Paper AI-343

# Large Language Models and AI Agents in Patient Outcomes Research: A Technical Review, Benchmarking, and Governance Framework

Sherrine Eid, MPH, SAS Institute, Inc.

## ABSTRACT

Large language models (LLMs) and AI agent frameworks are increasingly deployed in patient outcomes research (POR) for tasks spanning automated phenotype extraction, cohort construction, literature synthesis, analytic code generation, and regulatory dossier preparation. Yet the statistical validity, reproducibility, and governance of these systems remain insufficiently characterized for real-world data (RWD) applications intended to support regulatory and health technology assessment decisions.

This paper benchmarks four publicly available LLMs — OpenAI GPT-5, Google Gemini 2.5, Anthropic Claude 4, and Meta Llama 3.1 (open-weight) — alongside six retrieval-augmented generation (RAG) agent frameworks across five governance dimensions: statistical reproducibility, data governance and privacy, transparency and auditability, bias and drift management, and fitness-for-purpose in regulated workflows. Assessments are aligned with the NIST AI Risk Management Framework (AI RMF 1.0), TRIPOD+AI (BMJ 2024), CONSORT-AI (Nature Communications 2024), and FDA RWD/E regulatory guidance including the January 2025 draft credibility guidance and December 2025 final RWE policy.

Key findings: no current LLM, used standalone, meets the full credibility requirements for unassisted regulatory-grade evidence generation; RAG architectures substantially reduce hallucination rates and improve faithfulness to 95–100% in documented pilots; Claude 4 leads on transparency and auditability; GPT-5 leads on clinical benchmark performance; and Llama 3.1 uniquely enables on-premises deployment for HIPAA-sensitive environments. Only 27% of AI oncology randomized controlled trials self-report CONSORT-AI adherence, signaling a systemic gap between deployment speed and reporting rigor. The SAS Viya platform stack — encompassing Intelligent Decisioning, MCP Server, and AI Navigator — represents the most governance-mature solution for pharma-grade AI deployments. Governance-first deployment is the prerequisite for regulatory-grade POR.

## INTRODUCTION

The application of large language models (LLMs) and AI agent frameworks to patient outcomes research (POR) is no longer speculative. Sponsors, contract research organizations, and academic medical centers are actively deploying LLMs in phenotype extraction from electronic health records (EHRs), automated cohort construction, natural language literature synthesis, statistical analysis plan (SAP)-compliant code generation, and preparation of regulatory evidence packages. The efficiency potential is real: LLM-assisted literature review offers three-fold efficiency improvements over manual screening when applied to structured regulatory documents [28].

However, the governance infrastructure required to support these applications in regulatory contexts is lagging the technology. Patient outcomes research occupies a uniquely demanding regulatory niche: unlike research-grade AI applications, POR outputs may directly support IND submissions, label expansions, comparative effectiveness dossiers submitted to health technology assessment bodies, and postmarket surveillance obligations. The evidentiary standard is correspondingly high, and the consequences of governance failure — including hallucination-propagated errors in regulatory submissions or bias-embedded evidence supporting treatment decisions — extend to patient safety.

The regulatory landscape has shifted substantially in the past eighteen months. The FDA issued its first draft guidance explicitly addressing AI in drug and biological product regulatory submissions in January 2025 [9], establishing a context-of-use (COU)-specific credibility framework for AI models. The December 2025 final RWE policy removed the requirement for individual patient-level data in certain RWE submissions, directly enabling AI-mediated evidence generation from de-identified pooled databases [10]. These developments create both urgency and opportunity: sponsors who develop governance-mature AI

workflows are positioned to accelerate evidence generation, while those who deploy AI without adequate governance documentation face increasing regulatory scrutiny.

This paper addresses the following objectives:

1. Benchmark four leading LLMs (GPT-5, Gemini 2.5, Claude 4, Llama 3.1) across five governance dimensions relevant to regulated POR;
2. Evaluate six RAG agent frameworks for fitness-for-purpose in regulated clinical research workflows;
3. Map findings to NIST AI RMF, TRIPOD+AI, CONSORT-AI, and FDA RWD/E governance requirements;
4. Present an end-to-end six-stage workflow with Human-in-the-Loop (HITL) oversight roles at each stage;
5. Provide a twelve-point governance recommendation framework for institutional deployment.

This paper is written for a PharmaSUG audience of SAS-literate pharmaceutical statisticians, data scientists, regulatory affairs professionals, and biostatisticians who are either currently deploying or evaluating LLM tools for POR use cases.

# BACKGROUND

## Governance Frameworks for AI in Patient Outcomes Research

### NIST AI Risk Management Framework (AI RMF 1.0)

The NIST AI RMF 1.0, released January 26, 2023, organizes AI risk management across four core functions: GOVERN, MAP, MEASURE, and MANAGE [11]. GOVERN establishes organizational culture, policies, executive accountability, and cross-functional team structures for AI risk oversight. MAP identifies stakeholders, documents technical and societal risks, and establishes context-of-use. MEASURE encompasses performance metrics, fairness assessments, bias tracking, and ongoing evaluation. MANAGE covers continuous monitoring, model updates, version control, and risk mitigation throughout the AI lifecycle.

The framework is technology-neutral and sector-agnostic, making it directly applicable to LLM deployments in pharmaceutical settings. Critically, NIST AI RMF does not prescribe specific metrics or compliance thresholds — it provides a process architecture that organizations must populate with domain-specific requirements. For POR, this means sponsors must develop NIST-aligned governance structures that incorporate clinical validation, statistical reproducibility requirements, and regulatory submission standards.

### TRIPOD+AI (BMJ 2024)

TRIPOD+AI provides a 27-item checklist for reporting prediction model studies using regression or machine learning methods, superseding the 2015 TRIPOD statement [12]. Key additions relevant to LLM-augmented POR include: fairness items embedded throughout all 27 checklist items requiring model performance reporting in key demographic subgroups; open science subitems requiring study protocol registration, data sharing, and code sharing; a patient and public involvement item requiring disclosure of patient engagement in study design; and explicit requirements to evaluate model performance by sociodemographic characteristics.

An important limitation for the present review: the TRIPOD+AI statement explicitly acknowledges that "foundation and large language models (such as ChatGPT) that are rapidly gaining momentum were not considered — the TRIPOD+AI guidance is primarily aimed at non-generative models." This creates a reporting gap that institutions must address through supplementary documentation when using generative AI in POR. Practical compliance requires structured workflows with complete documentation of model version, temperature settings, prompt engineering procedures, validation datasets, and human review checkpoints.

