Paper ET02

AI + Metadata: An Agentic Approach to Standards-Based Automation

Kevin Burges, Veramed, Twickenham, United Kingdom

ABSTRACT

Clinical trial biometrics demand strict adherence to regulatory principles of Accuracy, Reproducibility, and Traceability (ART). However, the probabilistic nature of Large Language Models (LLMs) poses a challenge to integrating AI into GxP-compliant workflows.

This paper presents a practical framework for introducing AI into clinical trial biometrics, addressing the real-world challenges of policy, regulation, trust, and technology. The framework combines a risk-based governance approach with technical strategies including metadata-driven AI, Model Context Protocol (MCP), and deterministic-first principles with human-verified probabilistic fallback.

The discussion addresses the role of emerging CDISC standards - USDM, Biomedical Concepts, and Dataset Specialisations - in establishing the semantic framework necessary for reliable, autonomous agentic AI processes.

INTRODUCTION

Mid-2025 marked a pivotal moment in the AI landscape. Headlines proclaimed that agentic AI would transform enterprise workflows, that AI agents could autonomously complete complex multi-step tasks, and that the future of work was human-AI collaboration. It was in this environment of excitement and possibility that we began our journey toward AI-enabled clinical trial biometrics.

The vision was compelling: end-to-end automation from Protocol to Submission. Imagine generating machine-readable protocols from PDF or Word documents, automatically producing draft CRFs aligned to CDASH standards, generating SDTM metadata and mappings, creating transformation code, and producing ADaM datasets and TLFs - all orchestrated by intelligent agents working within a compliant framework.

This paper tells the honest story of that journey, from ambitious aspirations to practical reality. We acknowledge that introducing AI into clinical trial biometrics is complex, involving policy challenges, regulatory constraints, trust gaps, and technical hurdles. Rather than presenting finished solutions, we share a framework and early learnings that others can adapt for their own journeys.

THE REALITY OF AI IN REGULATED ENVIRONMENTS

THE TRUST CHALLENGE

A survey of 233 employees at our organisation revealed telling insights about AI readiness in clinical research settings. While 73% see opportunities for AI to free up time from repetitive tasks, and 72% express excitement or selective optimism about AI, 53% do not fully trust AI outputs for regulated activities.

The concerns are rational. Clinical trials operate under strict regulatory oversight where accuracy, reproducibility, and traceability are non-negotiable. The probabilistic nature of LLMs - where the same question might yield different answers - directly challenges these principles.

An encouraging insight from the survey is that 39% of respondents identify as an “early adopter” persona. This demonstrates a strong appetite for innovation within the organisation, suggesting a robust foundation for driving change and embracing new solutions.

AI STRENGTHS VS. GXP EXPECTATIONS

The tension between AI capabilities and GxP requirements manifests across several dimensions. AI’s probabilistic reasoning enables powerful inference and pattern detection, but GxP processes expect deterministic, validated behaviour with full traceability of how results were produced. AI systems that learn and improve over time conflict with the requirement for controlled, versioned systems where behaviour changes are documented and validated. AI’s ability to reduce human workload raises questions about accountability - GxP mandates qualified oversight and clear responsibility for decisions.

These tensions are real but not insurmountable. Recent regulatory guidance provides a path forward.

THE EVOLVING REGULATORY LANDSCAPE

Recent and upcoming regulatory guidance supports thoughtful AI adoption while maintaining patient safety and data integrity.

GAMP 5 SECOND EDITION

The GAMP 5 Second Edition (2022) describes a risk-based approach to validation that moves away from “one size fits all” thinking. It emphasises critical thinking by knowledgeable and experienced subject matter experts to determine appropriate validation approaches. Appendix M12 explicitly covers Critical Thinking, while Appendix D11 addresses AI/ML considerations. The core message is that validation effort should be scaled based on a deep understanding of how the system impacts the process and patient safety, not based on blanket requirements tied to software categories.

EU GMP ANNEX 22 (DRAFT)

The upcoming EU GMP Annex 22 makes an important distinction for AI applications in Good Manufacturing Practice (GMP) environments. For critical GMP applications (those with direct impact on patient safety, product quality, or data integrity) non-deterministic AI (including Generative AI and self-learning systems) must not be used. However, for non-critical GMP applications, non-deterministic AI can be used provided that personnel with adequate qualification and training are responsible for ensuring outputs are suitable for intended use - in other words, human-in-the-loop (HITL) oversight.

