Statistical Analysis Plan (SAP)

Title: Evaluating Generative AI as a Foundational Technology: Results from a Mixed Methods Randomized Controlled Trial

1. Introduction

This Statistical Analysis Plan (SAP) outlines the prespecified statistical methods for analyzing data collected during the randomized controlled trial (RCT) evaluating the GutGPT clinical decision support system (CDSS). This trial compares a generative AI-enhanced dashboard (GutGPT+dashboard) to a standard machine learning dashboard (dashboard) in a high-fidelity simulation environment.

2. Objectives and Hypotheses

Primary Objectives:

1. To evaluate the impact of GutGPT on clinicians' Behavioral Intention (BI) to adopt the technology, as measured by the Unified Theory of Acceptance and Use of Technology (UTAUT) framework.
2. To assess the clinical decision accuracy of GutGPT compared to the standard dashboard.

Secondary Objectives:

1. To assess the impact of GutGPT on secondary UTAUT constructs, including Effort Expectancy (EE), Performance Expectancy (PE), Social Influence (SI), and Facilitating Conditions (FC).

Primary Hypotheses:

1. Clinicians using GutGPT will demonstrate a greater improvement in Behavioral Intention (BI) compared to those using the standard dashboard.
2. Clinicians using GutGPT will achieve similar decision accuracy compared to the dashboard arm.

Secondary Hypotheses:

1. Secondary outcomes from UTAUT will also improve more for participants in the GutGPT arm compared to the dashboard arm

3. Study Design Overview

This is a two-arm, parallel-group, randomized controlled trial conducted at a high-fidelity simulation center. Participants include internal medicine residents, emergency medicine residents, and medical students at Yale School of Medicine. Each participant was randomized to either the GutGPT+dashboard arm or the dashboard arm. Participants were block randomized, and due to visible system differences, blinding was infeasible. Standardized proctor scripts were used to minimize bias.

4. Analysis Populations

All randomized participants who completed both pre- and post-intervention surveys will be included in the primary and secondary analyses.

5. Sample Size and Power Calculations

A total sample size of 104 participants (52 per arm) was calculated to achieve 80% power to detect a clinically relevant medium effect size (Cohen's d = 0.5) with a two-tailed significance level of 0.05 for the primary outcome, Behavioral Intention (BI). Decision accuracy findings were considered exploratory due to the study's powering for BI.

6. Outcomes

Primary Outcome:

- Change in Behavioral Intention (BI) from pre- to post-intervention, measured on a 5-point Likert scale.
- Clinical decision accuracy, defined as the proportion of scenarios in which participants made guideline-concordant decisions.

Exploratory Outcomes:

1. Change in Effort Expectancy (EE), Performance Expectancy (PE), Social Influence (SI), and Facilitating Conditions (FC).
2. Change in Trust, Perceived Risk, and Perceived Benefit.

7. Statistical Analysis

7.3 Primary Outcome: Change in Behavioral Intention (BI)

1. **Statistical Test:** Two-tailed Mann-Whitney U test to compare median changes in BI between the GutGPT+dashboard and dashboard arms.
2. **Effect Size:** Median difference and 95% confidence intervals calculated using bootstrap resampling with 1,000 iterations.
3. **Sensitivity Analysis:** Proportion of participants achieving a ≥1-point increase on the Likert scale.

7.4 Secondary Analyses

Clinical Decision Accuracy

1. **Statistical Test:** Independent samples t-test to compare mean decision accuracy between arms.
2. **Scenario-Specific Accuracy:** Chi-squared tests will be used to compare accuracy for low-, medium-, and high-risk cases. These findings will be considered exploratory.

UTAUT Constructs (EE, PE, SI, FC)

1. Mann-Whitney U test for between-arm comparisons of median changes.
2. Within-arm comparisons using Wilcoxon signed-rank tests to evaluate pre-to-post changes.
3. Proportion of participants achieving a ≥1-point change will be reported with chi-squared tests for between-group differences.

8. Software

All analyses will be conducted using Python (version 3.11), with the following packages:

- scipy: Statistical tests
- pandas: Data management
- sklearn: Inter-coder reliability and model validation