### CONSORT-AI (Nature Communications 2024)

CONSORT-AI adds 14 items to the standard 37-item CONSORT 2010 checklist, addressing algorithm version specification, input data selection methodology, handling of low-quality or missing data, performance error assessment, and data accessibility [13]. A 2024 systematic review in Nature Communications found median CONSORT-AI concordance of 90% in non-mandated journals and 100% in journals that mandated its use, confirming that editorial mandates significantly improve reporting quality. For LLM-augmented POR, CONSORT-AI compliance requires documenting the specific model version, all prompt templates, temperature and seed settings, validation against human-abstracted reference standards, and performance error rates by demographic subgroup.

**FDA RWD/E Regulatory Framework**

FDA's regulatory engagement with RWD and AI has accelerated meaningfully. The January 2025 draft guidance [9] — Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products — is the first FDA document explicitly addressing AI in drug and biological product regulatory submissions, drawing on experience with over 500 AI-component submissions since 2016. The guidance establishes: (1) a COU definition requirement specifying exactly how an AI model addresses a specific question; (2) risk-based credibility assessment scaled to model risk and COU; and (3) scope covering novel clinical trial designs, AI-based drug development tools, pharmacovigilance, AI-mediated RWD studies, and pharmaceutical manufacturing. Early pre-submission engagement is strongly recommended.

The December 2025 final RWE policy [10] removed the requirement for identifiable individual patient data in certain RWE device submissions, explicitly enabling use of de-identified databases — cancer registries, EHR networks, insurance claims databases — containing millions of patient records. This policy directly enables AI-mediated evidence generation at scale from de-identified pooled databases without sponsors needing to submit individual patient records.

**Why RWE Specifically Requires Governance**

Patient outcomes research occupies an unusually demanding governance environment. Regulatory submissions referencing POR outputs are subject to FDA/EMA scrutiny for statistical validity, data quality, and evidence credibility. Patient safety implications are direct: errors propagated through LLM-extracted RWD may affect label-supporting comparative effectiveness claims. Statistical validity requires that LLM outputs be reproducible, unbiased across demographic subgroups, and traceable to authoritative source data. Any POR workflow that shortcuts governance requirements is therefore not merely a process compliance failure — it represents an evidentiary integrity risk with downstream regulatory and safety consequences.

# METHODS

## Study Design

This paper presents a structured literature review and benchmarking synthesis; no primary data collection was conducted. The evaluation synthesizes peer-reviewed clinical benchmarks (2024–2026), vendor governance documentation, privacy architecture specifications, and regulatory alignment evidence for each model and framework. Scores and ratings represent assessed fitness for regulated POR workflows, not general-purpose AI capability rankings.

## Models Evaluated

Four LLMs were selected based on prominence in clinical benchmarking literature, healthcare enterprise adoption, and policy relevance:

- **GPT-5 / GPT-5.2** (OpenAI): Leading closed-source commercial model with enterprise healthcare deployments
- **Gemini 2.5 Pro** (Google): State-of-the-art multimodal model with independent clinical benchmark performance [1]
- **Claude 4 Sonnet / Opus** (Anthropic): Model with most comprehensive published AI governance framework [14]

- **Llama 3.1 70B/405B** (Meta): Reference open-weight model enabling on-premises deployment

### Evaluation Dimensions

Each model was evaluated across five dimensions directly relevant to regulated POR:

1. **Statistical Reproducibility:** Output consistency at temperature=0, seed-locking behavior, variability metrics from published benchmarks
2. **Data Governance and Privacy:** HIPAA eligibility, BAA availability, PHI handling architecture, data sovereignty options
3. **Transparency and Auditability:** Published governance documentation, training data disclosure, audit trail support, interpretability
4. **Bias and Drift Management:** Hallucination rates in clinical contexts, subgroup fairness evidence, temporal drift vulnerability
5. **Fitness-for-Purpose:** Task-specific performance on POR-relevant benchmarks, regulatory alignment evidence, deployment precedent in clinical settings

### RAG Framework Evaluation

Six RAG and agent orchestration frameworks were evaluated using the same five primary dimensions augmented by HITL capability and clinical/pharma track record: LangChain/LangGraph (LangChain Inc.), LlamaIndex (LlamaIndex Inc.), Haystack (deepset), AWS Bedrock Agents (Amazon), Microsoft AutoGen/Semantic Kernel, and SAS Viya with MCP Server and Intelligent Decisioning.

### Scoring Methodology

Scores (0–10 for radar analysis; star ratings for tabular comparison) reflect synthesis of published clinical benchmarks, governance documentation, privacy architecture, regulatory alignment evidence, and independent comparative evaluations. Hallucination rate estimates are drawn from the medical hallucination framework published on arXiv [8]. All scores represent assessed fitness for regulated POR workflows specifically.

### Human Expert Roles

The evaluation explicitly maps oversight responsibilities across the research team: the study clinician and principal investigator provide clinical validity assessment; the epidemiologist oversees cohort design and bias review; the biostatistician validates analytic code, reproducibility, and statistical interpretation; the data engineer manages ingestion, de-identification, and infrastructure; regulatory affairs manages COU documentation and FDA submission alignment; and the project lead coordinates HITL checkpoints and audit trail integrity.

## RESULTS

### LLM Performance Benchmarks

The table below presents clinical performance benchmarks drawn from peer-reviewed literature for each model evaluated. Readers should note important cross-study comparability limitations: benchmarks were conducted on different datasets, at different time points, and with varying methodologies. HealthBench Hard subset scores for GPT-5 originate from OpenAI's internal technical report and have not been independently peer-reviewed at time of writing.

**Table 1. LLM Clinical Performance Benchmarks for Patient Outcomes Research**

| Model | HealthBench Hard | Clinical QA Accuracy | Consistency (T=0) | Est. Hallucination Rate |
|---|---|---|---|---|
| GPT-5 (OpenAI) | 46.2%* | 93.3%† | ~92% | ~10% |
| Gemini 2.5 Pro (Google) | ~38% est. | 97.4%‡ | ~90% | ~12% |

| Model | HealthBench Hard | Clinical QA Accuracy | Consistency (T=0) | Est. Hallucination Rate |
|---|---|---|---|---|
| Claude 4 (Anthropic) | ~35% est. | 94.4–95.2%‡ | ~93% | ~8% |
| Llama 3.1 70B (Meta) | ~24% est. | 87.5%§ | >90%¶ | ~20% |

** HealthBench Hard subset figure from OpenAI internal HealthBench technical report; not independently peer-reviewed [22]. † GPT-4o on 1,181-question European critical care diploma examination [4]. ‡ Diagnostic hit rate from 5,921 MIMIC-IV clinical cases [1]. § Llama 3.1 70B on 1,181-question critical care MCQ benchmark [4]. ¶ Consistency ratio >90% at T=0 on FoundationOne genomic variant classification dataset [5]. Hallucination estimates derived from medical hallucination framework analysis [8].*

## Model-by-Model Assessment

### OpenAI GPT-5

**Clinical Performance:** GPT-5 scored highest of any model evaluated on HealthBench — a physician-designed benchmark of 5,000 realistic medical conversations graded by over 250 physicians [22]. The 46.2% Hard subset figure is drawn from OpenAI's internal HealthBench technical report; readers should note this has not been independently verified in peer review at time of writing. In critical care MCQ benchmarking, the predecessor GPT-4o achieved 93.3% on a 1,181-question European critical care diploma examination — the highest performance among four models evaluated, with the lowest output variability (standard deviation 2.3 across subdomain scores) [4].

