Paper ML34

Transparent Models or Black Box? Using Generative AI to Generate Real-World Intelligence

Sherrine Eid, SAS Institute, Inc, Cary, NC, USA
Samiul Haque, SAS Institute, Inc, Cary, NC, USA

ABSTRACT

The rapid adoption of large language models (LLMs) and machine learning (ML) in pharmaceutical research presents opportunities for accelerating clinical development, pharmacovigilance, and regulatory submissions, while underscoring the need for robust model governance. Using synthetic patient data (e.g., Synthea), descriptive summary statistics and automated analyses can be generated across Python, R, SAS, and Lua environments to ensure reproducibility and transparency. Governance frameworks must emphasize model documentation (e.g., model cards), data lineage, explainability, and validation under Good Machine Learning Practices (GMLP) and GxP requirements. In current compute environments, containerized MLOps pipelines with embedded audit trails, risk-based monitoring, and human-in-the-loop oversight enable regulatory-grade deployment. Practical approaches include bias testing, drift detection, and controlled retraining aligned with ICH quality guidelines. By integrating model transparency with risk management, pharma organizations can responsibly leverage AI while providing regulators with the traceability, reproducibility, and interpretability required for trust and approval

INTRODUCTION

Patient outcomes research (POR) is increasingly expected to deliver decision-grade, transparent, and reproducible evidence across clinical development, regulatory review, health technology assessment (HTA), and post-market evaluation. Real-world data (RWD) -including electronic health records, claims, registries, and other routinely collected sources—has become central to this mission because it can represent broader populations, longer follow-up, and care patterns not fully captured in randomized trials. Yet regulators have been explicit that the value of RWD depends on demonstrable relevance and reliability, including fit-for-purpose data provenance, appropriate study design, prespecified analyses, and audit-ready documentation. The U.S. Food and Drug Administration (FDA) has articulated these expectations in guidance on using RWD/RWE to support regulatory decision-making for drugs and biologics and has separately published submission-focused guidance on how sponsors should identify and package RWD/RWE components within IND/NDA/BLA submissions. In Europe, the European Medicines Agency (EMA) has operationalized an RWE framework through regulator-led studies and the DARWIN EU program, underscoring both the opportunity and the governance needed for RWD to support EU regulatory decisions.

LARGE LANGUAGE MODELS

Against this backdrop, generative AI—and in particular large language models (LLMs)—is rapidly entering POR workflows. LLMs can accelerate tasks that are traditionally labor-intensive and error-prone: (1) phenotype and outcome definition from clinical narratives via natural language processing; (2) protocol and feasibility intelligence, including eligibility logic checks and site/patient availability signals; (3) literature surveillance and synthesis, supporting faster evidence landscaping; and (4) analytic scaffolding, such as code templates, table shells, and documentation drafts. These capabilities can shorten cycle time from question to insight and improve consistency in early iterations of study planning—especially when paired with retrieval-augmented generation (RAG), where the LLM is constrained to a curated evidence corpus and produces outputs grounded in traceable sources.

RISKS

However, the same features that make LLMs powerful also introduce new and amplified methodological risks relative to conventional statistical tools. First, LLMs can generate hallucinated or unverifiable statements, which is particularly problematic in evidence synthesis and regulatory-facing narratives. Second, outputs can vary across prompts, contexts, and model versions, creating reproducibility and comparability challenges if model, prompt, and retrieval configurations are not version-controlled and archived. Third, LLM performance can degrade through data drift (changes in underlying clinical documentation patterns or coding), model drift (vendor updates), or shifts in case-mix—risks that are magnified when LLMs

are embedded in multi-step “agentic” workflows that plan, call tools, and iterate autonomously. Fourth, LLMs may inadvertently expose sensitive information (e.g., through prompt injection, retrieval leakage, or inappropriate logging) unless privacy-by-design controls are implemented. Finally, LLMs can entrench or amplify bias if the training data, retrieval corpus, or downstream labeling reflects inequities in access, documentation, or care pathways.