This distinction is crucial: context determines appropriateness. The same AI capability might be acceptable in one use case and not in another. This supports a risk-based, context-aware governance framework.

While these requirements are designed specifically for GMP settings, the underlying principles (namely the distinction between critical and non-critical applications, and the importance of qualified human oversight) are equally relevant to Good Clinical Practice (GCP) environments. In GCP, where patient safety and data integrity are paramount, a risk-based approach to AI adoption, with HITL controls for non-critical tasks, can help support robust compliance and responsible innovation.

A RISK-BASED GOVERNANCE FRAMEWORK FOR AI TOOLS

Our proposed governance framework employs a structured approach to risk assessment, determining appropriate controls for each AI use case. The framework applies equally to internally developed AI tools and third-party solutions such as Claude Code, with human oversight embedded at all decision points. It implements a multi-layer process:

Use Case → Usage Profile → Risks → Controls → Risk Zone Assignment

A partial example is shown below.

USE CASE

“Claude Code for developing assistive tools”

USAGE PROFILE

Autonomy: “Agentic”, Platform: “Anthropic-hosted Team plan”, Models: Claude (Opus, Sonnet, Haiku), Action: “Create/Modify non-GxP content”, Data: “Company Confidential”, Users: “Technical Solutions Group”

RISKS

R-0181: “Agent workflow orchestration governance”, R-0277: “AI systems making consequential decisions without human input”

CONTROLS

OPS-027: “Prompt and Context Quality Assurance” , OPS-023: “Human Validation Before Downstream Use”

RISK ZONE

Amber: Restricted use

USAGE PROFILE

The usage profile consists of several factors that relate to the potential risk, such as:

LEVEL OF AUTONOMY

AI applications can be classified into three tiers based on their level of autonomy.

Tier	Description	Examples
T1: Assistive	Human-initiated, single-task chatbot interactions. Output is suggestion only.	Prompt Optimiser, CDASH Expert agent, Research Assistant
T2: Automated	Defined workflows with AI components. Human oversight at checkpoints.	Low-code workflow automation, document processing pipelines

T3: Agentic	Autonomous goal pursuit with tool access. Minimal human intervention during execution.	Coding agents, Complex multi-step analysis