**Data Governance:** OpenAI for Healthcare offers HIPAA-eligible deployment with signed Business Associate Agreements (BAAs) across enterprise tiers. Standard consumer ChatGPT is not HIPAA-compliant and must not be used with any PHI-adjacent data in regulated POR workflows [6]. Organizational data is not used to train shared models under enterprise agreements.

**Regulatory Alignment:** GPT-5 does not natively output TRIPOD+AI-aligned study reports or CONSORT-AI documentation. Structured prompting templates and human editorial review are required. Setting temperature to 0 maximizes reproducibility for extraction and code generation tasks.

**Limitations:** Closed-source architecture limits full auditability of training data composition, memorization risk for clinical records, or exact parameter specification. Context-of-use credibility evidence is not publicly available in FDA-submission-ready format. As with all commercial APIs, model version changes between study initiation and regulatory submission represent a material reproducibility risk requiring explicit management via API version pinning.

**Best Suited For:** High-accuracy tasks on de-identified data; HealthBench-class clinical reasoning; literature synthesis tasks requiring broad medical knowledge depth.

### Google Gemini 2.5 Pro

**Clinical Performance:** Gemini 2.5 Pro achieved the highest diagnostic hit rate in an independent head-to-head evaluation of 5,921 MIMIC-IV clinical cases — 97.4% (95% CI 97.0–97.8%) — significantly outperforming GPT-4.1, Claude 4 Opus, and Claude 4 Sonnet using an LLM-as-judge methodology [1]. RAG augmentation further improved GPT-4o performance on the same benchmark ($p<0.006$), suggesting Gemini's already-high baseline may improve further with retrieval grounding. Gemini 2.0 Flash was evaluated as a component of a two-level RAG cohort generation system achieving F1-scores up to 0.75 in patient cohort identification from EHR data [7].

**Data Governance:** Google Cloud offers HIPAA-eligible Vertex AI deployments with BAA. Consumer Gemini (Gemini.google.com) does not carry healthcare-grade compliance. Enterprise deployments require explicit configuration within Google Cloud's HIPAA-eligible services, with data residency controls and VPC configurations available.

**Limitations:** Despite strong benchmark performance, Gemini 2.5 has limited published governance documentation relative to Anthropic's Constitutional AI framework. The training data composition for medical domain performance is not disclosed in detail — a gap that is material for regulatory credibility assessments.

**Best Suited For:** Large-scale diagnostic reasoning; high-volume de-identified data extraction where clinical accuracy is the primary optimization target.

## Anthropic Claude 4 (Sonnet and Opus)

**Governance Architecture:** Claude's Constitutional AI framework [14], updated January 2026, represents the most comprehensive publicly available AI governance document among the evaluated models. It establishes a four-tier priority hierarchy: (1) broadly safe, (2) broadly ethical, (3) compliant with Anthropic's guidelines, (4) genuinely helpful — directly supporting regulated deployment contexts. Anthropic signed the EU General-Purpose AI Code of Practice in July 2025, providing presumptive conformity with EU AI Act requirements for high-risk systems [26]. The framework includes hardcoded behavior prohibitions and published soft-coded defaults, enabling institutional compliance mapping.

**Clinical Performance:** Claude 4 Sonnet achieved 94.4% and Claude 4 Opus achieved 95.2% diagnostic hit rates in the 5,921-case head-to-head evaluation [1]. Claude 3.5 Sonnet was used as the parsing LLM in the highest-performing RAG cohort construction pipeline in the Bayer AG two-level RAG study [7]. The Constitutional AI framework's emphasis on calibrated uncertainty expression reduces overconfidence hallucination — a particularly valuable property for clinical evidence generation.

**Transparency and Auditability:** Claude's published governance architecture enables institutions to map model behavior to NIST AI RMF GOVERN function requirements. The model explicitly avoids undermining human oversight — a behavioral property directly aligned with FDA's HITL requirements for AI credibility. This is the most documentation-rich governance basis among the evaluated models.

**Best Suited For:** Regulated submissions where auditability is critical; workflows requiring mappable governance documentation; settings where EU AI Act compliance must be demonstrated.

## Meta Llama 3.1 (Open-Weight)

**Architecture and Data Sovereignty:** Llama 3.1 405B, released July 2024, was the first openly available model to rival frontier closed-source models on general knowledge benchmarks. Trained on over 15 trillion tokens with a 128K context window. As an open-weight model, Llama 3.1 can be deployed entirely on-premises or within institutional private cloud infrastructure. No data traverses external APIs — the critical differentiator for HIPAA-sensitive POR workflows where raw PHI may be involved. Multiple hospital systems have adopted on-premises LLM deployment approaches to ensure no PHI leaves institutional boundaries [6].

**Healthcare Fine-Tuning:** The open-weight architecture enables domain-specific fine-tuning. Saama's OpenBioLLM (8B and 70B), built on Llama 3.1 and deployed across clinical trial operations, generates clinical trial protocols and study reports. Healthcare fine-tuning on OMOP CDM and eMERGE phenotyping datasets can substantially improve domain-specific extraction accuracy over baseline.

**Clinical Benchmarks:** In critical care MCQ benchmarking, Llama 3.1 70B scored 87.5% versus GPT-4o's 93.3% on the 1,181-question examination [4]. In genomic variant classification (OncoKB/CIViC), GPT-4o outperformed Llama 3.1 (accuracy 0.73 vs. 0.50 on FoundationOne dataset), though RAG augmentation and prompt engineering substantially closed the gap [5].

**Best Suited For:** PHI-sensitive on-premises deployments; workflows where data sovereignty is non-negotiable; institutional fine-tuning for disease-specific or coding-standard-specific performance.

## RAG Agent Framework Comparison

RAG architectures substantially improve LLM fitness for regulated POR by grounding outputs in authoritative, verifiable source documents. In a regulatory compliance pilot, RAG systems achieved 95–100% faithfulness on clinical trial protocol queries against FDA E9 guidance, with ClinPharm accuracy metrics ranging 65.9–88% depending on query type [2]. RAG also significantly improved GPT-4o diagnostic hit rate in the MIMIC-IV clinical benchmark ($p<0.006$) [1].

