J&J Go/No-Go: AI-Powered Clinical Trial Decision Engine

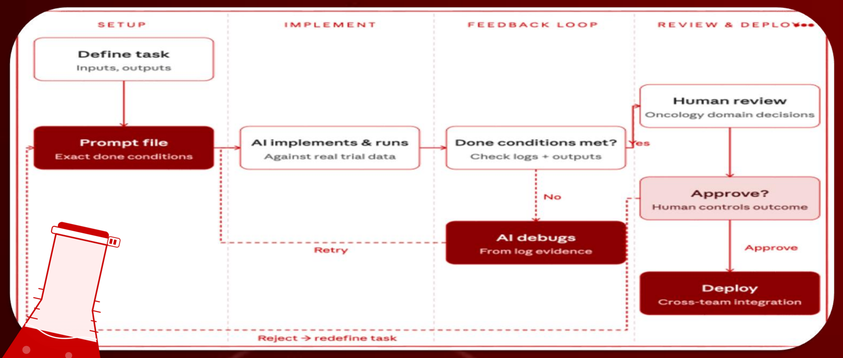
Names: Ryan Sierra, Bisti Potdar, Shaivi Mishra, Shilp Shah, Shivam Trivedi, Tanya Jain
Mentor: Fei Chen
TA: Tanisha Agarwal
Fall '25 Students: Shaivi Mishra, Shilp Shah, Hasan Sadiq, Mahrukh Muhaimeen, Kisal Wijesooriya, Yana Gomon, Daivin Desai

Overview

Last semester, we focused on establishing the fundamental structure of our methodology. Through continuous experimentation, we built a basic Go/No-Go application using simple components that served as a foundation for further development this semester.

The Objective

The Cost of Failure: Phase 3 trials cost $300M-$2B, take 3-5 years, and roughly 50% fail. Failures often happen because the Phase 2 signal was misread.
The Human Limit: The decision to advance is currently made by human reviewers without a consistent, quantitative framework. It is difficult to manually synthesize many variables across efficacy, toxicity, trial design, and patient baselines at scale.
Our Goal: Given a completed Phase 2 trial's patient-level data, predict the probability of Phase 3 success. Output a calibrated, explainable GO/NO-GO recommendation grounded in 93 historical oncology trials

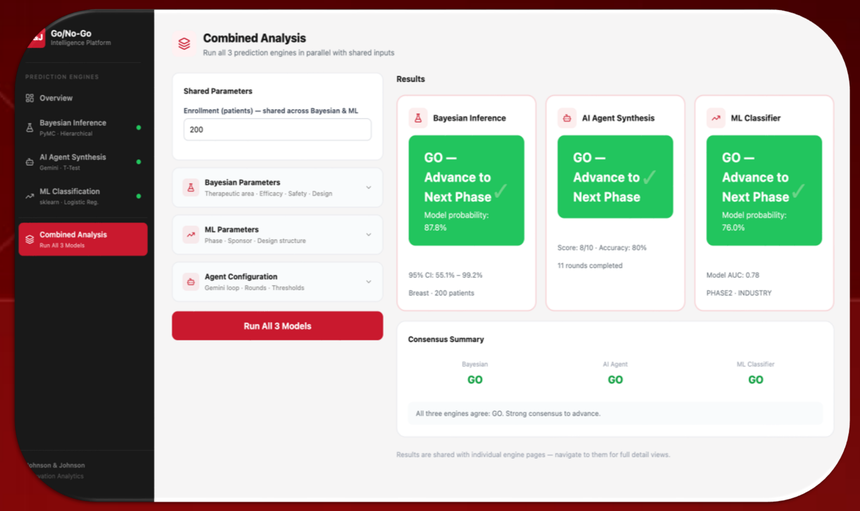


Methodology(AI workflow)

Checkpoint-Driven Execution:
- Built alongside agentic AI (Claude Code)
- Development guided by prompt files
- Exact inputs, outputs, and "done conditions" defined to prevent AI drift.