TYPE OF ACTION

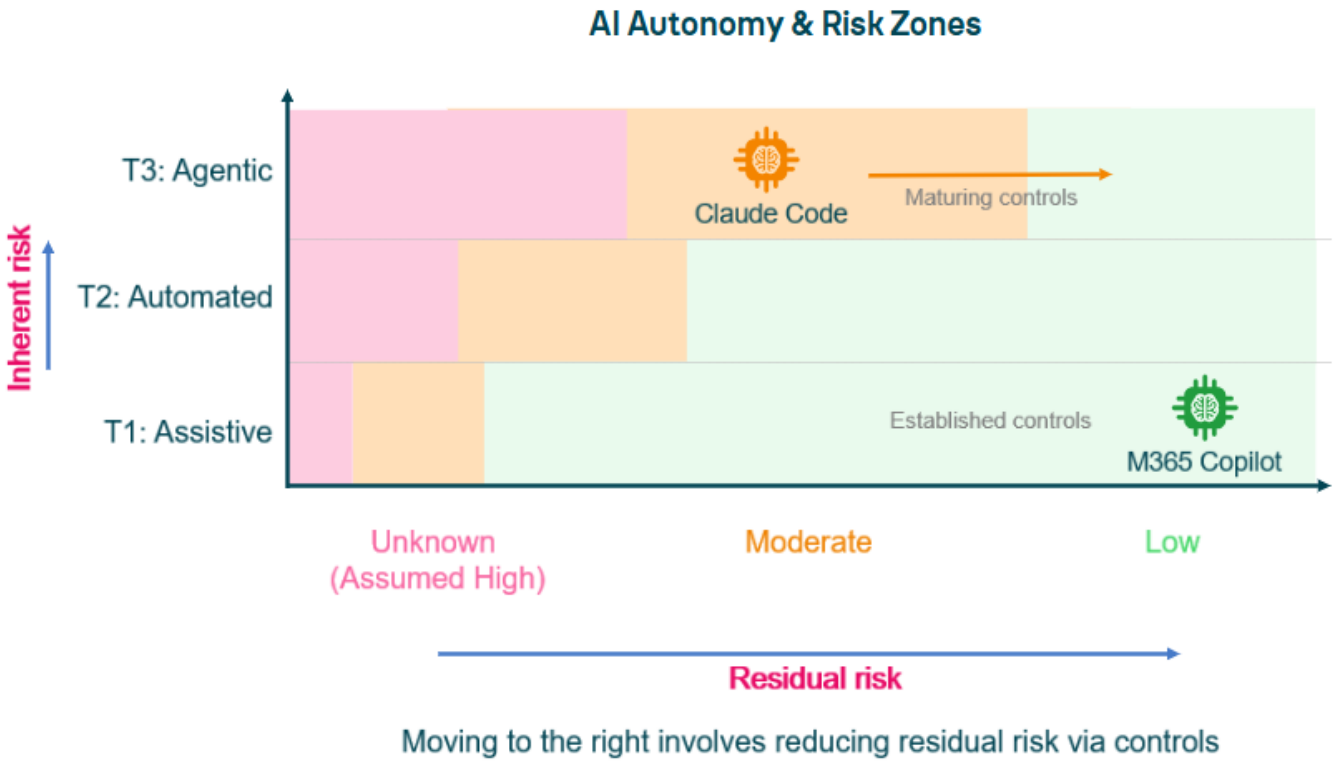
AI applications can be classified based on their type of action, e.g. “Read/Report”, “Create/Modify non-GxP content”, “Create/Modify GxP content”.

TYPE OF DATA

AI applications can be classified based on the type of data they act on, e.g. “Public”, “Company confidential”, “Client metadata”, “Client data”.

RISK ZONE

Each AI use case is assigned to a risk zone based on the “residual risk” remaining after the identified controls have been put in place. This assignment is performed through a combination of automated assessment and Governance Committee determination.



The risk zone determines what additional mandated controls must be in place to mitigate the residual risk.

The Red zone is the default entry state for unanalysed use cases, where residual risk remains unknown. Mandated controls such as sandboxed environments apply until formal assessment is completed.

The Amber zone indicates that risk analysis has been conducted and residual risk reduced through compensating controls, though not yet to a threshold permitting widespread deployment. Use cases may remain in Amber sustainably where technology immaturity or control limitations persist. Mandated controls such as regular review by the Governance Committee and enhanced monitoring apply.

The Green zone permits general organisational use, reflecting residual risk reduced to an acceptable minimum through established, typically automated controls.

Context determines risk - the same underlying tool may be Green zone in one use case and Amber zone in another.

Zone boundaries are defined by aggregate residual risk scores calculated for each use case following the application of relevant controls. The Governance Committee establishes threshold values that delineate zone transitions, providing objective criteria against which use cases are assessed.