**Table 2. RAG Agent Framework Comparison for Regulated POR**

| Dimension | LangChain/ LangGraph | LlamaIndex | Haystack | AWS Bedrock | MS AutoGen/ SK | SAS Viya + MCP |
|---|---|---|---|---|---|---|
| Audit Trail Quality | ★★★ (LangSmith) | ★★ | ★★★★ (SOC 2) | ★★★★★ (CloudTrail) | ★★★★ (Azure) | ★★★★★ (21 CFR) |
| 21 CFR Part 11 | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| HIPAA Eligible | Via cloud | Via cloud | SOC 2 | ✓ (BAA) | ✓ (Azure BAA) | ✓ |
| HITRUST Certified | Via cloud | Via cloud | ✗ | ✓ (177+ svc) | ✓ (Azure) | Enterprise |
| PHI/PII Controls | Requires config | Presidio int. | RBAC + VPC | Guardrails native | AI Content Safety | Built-in masking |
| Reproducibility | ★★★ | ★★ | ★★★ | ★★★ | ★★★ | ★★★★★ |
| Regulatory Alignment | Developing | Developing | Moderate | Strong | Moderate–Strong | Very Strong |
| HITL Native | ★★★★★ (LangGraph) | ★★ | ★★★ | ★★★ (A2I) | ★★★★ (SK Process) | ★★★ |
| Pharma Track Record | Emerging | Emerging | Growing | Strong | Strong | Established (decades) |
| CDISC Native | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Overall POR Fitness | **Moderate** | **Moderate** | **Moderate–Strong** | **Strong** | **Moderate–Strong** | **Strong** |

**Key Framework Differentiators:**

**LangChain/LangGraph:** The most widely deployed open-source framework, surpassing 1 billion cumulative downloads as of March 2026. LangGraph's first-class native HITL support — workflows can pause at any node, await human approval, then resume or roll back — is the strongest HITL implementation in this comparison. LangSmith has processed over 15 billion traces across enterprise customers. However, achieving 21 CFR Part 11 compliance requires substantial custom infrastructure engineering. Best used as the orchestration layer within a purpose-built, compliance-hardened stack.

**LlamaIndex:** Strong document-heavy RAG pipelines and native Microsoft Presidio integration for PHI masking of 18+ entity types before data reaches the LLM. Audit trail depth and regulatory certifications are limited. Best suited as the knowledge retrieval layer within a larger compliant architecture.

**Haystack/deepset:** Holds SOC 2 Type II, ISO 27001, GDPR, and HIPAA certifications — the strongest certification portfolio among open-source-origin frameworks. Serializable pipeline graphs enable deterministic replay. Recommended for organizations seeking flexible, open-source-compatible frameworks with enterprise governance.

**AWS Bedrock Agents:** Most mature certifiable compliance posture of any cloud-native platform. HIPAA eligibility with BAA, HITRUST CSF certification across 177+ services, FedRAMP HIGH authorization, and immutable CloudTrail audit logs. Amazon Bedrock Guardrails provide configurable PII/ePHI redaction, prompt injection protection, and contextual hallucination detection. Amazon Augmented AI (A2I) is a purpose-built HITL service routing low-confidence outputs to human reviewers.

**Microsoft AutoGen/Semantic Kernel:** The Semantic Kernel Process Framework provides stateful, durable HITL workflow modeling with human decision points as first-class workflow components. Azure ecosystem compliance (HIPAA, HITRUST, SOC 2, FedRAMP HIGH) provides a strong compliance foundation. AutoGen itself remains research-grade and is not independently compliance-certified.

**SAS Viya + MCP Server + Intelligent Decisioning + AI Navigator:** The most purpose-built and regulatorily mature platform for POR, with native 21 CFR Part 11 compliance, CDISC alignment (SDTM/ADaM/dataset-JSON), and a decades-long FDA-accepted track record. See dedicated section below.

## NIST AI RMF Alignment

The four NIST AI RMF functions map differently to each LLM's native capabilities.

**GOVERN Function:** Claude 4 scores highest on GOVERN alignment due to Anthropic's published Constitutional AI framework, EU AI Act Code of Practice signature, and documented four-tier priority hierarchy. For POR deployments, organizations should establish a cross-functional AI Governance Committee — including clinical leadership, CISO, regulatory affairs, biostatistics, and data science — with defined escalation paths for high-risk AI outputs. All models require institutional governance overlays; none provide complete GOVERN function compliance natively.

**MAP Function:** MAP requires identifying stakeholders, documenting risks across technical and societal dimensions, and establishing context-of-use. No model provides MAP documentation natively — this function is entirely institutional. Key MAP tasks for POR include: documenting intended use cases, identifying affected patient populations, characterizing training data representativeness, and assessing potential for algorithmic bias. Llama 3.1's inspectable model weights uniquely enable MAP-function training data analysis unavailable to closed-source models.

**MEASURE Function:** Gemini 2.5 and GPT-5 score highest on raw clinical performance metrics. Claude 4 leads on consistency and calibrated uncertainty. Llama 3.1 uniquely enables full model inspection for MEASURE documentation. Hallucination rates (estimated 8–20% in clinical contexts) must be treated as primary MEASURE metrics in regulated deployments. Subgroup performance disaggregation — by race, ethnicity, sex, age, and site of care — is required by TRIPOD+AI fairness items [12] and FDA's expectations for demographic transparency in AI-enabled submissions.

**MANAGE Function:** On-premises Llama 3.1 deployments enable the tightest MANAGE controls: model versioning, fine-tuning audit trails, no external API dependencies, and full lifecycle control. Commercial APIs introduce dependency on vendor update cycles. For all commercial models, API version pinning is required: calls must specify exact model version (not "latest") with the version documented in study records. Predetermined Change Control Plans adapted from FDA's December 2024 PCCP guidance for medical devices provide a template for AI change management in POR.

**Table 3. NIST AI RMF Alignment by Model**

| NIST Function | GPT-5 | Gemini 2.5 | Claude 4 | Llama 3.1 | Required Institutional Action |
|---|---|---|---|---|---|
| GOVERN | Moderate | Moderate | Strong | Moderate | AI Governance Committee; policy documentation |
| MAP | Moderate | Limited | Moderate | Strong (inspectable) | COU documentation; stakeholder analysis; bias inventory |
| MEASURE | Strong (benchmarks) | Strong (benchmarks) | Moderate–Strong | Moderate | Task-specific validation; subgroup analysis |
| MANAGE | Moderate (API) | Moderate (API) | Moderate (API) | Strong (on-prem) | Version pinning; PCCP adaptation; drift monitoring |

## TRIPOD+AI and CONSORT-AI Compliance

### TRIPOD+AI (27-Item Checklist)

TRIPOD+AI fairness requirements are embedded across all 27 checklist items, making demographic subgroup reporting a pervasive rather than isolated obligation [12]. For LLM-augmented POR, compliance requires: model version specification (e.g., gpt-5-20260101), temperature and seed documentation, prompt template versioning, validation methodology description, and human review checkpoint documentation — all items that must be designed into the study protocol before initiation.