IMPACT AND GOVERNANCE

These risks have direct implications for evidence standards. Regulators and HTA bodies do not evaluate “AI output” in isolation; they evaluate whether the totality of evidence is credible, transparent, and fit for purpose, with clear traceability from raw data to analytic decisions to results. FDA’s RWD/RWE guidance emphasizes the importance of documenting data provenance, study design, and analytic integrity in ways that enable regulatory confidence. EMA’s experience with regulator-led RWD studies likewise highlights the need for well-governed data pipelines and methodological discipline when translating RWD analyses into regulatory decision contexts. As a result, LLM-enabled POR cannot be treated as generic productivity tooling; it must be governed as part of the statistical and evidence-generation system.

GOVERNANCE FRAMEWORKS

A practical way to operationalize this governance is to align LLM use with established, citable frameworks for AI risk management, reporting transparency, and organizational accountability. The NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) provides a cross-sector structure to identify and manage AI risks across governance, mapping, measurement, and management functions. Recognizing that generative AI introduces distinct risk modes, NIST also published the Generative AI Profile (NIST AI 600-1) as a companion resource to the AI RMF to help organizations address GenAI-specific risks and controls. At the organizational level, ISO/IEC 42001—the first AI management system standard—sets expectations for establishing and continuously improving an AI governance system, including accountability, risk management, and lifecycle controls.

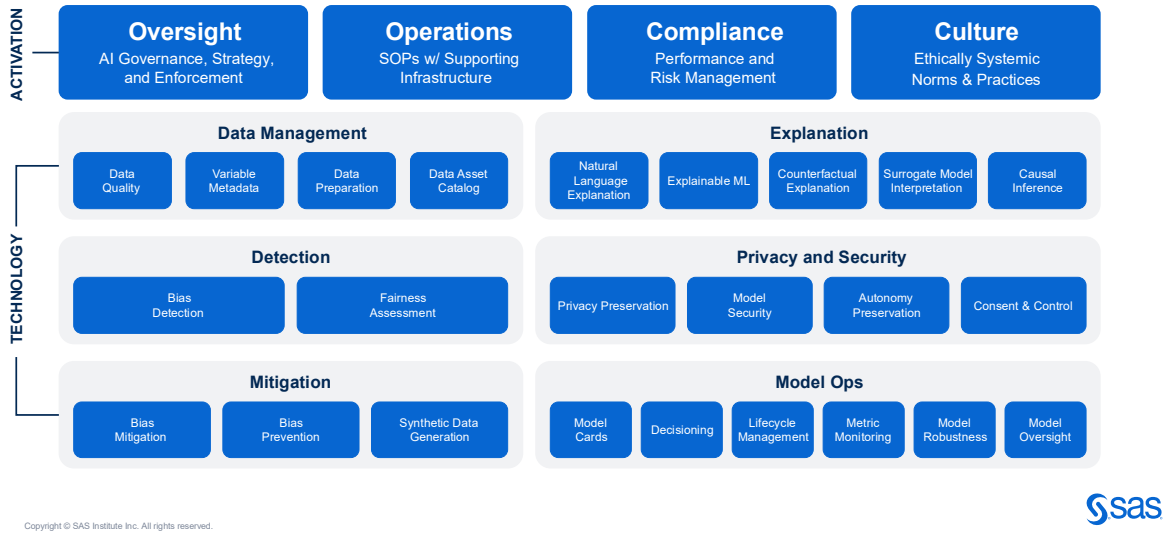

FIGURE 1: TRUSTWORTHY AI LANDSCAPE

POR SPECIFIC GUIDANCE

In parallel, POR increasingly relies on prediction models and AI-enabled methods that must be reported with rigor. The TRIPOD+AI statement updates guidance for reporting clinical prediction models using regression or machine learning, directly addressing transparency requirements that support reproducibility and critical appraisal. For AI interventions evaluated in clinical trials, CONSORT-AI (trial reporting) and SPIRIT-AI (protocol reporting) provide extensions that strengthen clarity about the AI component, intended use, data handling, and performance assessment—principles that are transferable to RWE contexts whenever AI contributes materially to endpoint ascertainment, cohort identification, or analysis. While these guidelines were not designed specifically for LLMs, their core emphasis—explicit specification of inputs, outputs, intended use, validation, and limitations—maps directly onto the governance needs of LLM-enabled POR pipelines.

