Decoding AI in Clinical Trials:

Practical Applications, Regulatory Reality, and the Road Ahead

phuse

Working Groups

phuse.global

Presenter

Marta Surlej

Agenda

- Practical Applications: reported metrics, realistic deliverables, limitations
- Shadow AI: use patterns and risks
- Regulatory Reality: FDA Framework vs FDA/ EMA Principles
- The Road Ahead
- Round Table discussion
- Q & A

Practical Applications

Reported Metrics, Realistic Deliverables, Limitations

Performance Examples

Concrete AI performance examples

DATA MANAGEMENT

5 min
Manual verbatim coding
per verbatim

→ AI →

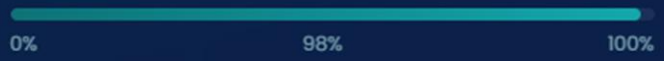
69h
saved per 1,000 verbatims at high-confidence threshold

Medical Coding AI (High-confidence threshold)

98%

SDTM mapping accuracy in AE domain

Zero human involvement required
Tomioka, 2018



PATIENT SCREENING

40%

less screening time for site staff

Accuracy fully maintained across all criteria
TrialGPT · NIH

42.6%

reduction via chain-of-thought LLM prompting

Structured reasoning prompts applied to eligibility criteria
Jin 2024 · Beattie 2024

Concrete AI performance examples

AI COPILOTS · SITE MGMT

50% fewer deviations

20% higher PI satisfaction

Site management AI copilots deployed across trial networks

McKinsey, 2025

SITE OPERATIONS

STUDY BUILD

~73 days

average industry study build time

→ AI reduces errors & accelerates deployment

Tufts CSDD, 2019

CONTRACT & BUDGET

4–6 wks

avg. contract/budget negotiation per site

35% of total startup time

AI contract review cuts this by up to 50%

PAYMENT & RETENTION

40%

of sites drop out due to payment delays

A key risk factor for trial failure

⚠ AI-automated payments directly address retention risk

ACRP, 2019

RBQM Today: Where AI Is Already Working

🗺️ **142 peer-reviewed studies** map AI deployment across clinical trial risk — safety, efficacy, and operational domains

📈 **Operational risk AI** — site selection, phase success prediction, trial termination — is the **fastest-growing category** since 2021

🎯 Phase success prediction models achieve **AUROC up to 92.7%**, trained on **75,000+** protocols from ClinicalTrials.gov

🔍 **ML predicts adverse event reporting rates** in centralized monitoring — validated in a clinical QA context (Ménard et al., 2019)

⚙️ **Traditional ML** (random forest, XGBoost) remains the workhorse — reliable, explainable, validated. **Not everything needs an LLM**

AI-Assisted Medical Data Cleaning: The Numbers Are In

- Controlled study (n=10 experienced medical reviewers, Phase III oncology data) compared AI-assisted vs. traditional spreadsheet review
- Every single participant improved on both speed and accuracy simultaneously

What the $5.1M breaks down to (Phase III oncology, 1,100 patients):

- ✓ **$4.4M** from 5-day earlier database lock ($840K/day in opportunity cost)
- ✓ **$420K** from reduced medical review hours
- ✓ **$288K** from query management savings

Eligibility Screening at Scale – RECTIFIER study

PATIENT IDENTIFICATION

Eligible Patients Missed by Manual Review → Found by AI

Manual screening is time-intensive and error-prone. Coordinators working through long eligibility criteria lists can miss borderline-eligible patients — reducing enrolment potential.

- Manual review: eligibility criteria applied by staff
- ↓
- Borderline or complex cases overlooked under time pressure
- ↓
- ✦ AI surfaces missed candidates — expanding enrolment reach

SCREENING EFFICIENCY

Ineligible Patients → Removed Faster

Screening ineligible patients wastes coordinator time, delays recruitment, and burdens sites. AI pre-filters records against exclusion criteria before staff review begins.

- Full patient list reviewed manually: slow & resource-heavy
- ↓
- AI eliminates clear non-candidates in seconds
- ↓
- ✦ Staff focus only on genuinely eligible candidates

HUMAN-IN-THE-LOOP

Investigator Retains Final Enrolment Decision

AI does not enrol patients — it supports the decision. The investigator reviews AI-ranked candidates and makes the final call, preserving regulatory accountability and clinical judgement.