The scope limitation is significant: TRIPOD+AI does not explicitly address generative AI models. This creates a documentation gap that institutions must fill with supplementary governance documentation when LLMs are used in any prediction model or evidence generation role.

**CONSORT-AI (14 Additional Items)**

A 2025 BMJ Oncology systematic review of 57 AI RCTs in oncology found only 27% self-reported adherence to CONSORT-AI, with particular gaps in methodology items related to reproducibility — specifically algorithm version specification, input data selection documentation, and performance error reporting [3]. This finding is alarming given that CONSORT-AI was published in 2020 and has been available for five years.

For sponsors using LLMs to generate or process evidence that appears in regulatory submissions, CONSORT-AI non-compliance represents direct regulatory risk. FDA reviewers evaluating AI-generated RWE can reasonably expect documentation meeting established reporting standards.

**Table 4. Reporting Framework Compliance Overview**

| Framework | Key Requirements for LLM Use | Current Compliance Gap |
|---|---|---|
| NIST AI RMF | Govern/Map/Measure/Manage across AI lifecycle | Institutional governance overlay required; not model-native |
| TRIPOD+AI | 27-item checklist; fairness; open science; model specification | Generative AI not explicitly addressed; supplementary docs needed |
| CONSORT-AI | Algorithm version, input data selection, error reporting | 27% self-reported adherence in published AI RCTs [3] |
| FDA RWD/E | COU definition; credibility plan; audit trail; HITL documentation | Jan 2025 draft not finalized; early FDA engagement recommended |
| HIPAA | PHI encryption, RBAC, BAA, audit logs | Consumer LLM use with PHI = violation; enterprise deployments available |

## SAS Platform Governance

SAS Institute has long represented the statistical computing standard for pharmaceutical regulatory submissions, and the company's 2024–2026 platform evolution positions SAS Viya, SAS Intelligent Decisioning, SAS MCP Server, and SAS AI Navigator as a governance-native ecosystem for AI-augmented POR.

**SAS® Viya and AI Life Cycle Governance**

SAS Viya provides the most comprehensive, natively compliant audit trail capability of any framework in this comparison [17]. SAS Clinical Acceleration, launched November 2025 and built on Viya, explicitly supports FDA 21 CFR Part 11 requirements via audit trails, electronic signatures, versioning, and role-based privileges. CDISC compliance (SDTM/ADaM/dataset-JSON/CDISC CORE) is supported natively — the only framework in this review with this capability.

SAS Viya for Healthcare includes built-in anonymization and masking of personally identifiable patient information for regulatory compliance, with HIPAA, 21 CFR Part 11, and GDPR compliance built into the platform configuration. The SAS Viya Data Maker capability supports synthetic data generation for model training and validation when real patient data access is restricted. Embedded model interpretability and fairness and bias monitoring provide continuous bias audits across the model lifecycle.

**SAS® Intelligent Decisioning**

SAS Intelligent Decisioning [15] enables configurable, customizable human-AI autonomy levels — a direct governance response to the variability in HITL requirements across different POR workflow stages. The Decision Flow Builder provides a visual interface for designing decision workflows with explicit epidemiologist sign-off gates, biostatistician review checkpoints, and regulatory affairs approval stages.

Real-time deployment capability exceeding 7,000 transactions per second supports production-scale deployment in pharmacovigilance and outcomes monitoring workflows. SAS announced these AI agents with customizable human-AI interaction capabilities at SAS Innovate 2025 [19].

### SAS® MCP Server

The SAS Model Context Protocol (MCP) Server [18] provides OAuth2/PKCE-governed LLM access to Viya analytics and data assets via standardized tool interfaces. This architecture enables AI agents and LLMs to execute controlled scoring, access the model inventory, and trigger job execution — all within governed access controls that prevent unauthorized model invocation or data exposure.

### SAS® AI Navigator

SAS AI Navigator [16] functions as an enterprise AI asset registry with structured policy packs specifically designed for pharmaceutical governance contexts. Policy packs are available for NIST AI RMF, FDA guidance alignment, TRIPOD+AI, and CONSORT-AI — providing a machine-readable governance layer that maps platform capabilities to regulatory requirements.

**Table 5. SAS Platform Governance Components**

| Component | Primary Function | Key Capability | Regulatory Alignment |
|---|---|---|---|
| SAS® Viya | End-to-end AI lifecycle governance | 21 CFR Part 11; CDISC; Data Maker | FDA, ICH E9, GAMP 5 |
| Intelligent Decisioning | Configurable HITL autonomy | Decision Flow Builder; >7,000 tx/sec; epi sign-off gates | FDA HITL requirements |
| MCP Server | Governed LLM-Viya access | OAuth2/PKCE; model inventory; controlled scoring | Access control compliance |
| AI Navigator | Enterprise AI asset registry | Policy packs (NIST, FDA, TRIPOD+AI, CONSORT-AI); audit trail | NIST AI RMF; FDA credibility |

## End-to-End Workflow with Human-in-the-Loop

The following six-stage pipeline maps AI-augmented POR from RWD ingestion through regulatory evidence submission, with HITL oversight roles at each stage.

**Table 6. Six-Stage AI-Augmented POR Workflow with HITL Roles**

| Stage | Key AI Tasks | Primary HITL Role | HITL Action | Governance Output |
|---|---|---|---|---|
| 1: RWD Ingestion | Data mapping to OMOP CDM; PHI de-identification; provenance tagging | Data Engineer | Validates de-identification; confirms BAA | Data provenance record; de-identification attestation |
| 2: Cohort Construction | Inclusion/exclusion criteria parsing; SQL generation; F1-based validation | Epidemiologist | Reviews cohort definition; validates temporal logic | Cohort definition document; F1 validation metrics |
| 3: Literature Review | RAG-based retrieval; relevancy scoring; evidence grading | Study Clinician / PI | Reviews evidence; validates clinical interpretation | Annotated evidence summary; faithfulness metrics |
| 4: Code Generation | SAP-compliant code; T=0; prompt versioning | Biostatistician | Reviews code; validates statistical methods | Code review log; SAP compliance attestation |
| 5: Bias/Drift Mgmt | Subgroup fairness analysis; drift detection; hallucination flagging | Epi + Biostatistician | Interprets fairness metrics; approves mitigation | Fairness report; drift monitoring log |

| Stage | Key AI Tasks | Primary HITL Role | HITL Action | Governance Output |
|---|---|---|---|---|
| 6: Reg. Submission | CAP compilation; TRIPOD+AI/CONSORT-AI checklists; IES prep | Reg. Affairs + Lead | Reviews credibility package; final sign-off | CAP; completed checklists; IES |

**Stage 1 — RWD Ingestion and Governance:** Real-world data sources — EHRs, medical claims, registries, and patient-reported outcomes — must be ingested under rigorous governance before any LLM interaction. FDA's December 2025 policy establishes clear expectations for data provenance documentation, quality assessment records, and audit trails from source to analysis [10]. Mapping source data to the OMOP Common Data Model (CDM) prior to LLM exposure standardizes terminology and reduces downstream extraction error. De-identification must follow HIPAA Safe Harbor or Expert Determination methodology. PHI must never be transmitted to public, multi-tenant LLMs [6].