Accordingly, the central methodological question is no longer whether LLMs can accelerate POR, but under what governance conditions LLMs and AI agents can produce outputs that are (1) statistically defensible, (2) reproducible across time and

teams, (3) privacy-protective, and (4) compatible with regulatory and HTA evidence expectations. This manuscript addresses that question by characterizing the methodological tradeoffs of widely used LLM ecosystems (closed, vendor-hosted models versus open-weight/self-hosted approaches) and by proposing practical controls that make LLM-enabled POR auditable. These controls include: curated retrieval corpora with provenance; prompt and model versioning; pre-specified use cases (e.g., concept extraction vs. evidence narrative); structured human adjudication for phenotype labels; evaluation harnesses for accuracy, bias, and stability; monitoring for drift; and documentation practices that align with regulatory submission norms for RWD/RWE content.

Ultimately, responsibly governed LLMs should be treated as statistical instruments within an evidence pipeline—subject to validation, traceability, and change control—rather than as general-purpose assistants. Establishing this posture enables POR teams to capture the upside of generative AI (speed, scalability, and consistency) while maintaining the evidence standards demanded for decisions that affect patient access, safety, and real-world outcomes.

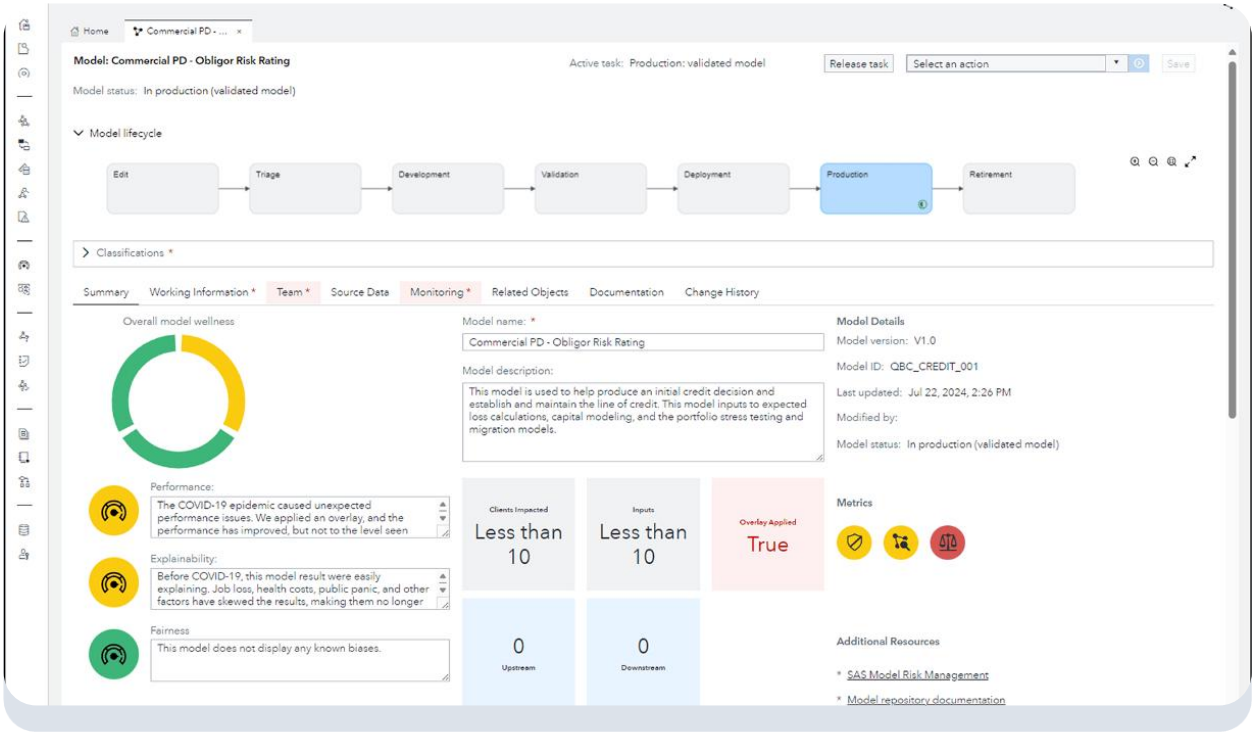

FIGURE 2: MODEL OPERATIONS AND GOVERNANCE

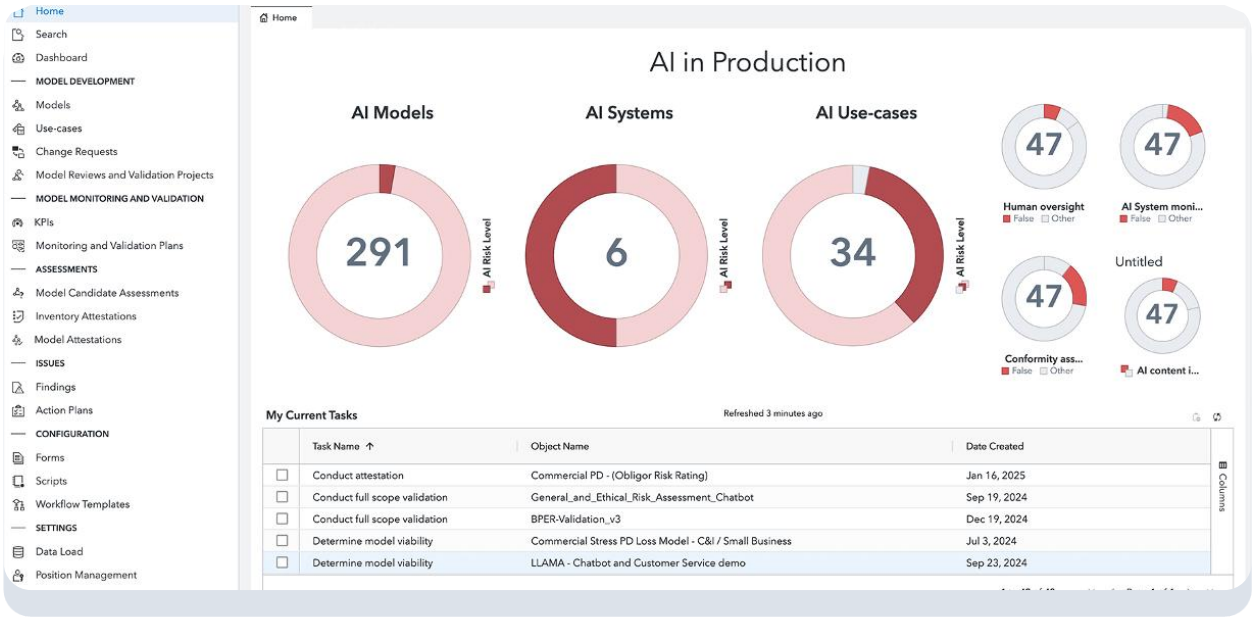

FIGURE 3: AI GOVERNANCE AND TRANSPARENCY

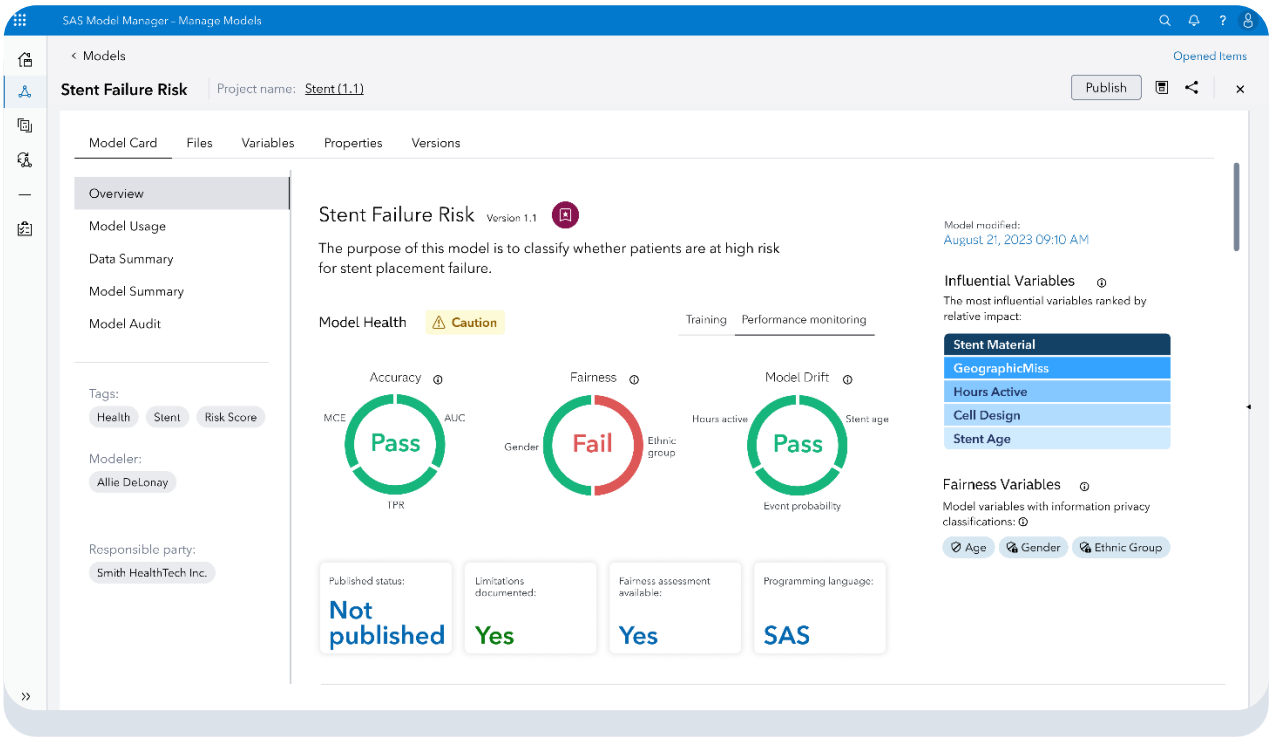

FIGURE 4: MODEL SCORECARD

MCP AND MODEL GOVERNANCE

The Model Context Protocol (MCP) is emerging as a foundational interoperability layer for large language models (LLMs), standardizing how models connect to external tools and data sources while maintaining contextual metadata during inference and execution. Originally introduced as a machine-agnostic protocol for tool integration and data access, MCP supports structured, bidirectional interactions between LLMs and external services through a JSON-RPC-based schema, enabling models to consume and reason over rich context beyond their native token input window.

In this landscape, we have developed a customized MCP server that exposes SAS procedural capabilities to LLM clients. It defines a set of SAS-oriented tool operations (e.g., `run_sas(code)`, `list_libs()`, `list_tables(libname)`) over MCP, which allows an LLM agent to request, execute, and receive results from backend SAS compute environments through a standardized protocol interface.

