)	Ground candidates in the requirements and context of the target study.
Reviewer feedback and official outcomes	Feed accepted and rejected decisions back into ongoing quality improvement.

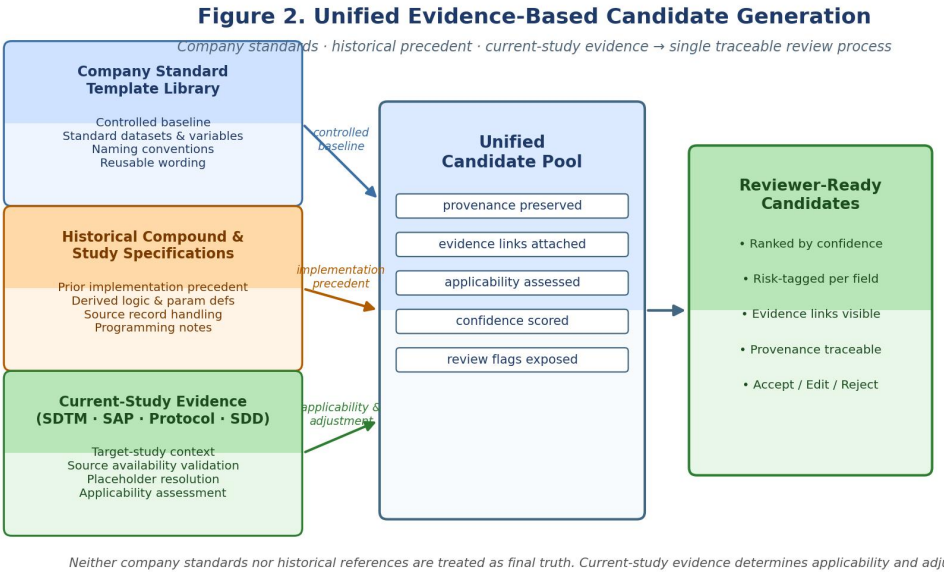

Figure 2. Evidence-based candidate generation: company standards and historical references as layered baseline; current-study evidence for grounding.

5. ADAM CANDIDATE AUTOMATION PIPELINE

The pipeline brings together company standards, historical study references, and current-study evidence to produce a reviewer-ready set of ADaM metadata candidates. Figure 3 illustrates the overall flow.

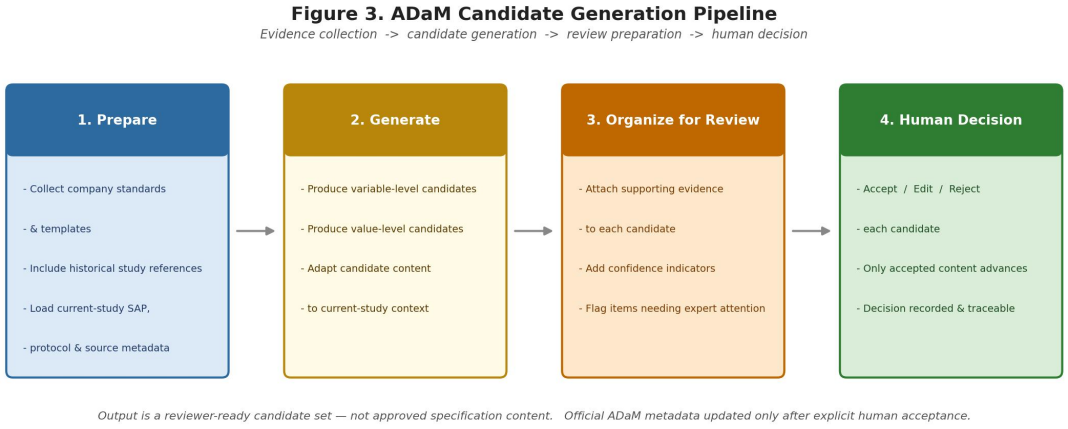

Figure 3. ADaM candidate generation pipeline.

A reviewer-ready candidate is not simply a copy of a template. It includes the key metadata fields for the variable or parameter, the evidence that supports the candidate, a confidence indicator, and any quality flags that indicate the reviewer should pay closer attention. This combination of content and context allows a reviewer to make an informed decision without having to search for supporting evidence separately.