- AI ranks & shortlists — does not decide
- ↓
- Investigator reviews AI output with full context
- ↓
- ✦ Final enrolment decision: always the investigator

● LOW-RISK AI APPLICATION

Safety Risk: From Reactive Review to Predictive Detection

- 55 of 142 reviewed studies focus on safety risk - the most mature AI application in clinical trials
- ADE prediction models achieve AUROC up to 96.6% on validated benchmarks
- Graph Neural Networks map drug-gene-protein-ADE relationships that no human reviewer can hold in working memory simultaneously
- DILI prediction - 16 dedicated studies – (#1 cause of post-market withdrawal)
- AI contextual intelligence can flag severity miscoding, timing mismatches between AE onset and concomitant medication start, missing supporting lab data

Real example of what AI catches that humans miss:

A patient narrative reads "severe nausea requiring hospitalization." The AE is coded as severity "mild." The AI cross-references the narrative text, the severity coding, and the hospitalization record simultaneously and flags the contradiction. In a 30-minute manual review, human reviewer processing dozens of patient records may never see it. (Purri et al., 2025)

Limitation to acknowledge openly:

The best ADE prediction model in the field outperforms a naive baseline by only 2 percentage points on AUROC (Zhong et al., 2024). Selection bias and imbalanced datasets remain real problems. AI safety tools should be positioned as a "second reader" layer, not a replacement for medical review.

Automated Endpoint Adjudication — The INVESTED Trial

- ❑ **NLP Adjudication** of HF Hospitalization (INVESTED Trial):
- ✓ Model trained at single center (Mass General Brigham)
- ✓ Externally validated in multicenter US + Canada trial
- ✓ Agreement with CEC: 87% on HF hospitalization
- ✓ Fine-tuning within INVESTED → **performance equal to human reviewers**

	Manual Adjudication	AI Adjudication
Speed	⊗ Weeks of delay	✓ Real-time signals
Cost	⊗ High cost	✓ Scalable cost
Consistency	⊗ Individual MD decisions may not be uniform	✓ Consistent criteria
Scalability	⊗ Scales poorly – impractical for large trials	✓ Scales to any trial size at marginal cost

LLMs used by Biostatistics Teams

Survey: 208 Biostatisticians, Duke + Stanford, Sept-Oct'24 (Grambow et al., 2025)

- **63.8%** use LLMs in their work
- **46.5%** use them daily; 37.2% weekly

What they use them for:

- Writing, debugging, documenting code (R/Python/SAS) - **77.5%**
- Editing and improving writing quality - **76.3%**
- Explaining statistical methods to non-experts - **57.5%**

The error problem is real:

- 70.7% encountered errors with significant potential consequences
- Incorrect code generation, statistical misinterpretation, content fabrication

Shadow AI

Use Patterns and Risks

Regulatory Compliance

phuse

phuse.global

Working Groups

‘Shadow AI’ –The Elephant in the Room

The scale:

⇒ **57%** of respondents had encountered or used an unauthorized AI tools

⇒ **50%** did it to work faster (**37%** of providers**, 60%** of admin staff use it frequently)

⇒ **17%** rely on AI on for most aspects of work vs **7%** providers which includes direct patient care)

⇒ **~ 33** pointed to either a lack of approved tools or the approved tools lacking the desired functionality

(Source: Wolters Kluwer Health, "Shadow AI: A Hidden Risk to Healthcare," January 2026, based on 518 HC workers)

- **Risk perception gap: 33%** rank **regulatory compliance as a top concern**
- **Policy gap: 27%** of providers **unsure or unaware of policies;**
- **“Trust" problem: 59%** judge AI by personal review - **"I tried it and it seemed fine"**

Shadow AI refers to employees using AI tools that haven't been authorized or vetted by their organizations.
In clinical trials and healthcare more broadly, this means researchers, clinicians, CRAs, or site staff using tools like ChatGPT or other generative AI for work-related tasks

‘Shadow AI’ Risk Spectrum in Clinical Trials

Lower risk (but still ungoverned):

- Using AI to proofread an internal email to a colleague
- Asking AI to explain a statistical concept for personal learning
- Using AI to draft an internal meeting agenda

Medium risk (entering the grey zone):

- Using AI to draft SOPs or work instructions
- Summarizing regulatory guidance documents with AI
- Using AI to prepare training materials for site staff