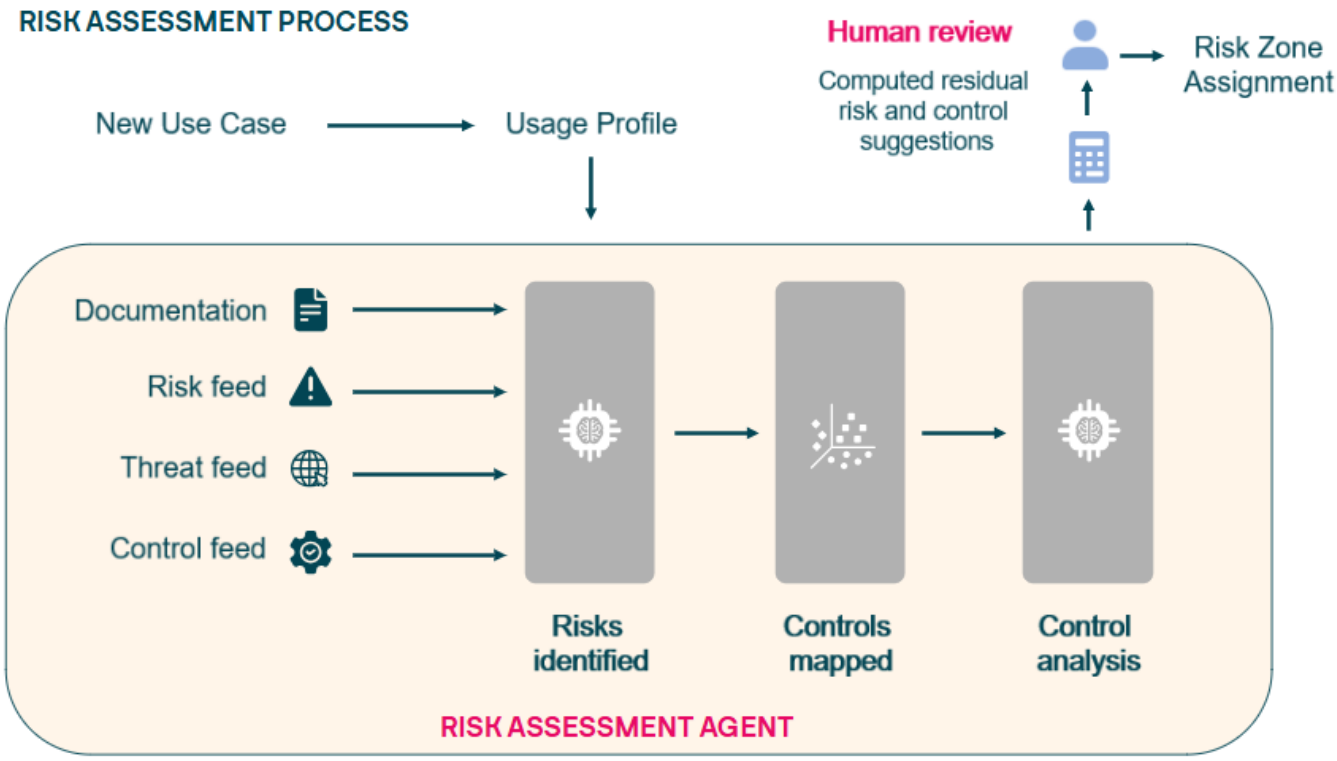
The varying sizes of the Red, Amber, and Green zones for each Tier of Autonomy indicate that the likelihood of a tool falling into a given zone depends on its Tier of Autonomy. For instance, most Tier 1 tools are likely to be found in the Green zone, while only a small number of Tier 3 tools will fall within the Green zone.

This framework remains a work in progress, with zone boundaries and control requirements being refined as practical deployment experience increases.

AI-ASSISTED RISK ANALYSIS

The rapid proliferation of AI tools, combined with an evolving threat landscape, presents a significant challenge for traditional risk management approaches. New vulnerabilities and attack vectors emerge at a pace that exceeds the capacity of manual assessment processes alone. Simultaneously, the volume of use cases requiring evaluation scales with organisational AI adoption.

To address these challenges, our framework incorporates an AI-powered risk analysis tool that augments human capacity to identify, assess, and monitor risks. The tool synthesises information from multiple sources to produce structured risk analyses, mapping identified risks to appropriate controls. This enables more comprehensive and consistent assessment than would be achievable through manual methods at equivalent scale.



Critically, the tool operates in an advisory capacity only. Human risk managers retain full authority to override, modify, or reject the tool's recommendations, and the Governance Committee makes all final zone determinations. This preserves accountability and ensures that contextual judgement, particularly regarding novel or ambiguous risk scenarios, remains with qualified professionals.

RELIABILITY: METADATA AND STANDARDS ARE CRITICAL

AI WITHOUT METADATA IS GUESSING

When a Large Language Model (LLM) is without access to authoritative sources, it may not always give the correct, or the same, response.

Consider a simple question: “What domains are in the SDTM 3.4 Interventions class?”

There is a definitive answer in the SDTM Implementation Guide: AG, CM, EC, EX, ML, PR, SU.