Candidates that cannot be adequately grounded in available evidence are surfaced for manual attention rather than silently included in the output.

No candidate is treated as approved content. Expert reviewers retain final authority before any candidate can be promoted to official metadata. The system records the outcome of every reviewer action, which allows retrospective analysis of where candidates required modification and why.

6. ASK AGENT FOR ADAM KNOWLEDGE QUERYING

The Ask Agent extends the same controlled AI principle to day-to-day ADaM consultation. It is a read-only interface that lets users ask natural-language questions about variable definitions, historical precedent, and study comparisons. The system answers by retrieving and summarizing structured ADaM knowledge within a defined evidence boundary; it cannot modify specifications or author new content.

The Ask Agent complements the authoring pipeline by making reviewed specification knowledge easier to find and reuse. Questions that previously required manual lookup across specifications and study documents can be answered directly from within the review workflow. For example, a reviewer inspecting a derivation candidate may want to know how the same variable was handled in a prior study. Instead of leaving the review interface to search manually, the reviewer can ask the Ask Agent directly. The answer is drawn from the same structured knowledge base that powers candidate generation, ensuring consistency between authoring and consultation.

Over time, the knowledge available through the Ask Agent grows as more studies are reviewed and accepted. Teams that have accumulated reviewed specifications across multiple studies can query that knowledge directly rather than searching document by document. This makes the Ask Agent particularly valuable during study startup and onboarding, when new team members need to understand how prior work was done without having access to the original programmers.

7. FRONT-END OPERATING MODEL

The system is delivered through a web-based review interface organized into five views, each corresponding to a step in the review workflow.

The **Configuration** view lets reviewers define the study scope before generation begins. They select which datasets to generate, which baseline sources to use, and which historical reference studies to include. Reviewers can also specify supplementary evidence sources if the study has characteristics not well represented in the standard templates. This configuration step gives reviewers control over what the system considers as evidence, rather than having the system make those choices autonomously.

The **Monitor** view provides a live progress view during candidate generation. Reviewers can see which datasets have been processed, track overall completion, and identify whether any issues have occurred. For larger studies with many datasets and variables, the monitor view makes it practical to start reviewing available results while the remaining datasets are still being processed, without waiting for full completion.

The **Results** view presents the full candidate list for a study. Reviewers can browse all generated candidates, see a confidence summary for each, and sort or filter by dataset, variable, or review status. Candidates with quality flags or lower confidence are easier to identify and prioritize from this view. This gives reviewers an overview of the full scope of the generation before they dive into individual candidates, and allows them to triage their time effectively.

The **Detail** view is the main review surface. It shows a single variable or value-level candidate side by side with the existing official specification, field by field. Attached evidence — such as relevant SAP paragraphs, protocol excerpts, or source variable information — is accessible directly within the view. Confidence signals and quality flags are visible at the top, so the reviewer immediately knows the system's assessment before reading the content.

From the detail view, reviewers can take three explicit actions: accept the candidate as is, open it for editing and then accept the edited version, or reject it. These are deliberate choices, not defaults. The system records each action along with a timestamp, the reviewer's identity, and the final content that was accepted or rejected. This creates a complete audit trail for every variable and parameter in the specification. A reviewer returning to an earlier variable can see exactly what was decided and why.

The detail view also shows where the candidate differs from the existing official record. Differences are highlighted field by field, so reviewers do not need to read both columns in full to find what has changed. This is particularly useful for studies that build on prior work: the reviewer can focus attention on the fields where the AI candidate diverges rather than reading every field from scratch.

The **Dashboard** view summarizes review progress across the full study scope, showing how many candidates have been accepted, edited, rejected, or are still pending. It also surfaces residual risk by highlighting which datasets have high proportions of pending or flagged candidates, or where the modification rate has been consistently high. This view gives the team a clear picture of how much work remains and where attention is most needed before the specification can be promoted.

Ask Agent access is available as a contextual panel throughout the interface. When a reviewer is working in the detail view for a specific variable, the Ask Agent can be queried in context, and the answer is drawn from the same evidence base that informed the candidate. This means the reviewer does not need to switch tools or context to look up historical usage or check a prior study's convention.