**Stage 2 — Feasibility and Cohort Construction:** A 2025 arXiv study by Bayer AG demonstrated that a two-level RAG system using Claude 3.5 Sonnet, Gemini 2.0 Flash, GPT-4o, and Llama 3.1 70B for SQL-based cohort generation from EHR data achieved F1-scores up to 0.75 in cohort identification from natural language inclusion/exclusion criteria [7]. GPT-4 produced the highest-quality phenotyping algorithm drafts for T2DM, dementia, and hypothyroidism in JAMIA 2024 benchmarking, though all models required human expert review before deployment against validated eMERGE-network benchmarks [21].

**Stage 3 — Literature Review and Evidence Synthesis:** RAG-based systems restricted to curated regulatory corpora demonstrate answer relevancy at 100% and faithfulness at 95–100% for regulatory compliance queries — substantially better than unconstrained LLM generation [2]. LLM-assisted literature review offers three-fold efficiency improvements over manual screening for structured regulatory documents [28].

**Stage 4 — Analytic Code Generation:** Key requirements for regulated use include: prompt versioning with immutable logging, temperature set to 0 for reproducible outputs, seed-locking in all stochastic operations, and independent biostatistician review of all generated code prior to execution on study data. A RAG pilot evaluating Phase 2a SAP compliance against FDA E9 guidance achieved 85.7% accuracy and 100% faithfulness, with primary gaps in dose-escalation handling [2].

**Stage 5 — Bias and Drift Management:** Flatiron Health's 2025 health equity study on over 25,000 patients with HR+/HER2- metastatic breast cancer found that LLM-extracted data replicated human-abstracted patterns of racial and ethnic health inequity — demonstrating both scalability and potential propagation of existing disparities [23]. LLMs have training knowledge cutoffs that create model drift as clinical guidelines, ICD codes, and drug labels evolve — a particular concern for longitudinal POR studies. RAG with regularly updated corpora mitigates but does not eliminate this risk.

**Stage 6 — Regulatory Evidence Submission:** The output layer must produce audit-ready documentation: a Credibility Assessment Plan (CAP) per FDA January 2025 draft guidance [9], completed TRIPOD+AI 27-item checklist, CONSORT-AI checklist for any interventional component, and an Integrated Evidence Summary (IES). FDA's December 2025 RWE policy explicitly enables de-identified database use for regulatory evidence generation at scale [10].

## DISCUSSION

### No Standalone LLM Meets Full Regulatory Credibility

The benchmark data presented in this review confirm meaningful clinical performance across all four evaluated models. However, clinical benchmark performance — even at the 97.4% diagnostic hit rate achieved by Gemini 2.5 [1] — does not constitute regulatory credibility for RWE applications. FDA's January 2025 draft guidance [9] requires COU-specific credibility plans, complete audit trails, and performance evidence commensurate with model risk. None of the evaluated models provides this documentation natively.

## RAG Is Non-Negotiable for Regulated Use

RAG systems operating over curated, authoritative corpora achieve 95–100% faithfulness in regulatory compliance tasks and substantially reduce hallucination rates relative to unconstrained generation [2]. A 2025 Nature review confirmed that RAG enables more reliable healthcare AI by leveraging retrieval of external knowledge [20]. Medical hallucinations in foundation models arise from autoregressive training objectives that prioritize token-likelihood over epistemic accuracy, producing fabricated medications, contraindicated drug recommendations, false imaging interpretations, and fabricated patient histories [8]. Uncontrolled hallucination rates of 15–30% in general contexts are operationally disqualifying for regulatory evidence generation; RAG with faithfulness evaluation and HITL escalation brings this risk to a manageable level.

## Model Selection Should Follow Use-Case Risk

The appropriate model selection framework for regulated POR is risk-stratified by data sensitivity and workflow criticality:

- **Highest-sensitivity PHI-touching workflows (raw EHR, identifiable data):** Llama 3.1 on-premises. No data traverses external APIs; full data sovereignty.
- **Audit-critical regulatory submissions:** Claude 4 with Constitutional AI governance documentation providing the richest basis for NIST AI RMF GOVERN function compliance [14][26].
- **Maximum clinical accuracy on de-identified data:** GPT-5 or Gemini 2.5 via HIPAA-eligible enterprise APIs with BAA.
- **High-volume de-identified extraction:** Gemini 2.5 Pro, given the highest validated diagnostic accuracy in independent benchmarking [1].

## The Benchmark Fallacy

The "Benchmark Fallacy" — relying on MedQA or HealthBench scores to assess LLM fitness for RWE generation — is a critical and common error in AI deployment for regulated research [27]. MedQA performance does not reflect clinical data extraction complexity under conditions of ambiguous clinical text, handling of conflicting coded information, or faithfulness to source documents when the ground truth is incomplete. GPT-5's HealthBench Hard score of 46.2% [22] — drawn from OpenAI's internal report, not independent peer review — reflects performance on a physician-designed conversation benchmark. It provides no direct evidence of fitness for OMOP CDM phenotype extraction or SAP-compliant code generation.

## Temporal Drift and Knowledge Cutoffs

Training knowledge cutoffs represent an underappreciated validity threat in longitudinal POR. A model's coding knowledge may not reflect ICD-11 updates, newly approved drugs, or revised guideline-concordant care definitions at time of study completion. RAG with regularly updated corpora mitigates this risk for literature synthesis and guideline queries; phenotype extraction algorithms and coding-based endpoints remain vulnerable. Model version locking at study initiation and prospective re-validation at submission are required study governance controls.

## TRIPOD+AI/CONSORT-AI Compliance Gap

The 27% CONSORT-AI adherence rate in published AI oncology RCTs [3] five years after CONSORT-AI publication signals a systemic failure of the field — not a minority problem. Regulatory sponsors, pharmaceutical companies, and their CROs must embed CONSORT-AI and TRIPOD+AI compliance as protocol-level design requirements, not post-hoc reporting additions. The FDA's January 2025 draft guidance [9] creates the regulatory pressure to make this shift; the institutional process design must follow.

