Safety and Efficacy Study of AI LVEF (EchoNet-RCT)

Statistical Analysis Plan

Principal Investigator:	David Ouyang, MD Staff Physician, Cedars-Sinai Medical Center Cedars-Sinai Medical Center
ClinicalTrials.gov identifier:	NCT05140642
Author:	Bryan He, BSc Computer Science PhD Candidate Stanford University

Table of Contents

Introduction

Recent advances in machine learning and image processing techniques have shown that machine learning models can identify features unrecognized by human experts and more precisely/accurately assess common measurements made in clinical practice. In echocardiography, this ability for precision measurement and detection is important in both disease screening as well as diagnosis of cardiovascular disease.

Echocardiography is routinely and frequently used for diagnosis and prognostication in routine clinical care, however there is often subjectivity in interpretation, heterogeneity in application, and variance with image acquisition and quality. The cardiac function, as described by the left ventricular ejection fraction (LVEF), is the focal measurement of echocardiography, and used to diagnose heart failure and determine various interventions and medical treatments.

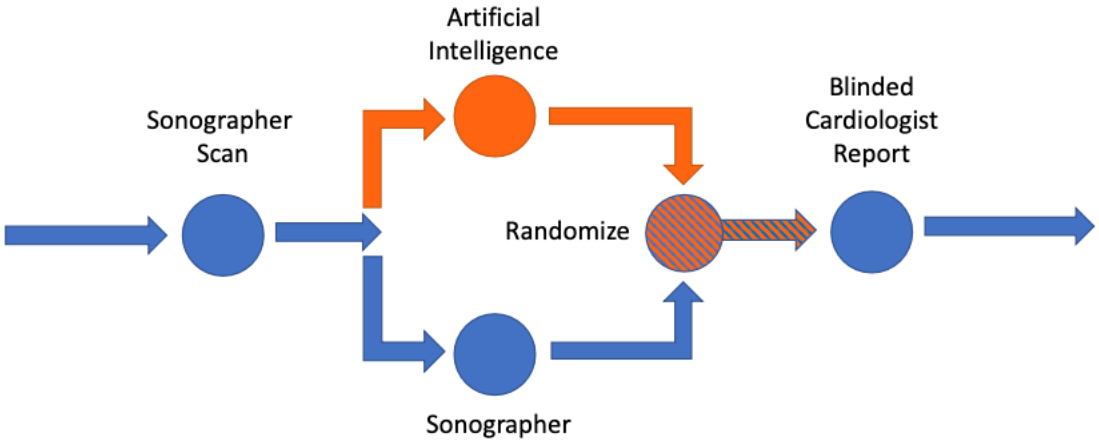
Echocardiography is the perfect place to study the impact of AI models, given the step-wise, time/location staggered evaluation by multiple independent clinical experts. Currently, sonographer technicians 1) acquire ultrasound images at the bedside and 2) provide preliminary interpretations prior to validation and overreading by cardiologists. This staggered, stepwise evaluation allows for the introduction of AI decision support at the time of the sonographer's preliminary interpretation and evaluation of which the cardiologist ultimately prefers.

While such AI algorithms have demonstrated improved precision on limited retrospective datasets, to date, there are no current cardiovascular AI technologies validated in blinded, randomized clinical trials. Additionally, human-computer interaction and the impact of AI prompting on clinical interpretations is under-explored in clinical studies. To address this need, we conducted a blinded, randomized non-inferiority clinical trial to prospectively assess the impact of initial assessment by AI vs. conventional initial assessment by a sonographer on final cardiologist interpretation of LVEF.

Study Design

Working with the image reporting system vendor, we've developed a blinded, automated process to change labels and storing initial sonographer tracings. Our workflow does not change the clinical workflow for cardiologists – they simply see a preliminary interpretation and/or measurement and are given the opportunity to change, adjust, correct, or keep the same as needed. The Cedars-Sinai Non-Invasive (Echocardiography) Laboratory is staffed by 10 cardiologists and ~approximately 50 sonographers, who will participate in the study. Sonographers are asked to use their standard clinical practice to annotate the left ventricle for either single-plane or bi-plane method-of-disks calculation of LVEF. Studies are excluded from the study if the study is low enough quality such that sonographer was unable to quantify LVEF.

With randomization, a proportion of the preliminary interpretations will be done by AI technology and the study team will assess how different this preliminary interpretation is from the final interpretation. Studies will be randomized 1:1 to either sonographer preliminary report finding or AI preliminary report finding with final adjudication by the cardiologist. The initial plan is for 1:1 allocation of randomization, with each study only having 1 preliminary report.



