# Assessing the Usability of GutGPT: A Simulation Study of an AI Clinical Decision Support System for Gastrointestinal Bleeding Risk

Colleen Chan[1], Kisung You[2], Sunny Chung[3], Mauro Giuffrè[3], Theo Saarinen[4], Niroop Rajashekar[3], Yuan Pu[3], Yeo Eun Shin[4], Loren Laine[3], Ambrose Wong[3], Leigh Evans[3], Allan Hsiao[3], René Kizilcec[5], Jasjeet Sekhon[1], Dennis Shung[3]

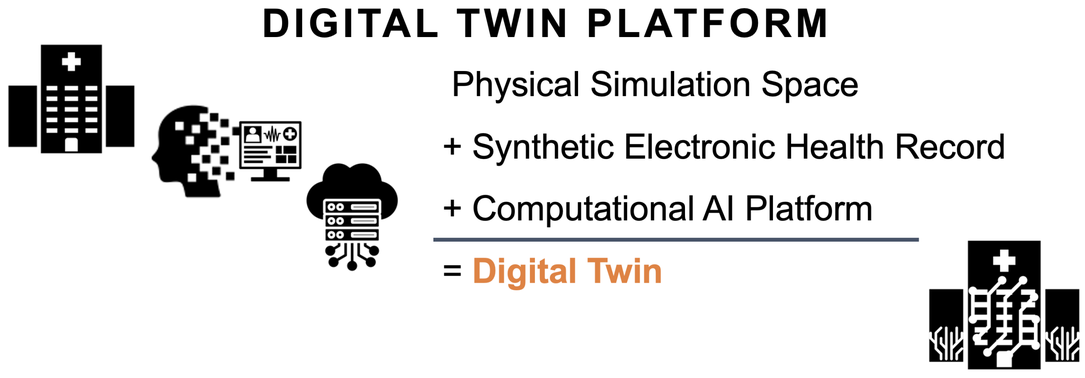
[1]Yale University, [2]CUNY Baruch College, [3]Yale School of Medicine, [4]University of California, Berkeley, [5]Cornell University

## Introduction

Machine learning (ML)-based risk stratification systems in upper gastrointestinal bleeding (UGIB) have been shown to outperform existing clinical risk scores. However, successful implementation of such systems into practice requires acceptance and trust of the systems by clinicians.

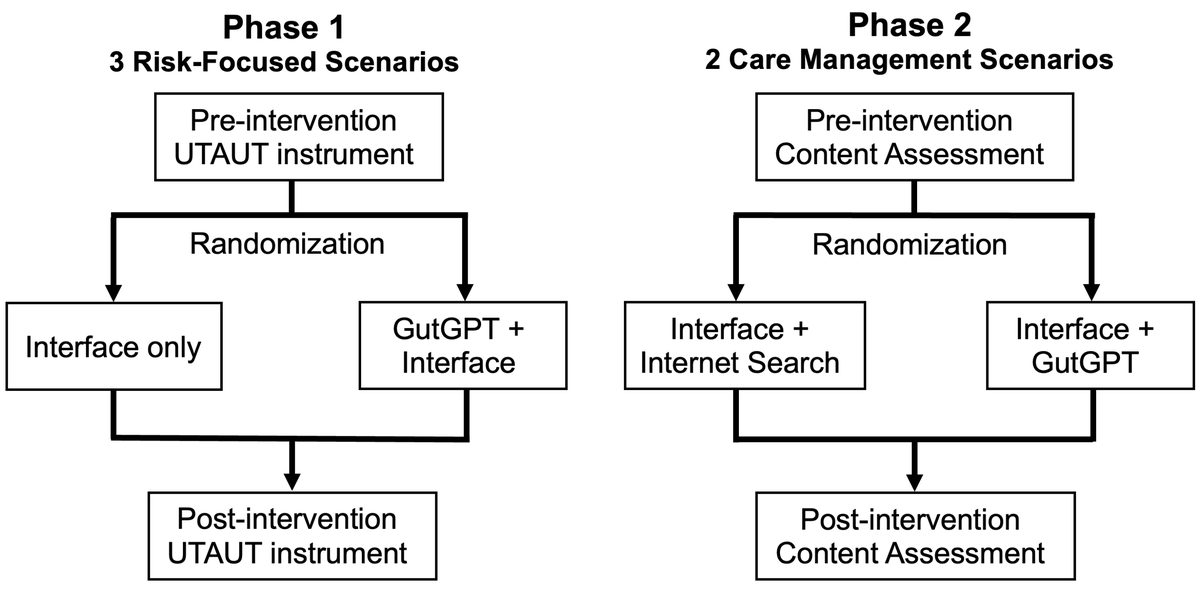
We built an interactive dashboard interface to explain ML risk predictions for UGIB. We also developed GutGPT, a large language model (LLM)-enhanced AI clinical decision support system (AI-CDSS) to better communicate output from our ML model and provide clinical management recommendations based on UGIB guidelines. We conducted a randomized controlled trial to test the effect of the dashboard with GutGPT on physician trust and acceptance compared to the dashboard alone using proctored scenarios in a digital twin setup at the Yale Center for Healthcare Simulation.