## SAS Platform Stack as a Governance Solution

For organizations requiring the deepest regulatory compliance baseline, the SAS Viya + Intelligent Decisioning + MCP Server + AI Navigator stack provides capabilities that no other framework in this review matches natively: 21 CFR Part 11 compliance, CDISC native support, FDA-accepted statistical outputs, configurable HITL autonomy levels, and governance policy packs aligned with NIST AI RMF and FDA guidance [15][16][17][18]. The emerging dominant enterprise pattern is SAS Clinical Acceleration for submission-grade analytics combined with AWS Bedrock or Azure AI for operational GenAI workflows and LangGraph or Semantic Kernel as the orchestration layer — achieving regulatory depth through SAS while maintaining development agility through open-source and cloud frameworks.

## GOVERNANCE RECOMMENDATIONS

The following twelve-point framework organizes governance requirements across three phases: pre-study, ongoing study, and regulatory submission.

### Pre-Study Governance (Points 1–4)

#### 1. AI Model Risk Classification

Classify each LLM use case per NIST AI RMF MAP function before study initiation. Literature search assistance is low risk; automated primary endpoint data extraction is high risk. Document the risk classification in the study protocol. Risk classification drives downstream validation intensity and HITL requirement specifications.

#### 2. Credibility Assessment Plan

Develop a COU-specific Credibility Assessment Plan (CAP) per FDA January 2025 draft guidance [9] before study initiation. The CAP must specify: model version and vendor, temperature and seed settings, prompt templates, intended use scope, validation plan including reference standard methodology, HITL procedures and personnel qualification requirements, and performance thresholds that trigger human escalation.

#### 3. Privacy Architecture Design

Select deployment architecture appropriate to data sensitivity tier: on-premises Llama 3.1 for raw PHI-touching workflows; HIPAA-eligible cloud enterprise APIs with BAA for de-identified EHR derivatives; any enterprise API for fully de-identified aggregated data [6]. Execute Business Associate Agreements before any data contact. PHI must never be transmitted to public, multi-tenant LLMs.

#### 4. Prospective Study Protocol Registration

Register the study protocol prospectively, including AI model specification, prompting approach, validation methodology, and HITL procedures, per TRIPOD+AI open science requirements [12]. Protocol registration establishes an immutable pre-study commitment to methodology that supports regulatory credibility assessment.

### Ongoing Study Governance (Points 5–8)

#### 5. Immutable Audit Trail

Log all LLM interactions — prompts, outputs, timestamps, model version, temperature, seed — in an immutable audit system accessible for regulatory review. For SAS Viya deployments, the native 21 CFR Part 11 audit trail fulfills this requirement. For other platforms, purpose-built audit infrastructure must be engineered at the deployment layer.

#### 6. HITL Checkpoint Documentation

Document all human expert review actions with reviewer identity, date, qualification, and a record of any modifications to LLM outputs. HITL documentation is the primary evidence that regulatory submissions will cite to demonstrate that AI-generated content has been validated by qualified human experts before use in regulatory decisions.

#### 7. Subgroup Fairness Monitoring

Monitor LLM performance by demographic subgroups — race, ethnicity, sex, age, site of care, insurance type — throughout the study. Flag any emergence of differential performance. A Nature npj Digital Medicine 2025 framework identifies availability bias and overconfidence as the primary bias mechanisms in clinical LLMs [24]. Systematic subgroup audits are required under TRIPOD+AI items.

### 8. Temporal Drift Detection

For studies exceeding six months in duration, re-evaluate model performance against a holdout validation set at defined intervals to detect temporal drift. Document re-evaluation results and any remediation actions. Version-lock the model at study initiation; if model updates are required, treat the update as a protocol amendment requiring governance committee review.

## Submission Requirements (Points 9–12)

### 9. CONSORT-AI Checklist Completion

Complete all 14 CONSORT-AI items plus standard CONSORT 2010 items for any randomized evaluation component [13]. Specific items requiring LLM-specific documentation: algorithm version (exact model version string), input data selection (prompt templates, retrieval configurations), low-quality data handling (missing data procedures), performance error reporting (hallucination rates, faithfulness scores by subgroup), and data accessibility (prompt template archiving).

### 10. TRIPOD+AI Checklist Completion

Complete all 27 TRIPOD+AI items for any prediction model component [12]. Where generative AI components exceed the scope of TRIPOD+AI's non-generative focus, supplement with institutional documentation covering: generative model architecture, training data representativeness, calibrated uncertainty expression methodology, and hallucination detection and mitigation procedures.

### 11. Model Credibility Evidence Package

Compile the full model credibility evidence package for regulatory submission: validation metrics against human-abstracted reference standards, faithfulness scores from RAGAS or equivalent framework, hallucination rates by task type, HITL documentation with reviewer qualifications, and cross-reference to study conclusions citing AI-generated evidence.

### 12. Early FDA Engagement

For novel LLM-supported regulatory strategies, request a Type B meeting or equivalent pre-submission interaction before data collection begins. FDA has engaged with over 500 AI-component submissions since 2016 [9] and has developed substantial expertise. Early engagement allows sponsors to align on COU scope, credibility evidence requirements, and submission format expectations before committing to a study design.

## CONCLUSION

LLMs and RAG-based AI agents represent a genuine acceleration opportunity for patient outcomes research — but one that requires governance-first deployment to achieve regulatory-grade validity.

**Governance-first deployment is the prerequisite.** No LLM should be deployed for regulatory POR without RAG grounding, HITL validation, and institutional governance documentation. The efficiency gains are real; the validity risks without these controls are disqualifying for regulatory submissions. RAG grounding, HITL validation, and institutional governance documentation are the three non-negotiables.

**RAG is not optional.** RAG systems operating over curated, authoritative corpora achieve 95–100% faithfulness in regulatory compliance tasks [2] and substantially reduce hallucination rates that would otherwise be disqualifying for regulatory evidence. Unconstrained standalone LLM generation introduces faithfulness risks that cannot be managed to a sufficient standard without retrieval grounding.

**Model selection must follow use-case risk.** For PHI-sensitive workflows: Llama 3.1 on-premises. For audit-critical regulated submissions: Claude 4 with Constitutional AI governance documentation [14]. For maximum clinical accuracy on de-identified data: GPT-5 or Gemini 2.5 via HIPAA-eligible enterprise APIs [1]. The Benchmark Fallacy — using HealthBench or MedQA scores as proxies for RWE fitness — must be actively resisted in institutional deployment decision-making [27].

**FDA's credibility framework is now operational.** The January 2025 draft guidance [9] and December 2025 RWE policy [10] create clear regulatory expectations and urgency for sponsors to develop COU-specific AI credibility plans. TRIPOD+AI and CONSORT-AI compliance requires institutional process design embedded at the protocol level, not post-hoc documentation. The 27% CONSORT-AI adherence rate in published AI oncology RCTs [3] represents the current state of the field — not the standard required for regulatory-grade submissions.