- **Higher risk (directly impacts regulated outputs):**
- Drafting monitoring visit reports, query text, or deviation narratives with AI
- Using AI to inform medical judgments (causality, eligibility, risk assessments)
- Creating or modifying patient-facing documents (ICFs, patient letters)
- Writing safety narratives or DSUR sections with AI assistance

Highest risk:

- Pasting identifiable patient data into unathorized AI tools
- Using AI-generated content in regulatory submissions without disclosure
- Relying on AI for endpoint adjudication or data-driven enrolment or dose-modification decisions
- Uploading unblinded treatment allocation data into a general-purpose AI tool

Regulatory Reality

FDA 7-Step Framework vs. EMA/FDA 10 Guiding Principles

Regulatory Compliance

Core Alignment: risk-based approaches , context of use definition , and lifecycle management

Framework and Joint Principles with 12 Months

FDA 7-Step Framework

January 2025 Draft Guidance

Purpose

Establish and evaluate AI model credibility for regulatory decision-making

Scope

Drug and biological product development across all lifecycle phases

Approach

Risk-based credibility assessment with proportionate validation

Foundation

Based on **800+ public comments** and **500+ submissions** with **AI components** since 2016

"With appropriate safeguards, AI has transformative potential to advance clinical research and accelerate medical product development."
– FDA Commissioner Robert Califf

EMA/FDA 10 Principles

January 2026 Joint Document

Purpose

Foundational principles for good AI practice across drug product lifecycle

Scope

Nonclinical, clinical, post-marketing, and manufacturing phases

Approach

Ethical and operational foundation for AI implementation

Foundation

International harmonization and consensus standards development

"These principles lay the foundation for developing good practice that addresses the unique nature of AI technologies." – EMA/FDA Joint Statement

FDA FRAMEWORK

FDA 7-Step Credibility Assessment Framework

1 Define Question of Interest

Articulate the specific regulatory question, decision, or concern the AI model addresses

2 Define Context of Use (COU)

Describe model's scope, role, data inputs/outputs, and workflow integration

3 Assess AI Model Risk

Evaluate based on model influence and decision consequence

4 Develop Credibility Assessment Plan

Outline model description, data sources, validation strategy, and acceptance criteria

5 Execute the Plan

Conduct model training, validation experiments, performance testing, and remediation

6 Document Results & Deviations

Compile credibility assessment report with performance data and rationales

7 Determine Model Adequacy

Evaluate if credibility criteria are met; implement mitigations if needed

Key Characteristics

Iterative & Cyclical

Continuous refinement and feedback loops throughout the process

Risk-Proportionate

Validation rigor matched to assessed risk level

Early FDA Engagement

Interactive feedback on risk and COU recommended

Documentation-Heavy

Comprehensive records for regulatory submissions

EMA/FDA PRINCIPLES

EMA/FDA 10 Guiding Principles Overview

1

Ethics

Human-Centric by Design

AI development aligns with ethical and human-centric values

2

Risk

Risk-Based Approach

Proportionate validation, risk mitigation, and oversight based on COU and model risk

3

Risk

Adherence to Standards

Compliance with legal, ethical, technical, scientific, and GxP standards

4

Risk

Clear Context of Use

Well-defined role and scope for why AI is being used

5

Dev

Multidisciplinary Expertise

Integration of AI technology and domain expertise throughout lifecycle

6

Dev

Data Governance & Documentation

Detailed, traceable, verifiable documentation of data provenance and processing

7

Dev

Model Design & Development

Best practices in model/system design with fit-for-use data, interpretability, and explainability

8

Perf

Risk-Based Performance Assessment

Complete system evaluation including human-AI interactions with fit-for-use metrics

9

Perf

Life Cycle Management

Risk-based quality management with scheduled monitoring and periodic re-evaluation

10

Comm

Clear, Essential Information

Plain language communication of COU, performance, limitations, and interpretability

FRAMEWORK ALIGNMENT

Direct Alignment: Both require precise definition of **AI's role and scope**.

Steps 1-2: Question Definition & COU Mapping

1 Define Question of Interest

FDA Step 1

- Specific regulatory question AI addresses
- Decision or concern being influenced
- Examples: patient risk stratification, QC pass/fail

2 Define Context of Use (COU)

FDA Step 2

- Model's scope, role, and operational boundaries
- Data inputs/outputs and integration workflow
- Whether AI informs or makes decisions

4 Clear Context of Use

EMA/FDA Principle 4

> "AI technologies have a well-defined context of use (role and scope for why it is being used)."