## Study Design




**Participants:** Emergency and internal medicine physicians and medical students, grouped into small teams of two to four for simulation scenarios using a high-fidelity mannequin and a playground version of the Epic electronic health record (EHR) system.

**Intervention:** Our study utilizes two main tools: GutGPT, an AI chatbot, and an interactive dashboard, both powered by a validated ML model for predicting GIB risk of a hospital-based intervention or 30-day mortality. GutGPT allows natural language interaction for guideline queries and risk prediction, while the dashboard enables risk prediction through adjustable patient covariates without natural language capabilities and displays interpretability plots.



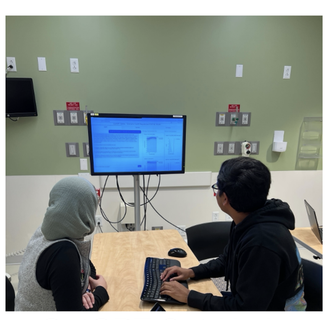

Our study comprised two phases:

**1. Trust and Acceptability Assessment**: Teams were randomized to use either GutGPT with the dashboard or the dashboard alone for risk assessment in patient scenarios. Surveys adapted from two established instruments, the Unified Theory of Acceptance and Use of Technology (UTAUT) and the System Usability Survey (SUS), were administered pre- and post-simulation.

**2. Knowledge of Clinical Management**: Teams were re-randomized to use GutGPT with the dashboard or the dashboard with additional online resources for managing GIB cases. Their decision-making and management skills were evaluated pre- and post-simulation.