Iterative Feedback Loop:
- AI implements and runs code against real clinical trial data.
- If done conditions not met, AI debugs directly from log evidence
- Iterations continue until the criteria are satisfied.

Human-AI Synergy:
- Allows AI accelerated implementation, debugging, and cross-team integration
- Humans control over all oncology domain decisions.
- Engineering bottleneck shifts from writing code to precisely defining the problem.

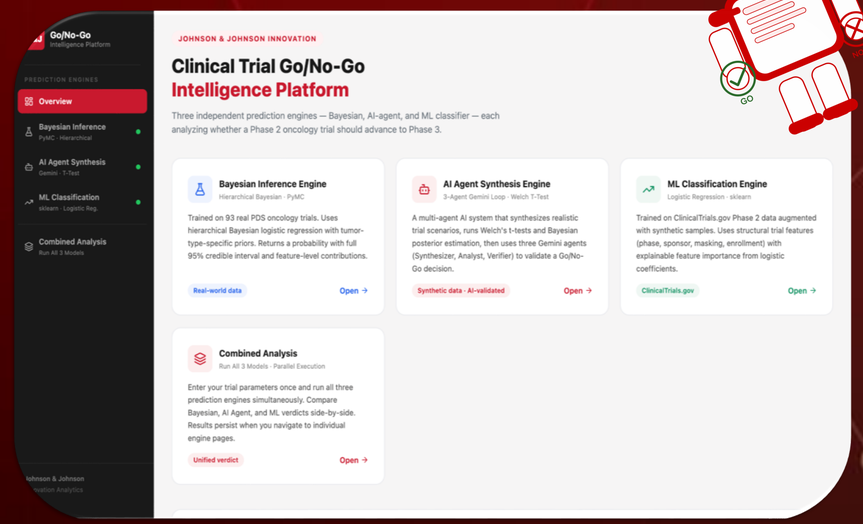


Architecture

Bayesian Inference Engine: Trained on 93 oncology trials using Hierarchical Bayesian Logistic Regression. It incorporates clinical priors and outputs probabilities with credible intervals to quantify uncertainty.

Multi-Agent Synthesis Engine: A Gemini-powered agent loop evaluates trial scenarios. The AI interprets deterministic statistics rather than generating them, ensuring hallucination-free, reproducible recommendations.

Classical ML Classifier: Ingests public trial metadata, using LLMs to categorize ambiguous outcomes. It applies rigorous feature selection and synthetic data augmentation to deliver explainable predictions.



Results

- The final output of the AI based workflow: (web application shown in the images) to house the models created
- Allows users to input clinical trial results and uses the three models to determine whether to go/no go on the trial
- Allows for parameter adjustment based on user preference
- Provides detailed explanations of how the models work as well as a simple, interactive layout

Conclusion

Performance: The Bayesian model showed strong held-out performance (AUC 0.799, Accuracy 79.2%, Brier Score 0.184) and achieved perfect MCMC convergence (R-hat 1.0008, zero divergences).

Predictive Features: Key drivers were counterintuitive. Small molecule modality (beta 3.05) and molecular weight (beta 1.94) were the strongest predictors of success, while ORR endpoints had a negative coefficient (beta -1.17).

Dataset Balance: The 93-trial dataset was balanced with 41 GO (44%) and 52 NO-GO (56%) decisions, preventing common clinical data imbalances.

Next Steps & Future Goals

- **Expand the Trial Corpus:** Combine the 93 internal PDS trials with the 100 trials recently ingested from the ClinicalTrials.gov API to double the training set.
- **Automate Live Ingestion:** Schedule the existing API client to automatically ingest newly completed trials, creating a self-updating prediction engine.
- **Expand Beyond Oncology:** Utilize the existing hierarchical architecture to ingest trials from new therapeutic areas (e.g., Immunology).
- **Validate the Ensemble:** Run a formal ablation study (Bayesian vs. ML vs. Ensemble) on a held-out test set to quantify the exact performance lift of the combined approach.