10 Clear, Essential Information

EMA/FDA Principle 10

> "Plain language is used to present clear, accessible, and contextually relevant information to the intended audience, including users and patients, regarding the AI technology's context of use, performance, limitations, underlying data, updates, and interpretability or explainability."

FRAMEWORK ALIGNMENT

Step 3: Risk Assessment Alignment

FDA Risk Matrix

Consequence ↓ \ Influence →	Low Influence	High Influence
	LOW Basic validation	**MEDIUM** Standard validation
High Consequence	**MEDIUM-HIGH** Moderate validation	**HIGH** **Extensive validation**

Model Influence
Degree of AI autonomy in decision

Decision Consequence
Impact of AI errors

2 Risk-Based Approach

EMA/FDA Guiding Principle

- → **Proportionate Validation**
  Validation rigor matched to assessed risk level
- → **Risk Mitigation**
  Targeted strategies to address identified risks
- → **Context-Dependent**
  Risk assessment based on COU and model characteristics
- → **Oversight Scaling**
  Regulatory oversight proportionate to risk

HIGH RISK Example

AI predicts serious adverse reactions; wrong prediction → patient harm

Requires extensive validation, continuous monitoring

LOW RISK Example

AI suggests vial fill volume; batch testing confirms

Basic validation sufficient

FRAMEWORK ALIGNMENT

Step 4: Credibility Planning & Development Practices

FDA Credibility Plan Components

A Model Description

- Inputs, outputs, architecture, features
- Loss functions, parameters, modeling approach
- Feature selection process and rationale

B Model Development Data

- Training data and tuning data characterization
- Fit-for-use data quality criteria
- Data management practices

C Validation Strategy

- Performance metrics and acceptance criteria
- Risk-appropriate testing methods
- Bias mitigation protocols

Early FDA Engagement
Interactive feedback on risk and COU recommended

EMA/FDA Principle Integration

3 Adherence to Standards

GxP compliance, legal, ethical, tech, cybersecurity standards

5 Multidisciplinary Expertise

Integration of AI and domain expertise throughout lifecycle

6 Data Governance & Documentation

Detailed, traceable, verifiable documentation of data provenance and processing

7 Model Design & Development

Best practices with fit-for-use data, interpretability, explainability, and predictive performance

FRAMEWORK ALIGNMENT

Steps 5-6: Execution, Validation & Documentation

5 Execute the Plan

Model Training & Tuning
Execute training protocols, hyperparameter optimization

Validation Experiments
Run test scenarios, edge cases, stress tests

Performance Metrics
Calculate accuracy, sensitivity, calibration, bias metrics

Error Remediation
Address identified issues, retrain if necessary

6 Document Results & Deviations

Credibility Assessment Report
- Performance tables, plots, and narrative
- Comparison to planned criteria
- Any deviations and their rationale

6 Data Governance & Documentation

EMA/FDA Guiding Principle

Detailed, traceable, verifiable documentation of data source provenance and processing steps. Analytical decisions in line with GxP requirements.

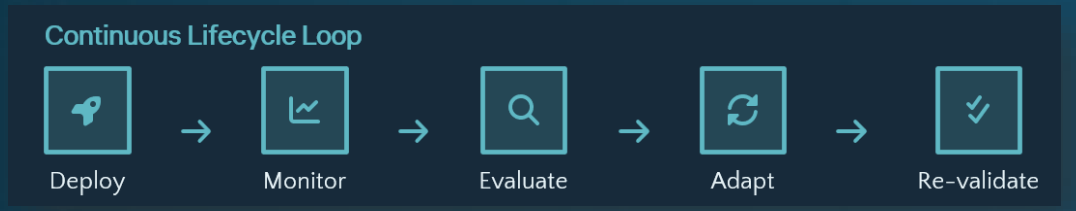
8 Risk-Based Performance Assessment

EMA/FDA Guiding Principle

- Use fit-for-use data and metrics appropriate for COU
- Evaluate complete system including human-AI interactions
- Validate predictive performance through appropriately designed testing
- Support with evaluation methods matched to risk level

FRAMEWORK ALIGNMENT

Step 7 & Principle 9: Lifecycle Management



7 Determine Model Adequacy

Adequacy Decision

- Evaluate if credibility criteria are met
- Compare performance against acceptance thresholds
- Assess residual risks and mitigations

If Not Adequate

- Implement additional mitigations
- Tighten validation requirements
- Integrate human oversight
- Revise model or COU

Iterative Nature

Continuous refinement and feedback loops throughout the process

9 Life Cycle Management

EMA/FDA Guiding Principle

Risk-Based Quality Management

Systems implemented throughout AI lifecycle

Scheduled Monitoring

Regular performance checks

Issue Management

Capture, assess, and address issues

Periodic Re-evaluation

Address data drift, ensure performance

FOUNDATIONAL PRINCIPLES

Human-Centric Design & Standards Adherence

1 Human-Centric by Design

EMA/FDA Guiding Principle

> "AI development and use align with ethical and human-centric values."