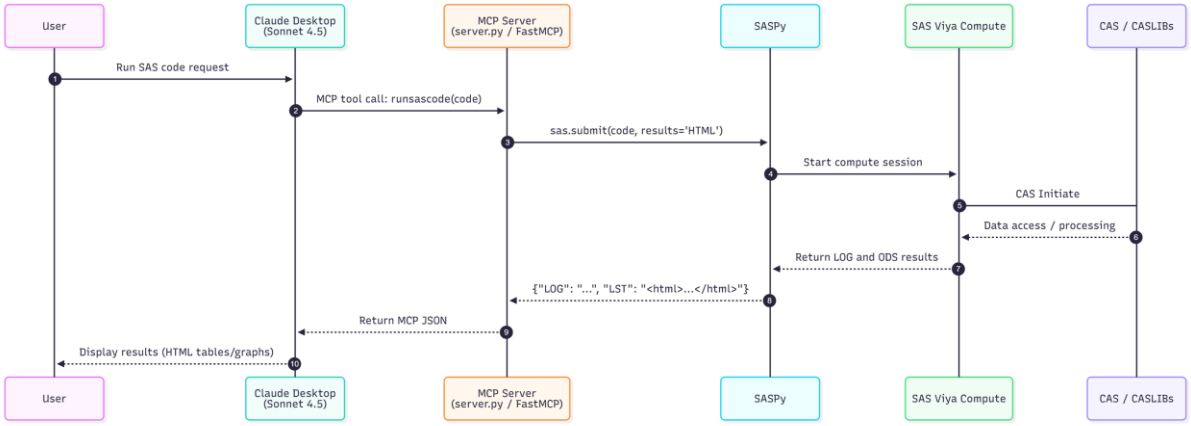

FIGURE 5: AGENTIC WORKFLOW WITH MCP SERVER AND SAS VIYA

Traceability & Governance through MCP Tool Calling
Governance of LLM behavior increasingly requires not only *control* over model actions but also *visibility* into *what* the model did, *why* it did it, and *what results* were produced. MCP tool invocations such as those implemented in sastool-mcp provide this critical traceability surface in several ways:

Structured Invocation Records. Each tool call in MCP is serialized as a structured request (e.g., tool name, function signature, arguments) and a corresponding structured response. By logging these interactions, governance frameworks can reconstruct the sequence of tool calls that led to a specific outcome, creating an auditable trail of decisions that goes beyond flat prompt→response logs.
Separation of Reasoning vs. Execution. When an LLM generates SAS code and transmits it via an MCP call (e.g., `run_sas(...)`, the protocol clearly segments *model reasoning* (code generation) from *external execution* (running the code). This separation enables compliance systems to independently evaluate the model's intent and the operational effects of that intent, improving transparency and accountability in automated analytics workflows.
Metadata Anchoring. MCP servers can propagate contextual metadata alongside tool requests and responses (including timestamps, data source identifiers, and execution logs). In practical governance settings, these metadata can be integrated with governance logs or policy engines to enforce access controls, provenance tracking, and post-hoc review of decisions that involve external computation.
Reproducibility. Because MCP interactions are explicit and machine-parseable, downstream systems can replay tool sequences with the same inputs to verify outputs. For example, a governance dashboard could re-execute a `run_sas(code)` request captured in MCP logs to validate results against regulatory expectations or internal audit requirements.