Figure 4. Front-End Operating Model and Mapping to the Controlled Workflow

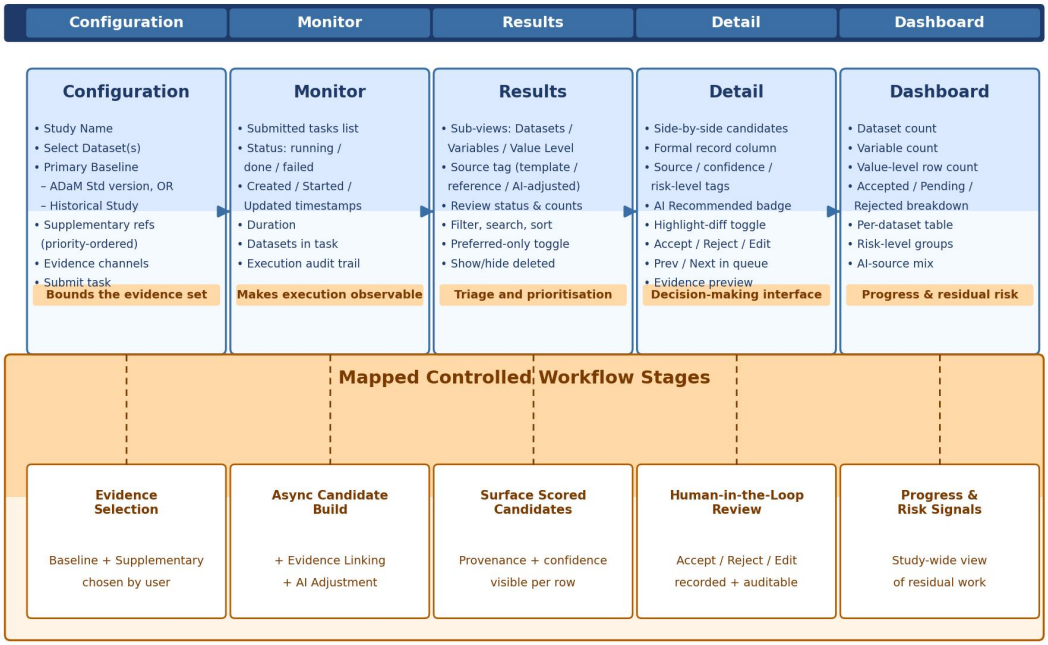

Figure 4. Front-end operating model.

Figure 6. Front-End Detail View — Reviewer UI Reference

Single AI candidate compared field-by-field against the formal record. Base template, SAP reference, source variable and confidence visible at all times.