**The field is moving faster than its reporting standards.** The pace of LLM deployment in pharmaceutical research is outrunning both reporting framework development and regulatory infrastructure. Only 27% of AI RCTs in oncology self-report CONSORT-AI adherence [3]; TRIPOD+AI does not yet address generative models [12]. Regulatory sponsors, journal editors, and HTA bodies must collaborate to close this gap before LLM-supported evidence becomes embedded in label-supporting submissions without adequate governance documentation. The tools for governance-grade AI deployment exist today. The institutional will to use them is the remaining variable.

## REFERENCES


[1] Goh E, et al. "Evaluating Large Language Models as Diagnostic Decision Support Tools Across a Range of Clinical Cases." JMIRx Med, August 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12396308/

[2] Triplett L, et al. "Retrieval-Augmented Generation for Regulatory Compliance in Clinical Pharmacology: A Pilot Study." CPT: Pharmacometrics and Systems Pharmacology, February 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC12917324/

[3] Singh H, et al. "CONSORT-AI Adherence in Artificial Intelligence Oncology Randomized Controlled Trials: A Systematic Review." BMJ Oncology, August 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12414185/

[4] Patel SB, et al. "Large Language Models in Critical Care: A Benchmark Against the European Diploma in Intensive Care Medicine Examination." Critical Care, February 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11809097/

[5] Wu J, et al. "Benchmarking Large Language Models for Genomic Variant Classification Using FoundationOne CDx Data." NPJ Precision Oncology, May 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12078457/

[6] Nehra A, et al. "Protected Health Information and Public Large Language Models: A Scoping Review." JMIR, November 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12680930/

[7] Pfohl S, et al. "Two-Level Retrieval-Augmented Generation for Cohort Construction from Electronic Health Records." arXiv, February 2025. https://arxiv.org/html/2502.21107v2

[8] Umapathi LK, et al. "Med-HALT: Medical Domain Hallucination Test for Large Language Models." arXiv (MIT/Harvard), 2025. https://arxiv.org/html/2503.05777v2

[9] U.S. Food and Drug Administration. "FDA Proposes Framework to Advance Credibility of AI Models Used in Drug and Biological Product Submissions." FDA Press Release, January 2025. https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions

[10] U.S. Food and Drug Administration. "FDA Eliminates Major Barrier to Using Real-World Evidence in Drug and Device Application Reviews." FDA Press Release, December 2025. https://www.fda.gov/news-events/press-announcements/fda-eliminates-major-barrier-using-real-world-evidence-drug-and-device-application-reviews

[11] National Institute of Standards and Technology. "Understanding the NIST AI Risk Management Framework." Data Brackets Blog, 2023. https://databrackets.com/blog/understanding-the-nist-ai-risk-management-framework/

[12] Collins GS, et al. "TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods." BMJ, 2024;385:e078378. https://www.bmj.com/content/385/bmj-2023-078378

[13] Liu X, et al. "Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension." Nature Communications, 2024. https://www.nature.com/articles/s41467-024-45355-3

[14] Anthropic. "Claude's Model Spec (Constitutional AI Framework)." Anthropic, January 2026. https://www.anthropic.com/constitution

[15] SAS Institute. "SAS Intelligent Decisioning." SAS Software, 2025. https://www.sas.com/en_us/software/intelligent-decisioning.html

[16] SAS Institute. "SAS AI Navigator." SAS Software, 2025. https://www.sas.com/en_in/software/ai-navigator.html

[17] SAS Institute. "SAS Viya AI Governance." SAS Software, 2025. https://www.sas.com/en_us/software/viya/ai-governance.html

[18] SAS Institute. "SAS Score MCP Server Tools Reference." GitHub — sassoftware, 2025. https://github.com/sassoftware/sas-score-mcp-serverjs/blob/main/sas-mcp-tools-reference.md

[19] SAS Institute. "SAS Unveils AI Agents with Customizable Human-AI Interaction for Transparent Decisioning." PR Newswire, SAS Innovate 2025. https://www.prnewswire.com/news-releases/sas-unveils-ai-agents-with-customizable-human-ai-interaction-for-transparent-decisioning-302447658.html

[20] Lewis P, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks in Healthcare." Nature, 2025. https://www.nature.com/articles/s44401-024-00004-1

[21] Wornow M, et al. "Ehrshot: An EHR Benchmark for Few-Shot Evaluation of Foundation Models." JAMIA, 2024;31(9):1994–2005. https://academic.oup.com/jamia/article/31/9/1994/7645319

[22] OpenAI. "OpenAI Launches GPT-5 with Healthcare Focus." HLTH Conference Report, August 2025. https://hlth.com/insights/news/openai-launches-gpt-5-with-healthcare-focus-as-altman-champions-medical-applications-2025-08-08

[23] Flatiron Health. "Assessing Bias in LLM-Extracted Real-World Data: A Health Equity Analysis of Access to Care and Outcomes in Metastatic Breast Cancer." Flatiron Health Publications, 2025.

https://resources.flatiron.com/publications/assessing-bias-in-llm-extracted-real-world-data-a-health-equity-analysis-of-access-to-care-and-outcomes-in-metastatic-breast-cancer

[24] Abramoff MD, et al. "Bias and Fairness in Clinical Artificial Intelligence: A Framework for Evaluation in Healthcare Settings." NPJ Digital Medicine, 2025. https://www.nature.com/articles/s41746-025-01786-w

[25] International Council for Harmonisation. "ICH E9(R1): Statistical Principles for Clinical Trials — Addendum on Estimands and Sensitivity Analysis." ICH, 2019. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf

[26] British Interactive Media Association / BISI. "Claude's New Constitution: AI Alignment, Ethics and the Future of Model Governance." BISI Reports, 2026. https://bisi.org.uk/reports/claudes-new-constitution-ai-alignment-ethics-and-the-future-of-model-governance

[27] Castor. "Automated Evidence Generation and Regulatory-Grade Real-World Data: Avoiding the Benchmark Fallacy." Castor Insight Briefs, 2025. https://www.castoredc.com/insight-briefs/automated-evidence-generation-regulatory-grade-real-world-data/

[28] Wong A, et al. "Large Language Models for Extracting and Summarizing Regulatory Intelligence from Health Authority Guidance Documents." DIA Global Forum, January 2024. https://globalforum.diaglobal.org/issue/january-2024/large-language-models-extracting-and-summarizing-regulatory-intelligence-from-health-authority-guidance-documents/

Correspondence: Sherrine Eid, MPH

Sherrine.Eid@sas.com

www.linkedin.com/in/sherrineeid

www.sas.com