Plain Language Summary: An Agentic AI Fidelity Evaluation Framework for Real-World Evidence Workflows

John D. Diaz-Decaro, PhD, MS
Black Swan Causal Labs

Abstract 116148

Background

- AI systems are beginning to support real-world evidence work, including protocol writing, code generation, analysis checks, and reporting.
- Current AI evaluation often focuses on whether the system functions technically, but not whether it produces correct real-world evidence methods.

Objective

- This poster proposes a way to evaluate agentic AI workflows in a manner aligned with emerging FDA expectations for AI credibility.
- Test two specific AI subagents:
  - a protocol peer reviewer that checks for study design and protocol errors;
  - a code inspector that checks for coding and analytic errors.

Methods

- A mock multi-agent real-world evidence workflow was developed, covering data feasibility, protocol development, protocol review, code generation, code inspection, analysis execution, and reporting.
- Human review was included at several points in the workflow to support oversight and decision-making.
- The study inserted 18 protocol errors and 18 code errors, then tested whether the AI subagents detected them.
- The experiments were repeated using two Claude models (Sonnet 4.6 and Opus 4.7) to assess consistency across models.

Results

- The protocol peer review subagent detected all 18 intentionally inserted protocol errors using both Claude models. It also passed the negative control test, meaning it did not flag errors when reviewing the clean protocol.
- The code inspector detected 17 of 18 intentionally inserted code errors using both Claude models.
  - The one missed coding error involved a switch between two related survival analysis models, which the code inspector treated as the same type of issue.
  - On clean code, the code inspector was not completely silent, but its flags were not invented defects.
- Opus 4.7 appeared somewhat more reliable than Sonnet 4.6 in the clean-code test.

Discussion

- Agentic AI workflows for real-world evidence can be designed with validation checks embedded at each step.
- The findings suggest that protocol review and code inspection subagents may be useful as quality-control tools in AI-assisted real-world evidence workflows.
- A key takeaway is that multi-agent AI systems should not be evaluated only as a single end-to-end system. Each subagent should be tested separately because each workflow step carries different methodological risks.
- This is a promising proof of concept, but not yet a full validation of the entire workflow.