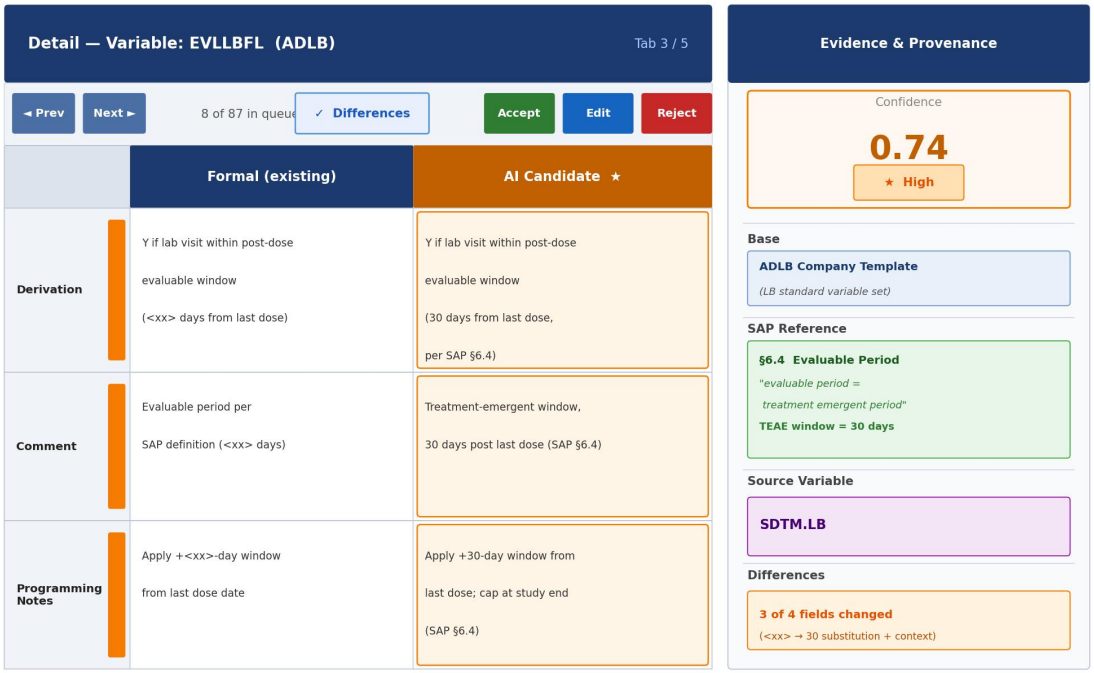

Figure 6. Reviewer UI reference: side-by-side comparison with evidence, confidence signals, and review actions.

8. PLATFORM OUTCOMES

The following outcomes reflect the practical effects of deploying the system across ADaM specification workflows.

Reduced routine drafting effort. ADaM teams spend significantly less time copying, adapting, and cross-referencing specifications for new studies. Reviewers start from evidence-grounded candidates rather than blank specifications, which means their effort shifts from searching and drafting to evaluating and deciding. The most routine variables — those with stable derivations well represented in historical studies — require the least reviewer time, freeing capacity for the variables where study-specific judgment is genuinely needed.

More consistent review across studies. Because all candidates are grounded in the same documented evidence base, specification decisions are more consistent across programmers and studies than when drafting from individual recall. Reviewers working on different datasets within the same study start from the same evidence, reducing the variation that arises when multiple people independently research the same references.

Faster study startup and onboarding. New team members can query the specification knowledge base directly using the Ask Agent rather than relying on colleagues or searching through archived documents. Questions that previously required senior programmer input — how a variable was handled in a prior study, whether a derivation has historical precedent — can be answered in seconds from within the review workflow. This reduces the time between study assignment and productive contribution.

Audit-ready documentation from day one. Every specification decision is recorded at the point of acceptance, with the evidence that informed it and the identity of the reviewer who made the call. No additional documentation steps are needed to produce a complete audit trail. When a discrepancy surfaces later, the review history is available for investigation without reconstructing what was decided or why.

Continuous improvement through accumulated knowledge. As more studies are reviewed and accepted, the knowledge available for future studies grows. The modification reason analysis provides structured feedback that helps teams understand where candidate quality is strong, where it needs improvement, and what type of improvement is needed. Over time, the system becomes more useful as the evidence base deepens and quality gaps are systematically addressed.

9. EVALUATION DESIGN

The evaluation compared AI-generated ADaM candidates with official specification content across multiple ADaM domains, covering both variable-level and value-level metadata. Reviewed and accepted AI candidates were compared at row level and at field level. Pending and rejected candidates were excluded from accuracy calculations, because these represent cases where the candidate was not accepted as a valid starting point — which is a different question from whether accepted candidates were accurate.

The primary metrics were: coverage (proportion of official rows with a matching reviewed AI candidate), cell-level accuracy (proportion of comparable fields that did not require modification), modified rate, and modification reason distribution. Coverage and accuracy are reported separately because they reflect different dimensions of candidate quality and point to different improvement paths.

The evaluation also distinguishes authoring quality from governance quality. For authoring, the key question is whether reviewers found the candidates to be useful starting points for decision-making. Governance metrics track whether the review and acceptance process was followed consistently and whether evidence traceability was maintained throughout. These are separate dimensions: strong candidate quality and strong governance compliance are both necessary, and each is measured independently.

No numeric claims are reported until validated against a defined evaluation dataset.

10. MODIFICATION REASON ANALYSIS AND QUALITY MONITORING

Field-level modifications between AI candidates and official specifications should not all be interpreted as AI errors. Differences may reflect missing study-specific context, official supplementary notes added during review, project conventions, or reviewer judgment calls that cannot be derived from documents alone.

The purpose of analyzing modification reasons is to give teams a clear picture of their quality profile rather than a single opaque accuracy number. Without this analysis, all differences look identical in an aggregate count — regardless of whether they reflect a substantive derivation gap, a stylistic preference, or a project team convention. Teams cannot tell where improvement effort is most needed or whether a high modification rate is a signal of quality concern or simply a reflection of normal human variation in expert review.