However, when this question was posed twice to an AI assistant (OpenAI, 2026), inconsistent responses were provided: one response listed “CM, EC, EX, PR, SU” (missing AG and ML), while another listed “CM, EX, ML, PR, SG, SU” (missing AG and EC, and incorrectly including SG). The AI may not have access to the required information or may not know how to interpret it correctly.

This illustrates a fundamental point: AI without reliable metadata is guessing. AI with structured access to authoritative standards can provide traceable, justified answers.

LAYERS OF STANDARDS

Standards in clinical data management exist in multiple layers, each adding value.

Layer	What It Provides	Examples
Format Standards	Common way to express content; interoperability between tools and organisations	CDISC ODM, Define-XML, Dataset-XML, Dataset-JSON
Content Standards	What to collect and how to structure it; industry consistency	CDASH, SDTM, ADaM, NCI Controlled Terminology
Semantic Standards	Meaning and relationships; enables genuine AI reasoning about studies	USDM, Biomedical Concepts, Dataset Specialisations
Organisational Standards	Organisation-specific implementations; consistency and data quality within organisation	Organisational CRF standards, Organisational SDTM/ADaM standards

WHERE WE ARE TODAY

The standards most adopted across the industry are **format standards** and **content standards**. The wide adoption is in no small part due to regulatory requirements, which mandate the use of standards such as SDTM, Define-XML, and NCI Controlled Terminology.

Increasingly, organisations are also implementing **organisational standards** to enhance data quality, improve consistency, and accelerate processes within their own environments. However, adoption of organisational standards is not ubiquitous. This is often because effective standards management demands dedicated expertise and resources, which can be difficult for organisations to justify or prioritise - especially smaller ones where such investments may not be feasible or are seen as lower priority compared to immediate operational needs.

THE SEMANTIC FUTURE

The real transformational shift, though, is the introduction of **semantic standards**. Emerging CDISC standards are unlocking the potential for end-to-end automation by adding a crucial layer of meaning to study data.

The **Unified Study Definitions Model** (USDM) provides machine-readable protocol definitions, encompassing areas such as study structure, objectives, endpoints, activities, and eligibility criteria.

USDM references **Biomedical Concepts** (BCs), which provide semantic definitions of what is being measured, such as “blood pressure” with all its attributes and relationships.

Dataset Specialisations provide an additional layer that operationalises BCs, guiding how concepts map to CRFs and SDTM for specific use cases.

By introducing semantic standards, we move beyond simply capturing and formatting data. These standards create a machine-understandable backbone for the entire study lifecycle, allowing AI systems to genuinely understand the meaning and relationships within a study, not just the surface structure of the data. As a result, AI can make intelligent, context-aware decisions, reliably generate CRFs, SDTM mappings and transformation code, and support advanced reasoning about studies.

This marks a significant step forward for clinical research automation and intelligence. While industry adoption is still in the early stages, organisations investing in semantic standards today will be well-positioned to benefit from genuine AI-driven automation and decision-making in the future.

THE TECHNICAL APPROACH

DETERMINISTIC-FIRST PRINCIPLE

A core technical principle is: deterministic-first, with human-verified probabilistic fallback. Consider the example of SDTM classification of CRF fields. First, we apply deterministic rules derived from standards metadata - CDASH and SDTM-IG classification rules encoded in Python. If a high confidence match is found, we return the classification, confidence level, and rule-based justification. If no high confidence deterministic match exists, we fall back to probabilistic LLM inference with documented reasoning. Either way, a human reviews and approves, especially for lower-confidence classifications.

This approach ensures that every classification is backed by justification and confidence level. Humans know where to focus their review. The audit trail shows exactly how each decision was made. And critically, the human retains responsibility.

MODEL CONTEXT PROTOCOL (MCP)

Model Context Protocol (MCP) is a standard way to connect LLMs to external data sources and tools. Instead of the LLM "remembering" or "guessing" information, MCP lets it query authoritative sources directly and deterministically.

For example, consider a user asking, "What domains are in the SDTM 3.4 Interventions class?". This sets in motion the following chain:

- User asks LLM: "What domains are in the SDTM 3.4 Interventions class?"
- LLM recognises it needs authoritative metadata → identifies the CDISC Library MCP tool
- LLM calls the MCP tool → MCP queries CDISC Library API directly
- MCP returns the authoritative answer → AG, CM, EC, EX, ML, PR, SU
- LLM responds to user with correct answer and source citation

This transforms unreliable inference into deterministic retrieval with full traceability.