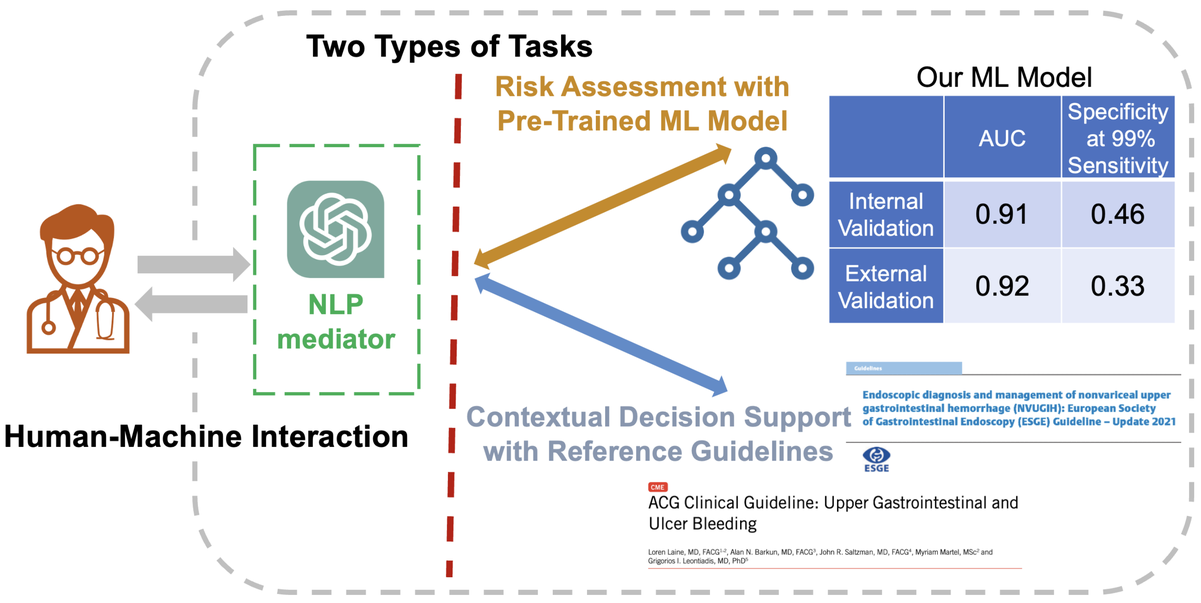
Control Room

Simulation Room Mannequin



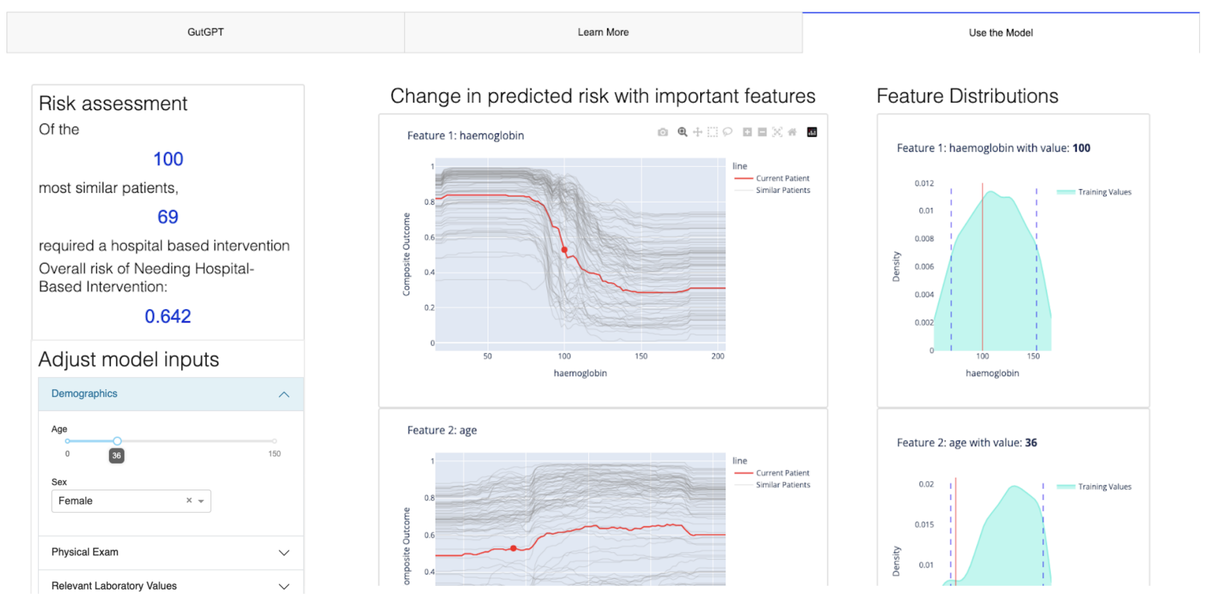
GutGPT Set-Up in Simulation Room

## GutGPT



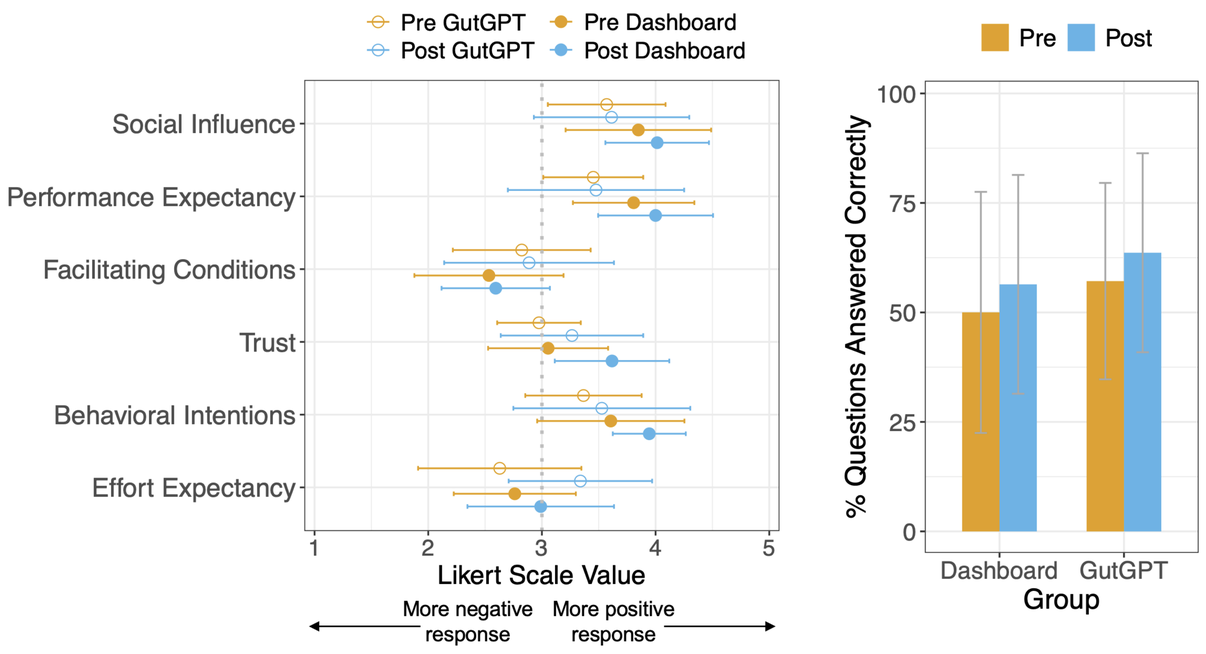

GutGPT, using OpenAI's GPT-3.5 Turbo API, processes user queries through in-context learning with a classifier LLM that directs queries to two main models: a Model LLM for GIB risk prediction and feature importance and a Guidelines LLM for medical guideline queries. Each uses a specialized model, with the Model LLM accessing our machine learning model, and the Guidelines LLM performing context extraction on a UGIB guidelines document. If a query falls into both categories, a final LLM synthesizes the responses from each model.

## Interactive Dashboard




## Results

Preliminary results from 55 participants suggested exposure to either GutGPT or the interactive dashboard maintained positive perceptions of Trust, Behavioral Intentions, Social Influence, and Performance Expectancy. Interestingly, Effort Expectancy, which corresponds to perceived ease of use, particularly increased for GutGPT arm participants.




## Discussion

We demonstrate the feasibility of medical simulation and a digital twin environment with an EHR and computational infrastructure to deploy generative AI for safety testing AI-CDSS with or without LLMs. Our study suggests that physicians have positive feelings with regards to trust and acceptance towards AI that persist or slightly improve after exposure to AI-CDSS in simulation scenarios. Deployment of LLMs in clinical processes may benefit from similar usability testing that evaluate human-algorithmic interaction.