The framework distinguishes six categories of difference. Each category has a different implication for quality monitoring. Some indicate that the platform's knowledge coverage can be improved. Others show that the differences are largely a feature of the collaborative review process — reviewer note additions, team writing conventions, or official supplementary content — rather than failures in candidate quality. This breakdown allows teams to evaluate their quality profile honestly and to set meaningful improvement goals.

This categorization also changes how results are communicated. An organization that sees a notable share of accepted candidates modified might interpret this as low quality, but if the reason analysis shows that most changes reflect convention differences and reviewer additions rather than incorrect derivations, the interpretation changes significantly. The system is performing well; the remaining differences are evidence that human judgment is actively engaged rather than absent.

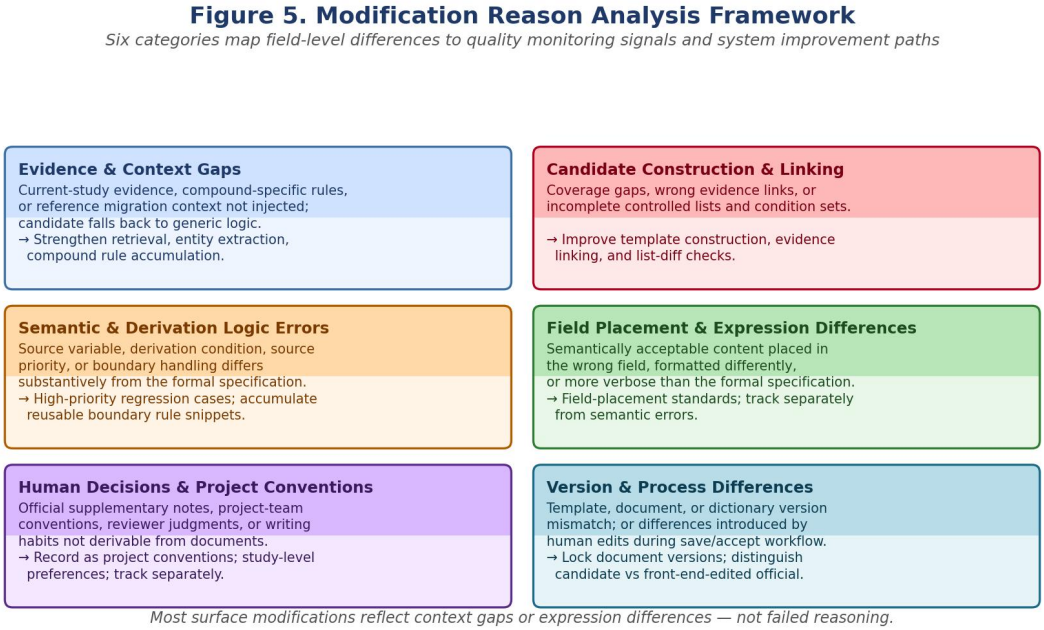

Figure 5. Modification reason analysis framework.

Table 5. Modification reason analysis framework: difference categories and their implications.

Category	Meaning	Implication
Evidence and Context Gaps	Study-specific rules or context was not sufficiently reflected in the candidate.	Indicates an opportunity to improve the coverage of study- and compound-specific knowledge used during generation.
Candidate Construction and Linking Issues	Candidate coverage is incomplete or evidence links are incorrect.	Indicates a generation coverage gap rather than an accuracy problem on included fields.
Semantic and Derivation Logic Errors	Derivation condition or boundary handling differs substantively from the official specification.	High-priority differences that affect clinical or statistical meaning; should be separated from other categories.
Field Placement and Expression Differences	Candidate content is semantically acceptable but placed in the wrong field or formatted differently.	Low-risk differences that do not affect meaning; should be tracked separately from semantic errors.
Human Decisions and Project Conventions	Differences reflect project-team conventions or reviewer judgments that cannot be derived from documents.	Not attributable to system performance; better managed through review guidance and convention documentation.
Version and Process Differences	Template, document, or dictionary version mismatches, or differences introduced during the accept workflow.	Evaluation artifact rather than a generation error; requires version control discipline rather than model improvement.