LLM SKILLS

Skills represent a modular pattern for extending LLMs[1]. These transform the general-purpose LLM into a domain specialist. Each Skill is a directory containing instructions, reference data, and scripts that can include both **deterministic logic** (Python) and **probabilistic inference**. Skills **load context on demand** rather than all at once, helping the model stay focused while reducing costs.

A proposed Skill hierarchy for SDTM Classification illustrates the approach:

- **SDTM Classification Orchestrator** skill
  Orchestrates the high level process flow for SDTM Classification.
- **CRF Parser** skill
  Reads CRF metadata (ODM, Excel) into common format using **deterministic** Python code.
- **GOC Classifier** skill
  Classifies into a General Observation Class using rules implemented in **deterministic** Python code.
  Falls back to classifying using **probabilistic inference** based on metadata-derived rules.
  Loads the minimum context required to classify into a GOC.
- **<GOC-type> Classifier** skills:
  Classifies into a specific Domain using rules implemented in **deterministic** Python code.
  Falls back to classifying using **probabilistic inference** based on metadata-driven rules.
  Loads the minimum context required to classify the Domain in the relevant GOC.

Each step minimises token usage while maintaining full justification and confidence levels for human review.

The Skills use structured metadata to provide a known, consistent pathway from input to output.

- **CRF Metadata**
  The CRF Parser skill converts multiple possible input formats into a single metadata format that is passed to the classifiers.

- **Classification Metadata**
  Classifications are stored as structured metadata, with confidence levels and justifications for human review.

EXAMPLE SKILL STRUCTURE

Proposed filesystem structure for SDTM classifier skills:

```
cs_sdtm-classifier/
├── skills/                    # Claude Skills (SKILL.md files)
│   ├── orchestrator/          # Orchestration of deterministic and probabilistic classification
│   ├── goc-classifier/        # Classify into a GOC
│   └── {findings|events|interventions|special-purpose}-classifier/   # Classify into a Domain
├── src/                       # Python modules (deterministic)
│   ├── parsers/               # ODM, Define-XML, Excel
│   ├── classifiers/           # GOC, domain
│   ├── confidence/            # Scoring, thresholds
│   └── utils/
├── reference/                 # Reference metadata (plugin-like)
│   ├── config/thresholds.json
│   ├── goc-definitions/{goc}/[metadata.json, patterns.json]
│   └── domain-definitions/{sdtm-ig|custom}/{goc}/{DOMAIN}/[metadata.json, patterns.json]
├── tests/                     # Unit, integration, benchmarking
└── docs/                      # Documentation
```

- **Skills**
  Including orchestration of deterministic and probabilistic classification.

- **Python code**
  Including deterministic classification, parsers, confidence thresholds.

- **Reference metadata**
  Including standards and derived rules. Plugin-like architecture allows new standards to be incorporated.

EXAMPLE CLASSIFICATION METADATA

Proposed JSON structure of a classification result:


```
{
  "classifications": [
    {
      "id": "550e8400-e29b-41d4-a716-446655440000",
      "source": {
        "id": "VS.FORM",
        "type": "odm",
        "name": "Vital Signs"
      },
      "goc": {
        "name": "Findings",
        "confidence": 0.95,
        "method": "llm-reasoning",
        "reasoning": "Form contains test-result-unit structure...",
        "rules_matched": [],
        "contra_indicators": []
      },
      "domain": {
        "name": "VS",
        "confidence": 0.98,
        "method": "rules",
        "reasoning": "Variables VSTESTCD, VSORRES, VSORRESU...",
        "rules_matched": [],
        "contra_indicators": []
      },
      "domain_alternatives": [],
      "needs_review": false,
      "metadata": {
        "token_count": 1250,
        "processing_time_ms": 145,
        "reference_files_loaded": [...]
      }
    }
  ]
}
```


- **Source identification**
  Identifies the metadata that has been classified.
- **GOC / Domain classification**
  Provides confidence, classification method, prose explanation, rules matched, contra-indicators.
- **Alternatives**
  A list of alternative classifications to consider.
- **Metadata**
  Debug information to help with optimisation.