In blinded fashion, cardiologists will re-review clinical measurements and finalize an interpretation which will be reviewed with initial preliminary report (either by sonographer or AI) as well as original clinical interpretation as a metric of retest variation. At the end of the study, participants will be asked whether the blinding and randomization was successful. Questions will be targeted to seeing if participants could identify which preliminary reports were made by software.

Sample Size Calculation

Sonographers are expert adjudicators who create preliminary reports before review by the cardiologist. While there might be small disagreements within measurement error,

we define meaningful change as changes beyond 5% of a metric (IE, going from 45 to 50% in the ejection fraction) or a change in classification (going from mild to moderate). In this setting, we estimate metrics and measurements are meaningfully changed from the preliminary sonographer report by the cardiologist **8% of the time.** We hope to find a statistically meaningful decrease in the frequency of change when using AI decision support, such that the preliminary AI report is changed by the cardiologist **5% of the time.**

With an alpha of 0.05, power of 0.9, and a 1:1 enrollment ratio, we anticipate needing a sample size of 2834 studies, with 1417 studies in each arm. As a buffer against dropout and other issues, we anticipate enrolling 1750 patients in each arm for 3500 total.

Outcomes

Primary Outcome: Frequency and degree of change from the preliminary LVEF assessment to the final LVEF assessment report

A. Statistical Analysis: Comparison of proportion of studies for which the difference in LVEF between preliminary to overread is greater than 5% between the two arms
B. Statistical Analysis: Comparison of mean absolute error (MAE) of LVEF between preliminary to overread between the two arms

Secondary Outcomes:

1. Comparison of mean absolute error (MAE) of LVEF between cardiologist overread vs. historical clinical report LVEF
    A. Mean absolute error (MAE) of LVEF between cardiologist overread with initial sonographer interpretation vs. prior clinical report LVEF (Human Test-Retest Comparison)
    B. Mean absolute error (MAE) of LVEF between cardiologist overread with initial AI interpretation vs. prior clinical report LVEF (Safety Endpoint)

2. Time to complete each imaging study
    A. Comparison of time (in seconds) for AI vs. sonographer preliminary assessment
    B. Comparison of time (in seconds) for cardiologist overread between the two arms

3. Effects of the AI systems integration with computer-human interaction (Blinding)
    A. Evaluation of whether cardiologists can distinguish between AI vs. sonographer preliminary measurements.

Populations and Subgroups

The analysis will be performed with the intent to treat population (all randomized studies).

The primary outcome will be analyzed in subgroups including:

A. Method of LVEF (Single Plane or Biplane assessment)
B. Image quality of echocardiographic study
C. Inpatient vs. Outpatient echocardiogram
D. Race of Patient
E. Sex of Patient
F. Stratified by whether the cardiologist is able to tell initial agent of assessment

Analyses

The analysis will be performed with the intent to treat population (all randomized studies). Two sided Fisher's exact test will be used to assess for differences between groups for categorial outcomes and two-sided t-test will be used to assess for differences between groups for quantitative outcomes.

Outcome	Statistical Test
Initial vs. Final Assessment	
Substantial Change	Fisher's exact (two-sided) for superiority; T-Test (two sided, independent) with margin of 5% for non-inferiority
Mean Absolute Difference	T-test (two-sided, independent)
Final vs. Prior Cardiologist Assessment	
Substantial Change	Fisher's exact (two-sided)
Mean Absolute Difference	T-test (two-sided, independent)
Other Outcomes	
Sonographer time (s), median (IQR)	T-test (two-sided, independent)
Cardiologist time (s), median (IQR)	T-test (two-sided, independent)
Any Change	Fisher's exact (two-sided)

Primary Outcome: Frequency and degree of change from the preliminary LVEF assessment to the final LVEF assessment report

B. Statistical Analysis: Comparison of proportion of studies for which the difference in LVEF between preliminary to overread is greater than 5% between the two arms
C. Statistical Analysis: Comparison of mean absolute error (MAE) of LVEF between preliminary to overread between the two arms

Secondary Outcomes:

Comparison of mean absolute error (MAE) of LVEF between cardiologist overread vs. historical clinical report LVEF

C. Statistical Analysis: Mean absolute error (MAE) of LVEF between cardiologist overread with initial sonographer interpretation vs. prior clinical report LVEF
D. Statistical Analysis: Mean absolute error (MAE) of LVEF between cardiologist overread with initial AI interpretation vs. prior clinical report LVEF (Safety Endpoint)

Time to complete each imaging study

D. Statistical Analysis: Comparison of time (in seconds) for AI vs. sonographer preliminary assessment
E. Statistical Analysis: Comparison of time (in seconds) for cardiologist overread between the two arms

Effects of the AI systems integration with computer-human interaction (Blinding)

F. Evaluation of whether cardiologists can distinguish between AI vs. sonographer using Bang's blinding index.