Implicit in FDA Framework

- Human oversight emphasized throughout 7 steps
- Model influence assessment (Step 3) considers human-AI interaction
- Interpretability and explainability requirements (Step 4)
- Patient safety as ultimate decision criterion (Step 7)

Human-AI Interaction Points

COU definition (human review)	Risk assessment (oversight level)
Validation (human evaluation)	Adequacy decision (final authority)

3 Adherence to Standards

EMA/FDA Guiding Principle

> "AI technologies adhere to relevant legal, ethical, technical, scientific, cybersecurity, and regulatory standards, including Good Practices (GxP)."

Embedded in FDA Framework

- GxP requirements in documentation (Step 6)
- Data governance and quality standards (Step 4)
- Model development best practices (Step 4)
- Performance assessment standards (Step 5)

Standards Integration

Legal	Ethical	Technical
Scientific	Cybersecurity	GxP

⚠ Framework and Principles are complementary, not alternative. Apply both simultaneously from project inception for regulatory success.

Integrated Framework: Practical Implementation

Unified Implementation Workflow

1-2 — **Define Question & COU**
Aligns with Principles 4, 10 (Clear COU, Communication)

↓

3 — **Assess Risk**
Aligns with Principle 2 (Risk-Based Approach)

↓

4 — **Develop Credibility Plan**
Aligns with Principles 3, 5, 6, 7 (Standards, Expertise, Data, Design)

↓

5-6 — **Execute & Document**
Aligns with Principles 6, 8 (Data Governance, Performance Assessment)

↓

7 — **Determine Adequacy + Lifecycle**
Aligns with Principle 9 (Life Cycle Management)

Key Success Factors

Early FDA Engagement
Interactive feedback on risk and COU

Multidisciplinary Teams
AI + domain expertise integration

Fit-for-Use Data
Quality, representativeness, governance

Risk-Proportionate Validation
Validation rigor matched to risk

Continuous Monitoring
Lifecycle management and re-evaluation

Regulatory Engagement Pathways

Clinical Trial Design (before IND submission)

→ C3TI — Center for Clinical Trial Innovation

Novel or complex clinical trial designs

→ CID — Complex Innovative Trial Design Meeting Program

Qualifying AI-based algorithms (biomarkers, endpoints, patient evaluation)

→ DDTs / ISTAND Pilot Program

AI-enabled digital health technologies (wearables, sensors) in drug development

→ DHTs Program

Model-informed drug development (e.g., PK/PD, exposure-response)

→ MIDD Paired Meeting Program

Real-world data sources to support clinical evidence

→ RWE Program

Road Ahead

What's Coming in the Nearest Future

phuse Working Groups

phuse.global

Looking Ahead: Realistic 3–5 Year Advancements

- More mature integration of **ML-driven CM, data analytics** into RBQM workflows
- **Agentic AI** entering operational workflows: autonomous planning and coordination across feasibility, site ID, operational planning, and risk management
- Improved **NLP interfaces** for clinical data exploration, reducing reliance on specialized tools and democratizing data access
- Better-defined **industry consensus on AI validation** in clinical settings
- Increased **regulatory clarity** as the EMA/FDA guiding principles mature into official guidance
- **Parallel workflow execution** through AI coordination of traditionally sequential trial conduct tasks

NOTE: Fully autonomous AI decision-making in trial conduct remains unlikely and inadvisable

Round Table Discussion Contributors

Name	Organization Type
Jeremy Wildfire and **Jenn Krohn**	Pharma
Madhumita Ghosh	Biotech
Neha Agarwal	CRO

Round-Table Discussion Topics

⇒ Where should the line be drawn between ‘operational efficiency AI’ and ‘decision-influencing AI?

⇒ What is a realistic approach to ‘Shadow AI’ use that could reduce risk without killing teams’ productivity?