This structure provides documented evidence of how the classification was made and provides humans with what they need to guide their review. A "`needs_review`" flag is provided to explicitly state if review is required, based on a configured confidence level.

BENEFITS OF THE SDTM CLASSIFIER SKILLS

- Deterministic where possible, probabilistic where necessary
- Can handle novel CRF content (probabilistic fallback)

- Full justification for every classification - auditable
- Progressive context loading minimises token usage and cost

LEARNINGS

Governance first, technology second. Deploying AI tools without a clear governance framework leads to either paralysis ("we can't approve this") or shadow IT ("I'll just use my personal account"). The framework enables, not constrains.

People are more ready than you think, but need clarity. In our survey, 73% of staff see opportunities for AI. The major barrier is unclear guidance (21%), not resistance. Give people guardrails and they'll run.

Start simple. Build confidence with low-risk use cases (Tier 1/Tier 2, non-client data) before tackling GxP-critical applications. Quick wins create momentum.

Approval and permissions can be harder than the technology. Getting tools approved and deployed in an enterprise environment can take longer than building the solution.

People are probabilistic too. Ask a person to classify something multiple times - they won't always do it the same way. Ask multiple people to classify something - they'll give different answers. Yet we manage this through review, documentation, and quality processes. Using probabilistic AI doesn't inherently add new risk - it can be managed within existing quality frameworks, often with better documentation than human-only processes provide.

CONCLUSION

AI in GxP environments is possible. Regulatory guidance supports risk-based approaches that can accommodate AI appropriately, distinguishing between critical and non-critical applications and recognising the value of human-in-the-loop oversight.

Metadata and standards are the key. They remove the guesswork and provide the semantic foundation that transforms AI from unreliable inference to governed, traceable automation. The emerging CDISC standards - USDM, Biomedical Concepts, and Dataset Specialisations - unlock the future of standards-based automation.

The deterministic-first approach with human-verified probabilistic fallback provides a practical pattern for compliance. It combines the power of AI to handle novel situations with the traceability and accountability that regulated environments require.

The journey from ambitious agentic AI vision to practical compliant implementation is ongoing. The abstract written during the 2025 hype cycle was ambitious - and that's okay. The vision is right; the timeline was optimistic. Being honest about where we are builds more trust than overpromising.

For organisations considering this journey, we encourage starting now. Build the governance framework. Invest in metadata and standards. Start with low-risk use cases and build toward the agentic, semantic future. The opportunity to transform clinical trial biometrics through AI-enabled, standards-based automation is real, and organisations building toward it now will be well-positioned to lead.

REFERENCES

Good Automated Manufacturing Practice (GAMP) 5, Second Edition. A Risk-Based Approach to Compliant GxP Computerized Systems. International Society for Pharmaceutical Engineering (ISPE), 2022.

European Commission. (2025). Draft – EU GMP Annex 22: Artificial Intelligence [Consultation guideline]. Retrieved from https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en

CDISC. Unified Study Definitions Model (USDM). https://www.cdisc.org/ddf

CDISC. Biomedical Concepts & Dataset Specialisations. https://www.cdisc.org/cdisc-biomedical-concepts

Anthropic. Model Context Protocol (MCP). https://modelcontextprotocol.io/

Anthropic: What are Skills? https://support.claude.com/en/articles/12512176-what-are-skills

OpenAI. (2026). M365 Copilot Chat (GPT-5.2 Auto, January 2026) [Large Language Model].

ACKNOWLEDGEMENTS

The author acknowledges the use of AI assistants - Claude (Anthropic) and Microsoft 365 Copilot (Microsoft) - for assistance with drafting and refining the manuscript text. All content was reviewed and edited by the author.

CONTACT INFORMATION

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

Author Name: Kevin Burges

Company: Veramed

Email: kevin.burges@veramed.com

[1] Skills are currently specific to Anthopic's Claude platform, however the underlying pattern of on-demand context loading, deterministic code execution, and structured instructions could inform similar approaches in other LLM platforms.