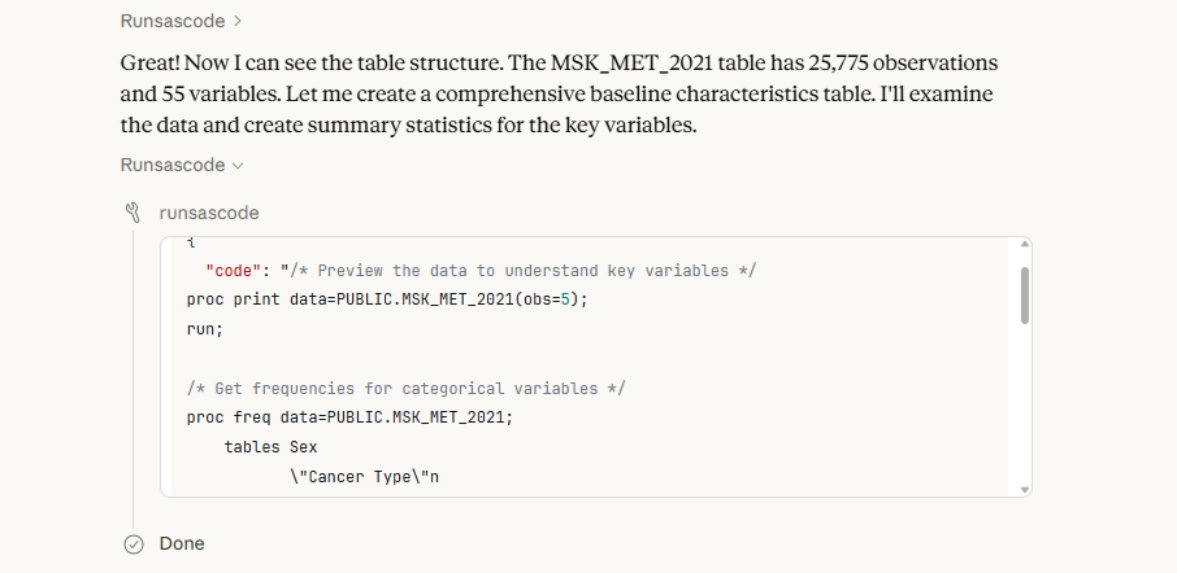

FIGURE 6: EXAMPLE OF MCP TOOL CALLING FROM A CHAT INTERFACE (*THIS EXAMPLE SHOWS AN EXAMPLE FROM ANTHROPIC CLAUDE CHAT INTERFACE*).

CONCLUSION

The future of real-world evidence is not simply AI-enabled—it is governed, traceable, and defensible by design.

As demonstrated through enterprise model registries, lifecycle management workflows, production monitoring dashboards, AI system inventories, and validation pipelines, AI in RWE must operate within a structured control environment. From model development to validation, deployment, monitoring, and retirement, every step must be documented, auditable, and reproducible. This is not operational overhead; it is methodological integrity.

Regulatory authorities have made expectations clear. FDA guidance on RWD/RWE emphasizes relevance, reliability, transparency, and auditability in submissions. The NIST AI Risk Management Framework and ISO/IEC 42001 establish structured approaches for AI governance, risk identification, monitoring, and accountability. Reporting standards such as TRIPOD-AI and CONSORT-AI reinforce reproducibility and methodological clarity in AI-enabled analyses. Together, these frameworks define the foundation for decision-grade AI-enabled evidence.

The dashboards and workflows illustrated here represent more than technical infrastructure. They embody a shift:

- From black-box modeling to transparent model cards and lineage
- From static validation to continuous performance and drift monitoring
- From ad hoc AI use to cataloged, risk-classified, human-oversight systems
- From productivity tools to statistical instruments

AI can dramatically accelerate cohort identification, signal detection, feasibility intelligence, and evidence synthesis. However, acceleration without governance creates fragility. Sustainable impact requires embedding fairness evaluation, version control, bias monitoring, explainability, and documented oversight directly into the RWE lifecycle.

The Trust Dilemma is resolved not by slowing innovation—but by governing it.

Defensible AI-enabled RWE is achieved when methodological rigor, regulatory alignment, and operational transparency are inseparable from technological advancement. In this paradigm, AI does not replace epidemiologic discipline—it amplifies it, responsibly.

RECOMMENDED READING

Data and AI Impact Report: the Trust Imperative
https://www.sas.com/en_us/gc/data-and-ai-impact-report.html

CONTACT INFORMATION

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

Author Name: Sherrine Eid, MPH
Company: SAS Institute, Inc
Email: Sherrine.Eid@sas.com
Website:www.sas.com

Author Name: Samiul Haque, PhD
Company: SAS Institute, Inc
Email: Samiul.Haque@sas.com
Website:www.sas.com