11. GOVERNANCE AND COMPLIANCE CONSIDERATIONS

Governance is built into the workflow rather than added afterward. AI outputs are candidate records, not final regulatory decisions. Official metadata updates require human approval. Reviewers remain accountable for interpretation, verification, and final acceptance.

Every candidate is traceable to its source: reviewers can see which evidence was used, what the confidence level is, and whether any review flags were raised. No candidate enters the official specification without an explicit accept action from a qualified reviewer. For any variable in the official specification, the record shows what was generated, what evidence was attached, what the reviewer decided, and when. This means that the complete decision trail is available for audit without requiring any additional documentation beyond what the system captures in normal operation.

This traceability makes the system suitable for regulated clinical programming environments where accountability for every metadata decision must be clear. The review process is not an informal check; it is an integral step in the metadata approval workflow, enforced by the system. Teams can demonstrate that every official specification entry was reviewed and accepted by a named reviewer with access to the supporting evidence.

The governance model also supports retrospective analysis. When a discrepancy is found in a submitted specification, the audit trail allows the team to trace back to the original candidate, the evidence that was attached, and the reviewer action that resulted in the official content. This makes issue investigation faster and more transparent than systems where the review history is recorded only informally.

12. DISCUSSION

The results of applying this system suggest that controlled AI is a practical approach to ADaM specification authoring. Candidates generated by the system provide useful starting points that reduce routine drafting, and the review interface makes it straightforward to identify which candidates need modification and why.

A key observation from the modification reason analysis is that not all field-level differences represent AI quality issues. A substantial share reflects study-specific conventions, project team preferences, and official notes added during review. Understanding this distinction matters for how the system is assessed and improved. When modification reasons are categorized rather than aggregated, teams can develop a precise quality profile — one that separates substantive derivation gaps from stylistic differences and human convention choices — and focus improvement effort accordingly.

The Ask Agent adds value by making the knowledge captured in reviewed specifications easier to access during daily work. Teams that have accumulated reviewed specifications across multiple studies can query that knowledge directly rather than searching document by document. This shifts consultation time from manual search to focused question-answering, which is particularly valuable during study startup and onboarding. Each reviewed study adds to the pool of queryable knowledge, so the system becomes more useful over time.

One broader implication of this design is that the value of AI in regulated metadata work is not primarily about generating novel content. It is about consistency, coverage, and accessibility of evidence. The most impactful contribution the system makes is surfacing what is already known — from standards, prior studies, and reviewed specifications — in a structured, comparable form at the point where a reviewer needs it. The AI component helps interpret and adapt that evidence; the structured workflow ensures the result is reviewed and traceable.

The front-end operating model plays a direct role in these outcomes. Because reviewers always see candidate provenance, evidence, confidence, and field-level differences, modifications are not silent. Each modification is a recorded decision that feeds back into the modification reason framework and informs ongoing quality improvement. This feedback loop is what makes the system continuously useful rather than static.

13. CONCLUSION

Controlled AI is well suited to ADaM specification automation when the system is designed around human review rather than autonomous generation. By bringing together company standards, historical study experience, and current-study evidence, generating structured candidates for expert review, and providing a read-only knowledge query interface, the system reduces routine drafting effort and makes ADaM knowledge easier to reuse.

The key lesson is that the quality of AI assistance in regulated clinical metadata work depends not only on the AI model, but on how well the system organizes evidence, prepares candidates for review, and keeps expert reviewers in control of final decisions. The review interface and the governance model are not secondary features; they are the foundation that makes AI assistance practical and auditable in a regulated environment.

Looking ahead, the system's value grows as the accumulated evidence base deepens and as quality feedback from reviewed studies is used to close remaining gaps. The modification reason framework provides the structure needed to turn routine review activity into systematic quality improvement — ensuring that each study completed contributes to better outcomes for the next.

RECOMMENDED READING

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems.

Amershi, S., et al. (2019). Guidelines for human-AI interaction. Proceedings of the CHI Conference on Human Factors in Computing Systems.

Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems.

CDISC. Analysis Data Model Implementation Guide.

CONTACT INFORMATION

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

Shuang Gao
shuang.gao@beonemed.com

Yaohui Zhu
yaohui.zhu@beonemed.com

Tingting Zeng
tingting.zeng@beonemed.com

Jinling Li
jinling.li@beonemed.com

Wei Wang
wei.wang@beonemed.com