The YODA Project
Research Proposal Review

The following page contains the final YODA Project review approving this proposal.

The Yale University Open Data Access (YODA) Project
Yale University Center for Outcomes Research and Evaluation (CORE)
195 Church Street, 5th Floor, New Haven, CT 06510

The YODA Project
Research Proposal Review - Final
(Protocol #: 2025-0804)

Reviewers:

- ☐ Nihar Desai
- ☐ Cary Gross
- ☐ Harlan Krumholz
- ☒ Richard Lehman
- ☒ Joseph Ross
- ☒ Joshua Wallach

Review Questions:	**Decision:**
1. Is the scientific purpose of the research proposal clearly described?	Yes
2. Will request create or materially enhance generalizable scientific and/or medical knowledge to inform science and public health?	Yes
3. Can the proposed research be reasonably addressed using the requested data?	Yes, or it's highly likely
4. Recommendation for this data request:	Approve

Comments:

No additional comments.

The YODA Project
Research Proposal Review

Revisions were requested during review of this proposal.
The following pages contain the original YODA Project review and the original submitted proposal.

The Yale University Open Data Access (YODA) Project
Yale University Center for Outcomes Research and Evaluation (CORE)
195 Church Street, 5th Floor, New Haven, CT 06510

The YODA Project
Research Proposal Review - Revisions Requested
(Protocol #: 2025-0804)

Reviewers:

- ☐ Nihar Desai
- ☐ Cary Gross
- ☐ Harlan Krumholz
- ☒ Richard Lehman
- ☒ Joseph Ross
- ☒ Joshua Wallach

Review Questions:	**Decision:**
1. Is the scientific purpose of the research proposal clearly described?	No
2. Will request create or materially enhance generalizable scientific and/or medical knowledge to inform science and public health?	Unsure, further clarification from requestor is needed
3. Can the proposed research be reasonably addressed using the requested data?	Yes, or it's highly likely
4. Recommendation for this data request:	Not Approve

Comments:

The Brief Project Background and Statement of Project Significance is too sparse. It would be helpful to have additional information about the observations motivating this work, along with references. The current description is rather informal. How will economic losses be quantified? It is difficult to evaluate how the proposed analyses will address the stated focus on economic losses. How will the analyses demonstrate that the estimates are valid estimates of the average treatment effect?

Principal Investigator

First Name: Gregory
Last Name: Sun
Degree: PhD Candidate
Primary Affiliation: Washington University in St Louis
E-mail: greg.s@wustl.edu
State or Province: MO
Country: United States

General Information

Are external grants or funds being used to support this research?: No external grants or funds are being used to support this research.
How did you learn about the YODA Project?: Data Holder (Company)

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2025/11/SV_57KskaKADT3U9Aq-R_3M0fGw7rYOxVwZ2-1.pdf

Certification

Certification: All information is complete; I (PI) am responsible for the research; data will not be used to support litigious/commercial aims.
Data Use Agreement Training: As the Principal Investigator of this study, I certify that I have completed the YODA Project Data Use Agreement Training

1. NCT00091715 - AC-052-364 - A Randomized, Double-blind, Placebo-controlled, Multicenter Study to Assess the Efficacy, Safety, and Tolerability of Bosentan in Patients With Mildly Symptomatic Pulmonary Arterial Hypertension (PAH)
2. NCT01323192 - JNS001-JPN-A01 - A Double-blind, Placebo-controlled, Parallel-Group Study to Evaluate the Efficacy and Safety of JNS001 in Adults With Attention-Deficit/Hyperactivity Disorder at Doses of 18 mg, 36 mg, 54 mg, or 72 mg Per Day
3. NCT00236665 - TOPMAT-OBHT-001 - A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Parallel Group Study of the Efficacy and Safety of Topiramate in the Treatment of Obese Patients With Mild to Moderate Essential Hypertension

What type of data are you looking for?: Individual Participant-Level Data, which includes Full CSR and all supporting documentation

Research Proposal

Project Title

On the use of baseline covariates in test retest designs

Narrative Summary:

Many clinical trials have a primary endpoint which can be described as the difference between an outcome measured at end-line compared to baseline. In all such settings, treatment effect estimators can have lower variances, provided that the baseline is adjusted for in a covariate adjustment. Using three example trials, I plan to demonstrate the degree of variance reduction typically available by doing so.

Scientific Abstract:

Background: Many current clinical trials compare baseline and endline measures in their primary endpoints. When doing so, the baseline is always available as a covariate for covariate adjustment, and adjusting can decrease treatment effect estimator variance. Despite this, few studies with this design actually adjust. The statistical efficiency losses from failing to do so are substantial.

Objective: I aim to quantify, using three example studies, how large the variance losses from failing to adjust are. For each study, I will compare estimators with and without covariate adjustment. I will also look at how variances vary by sample size.

Study Design: I will re-analyze data from three clinical trials with endline-minus-baseline outcomes. I compare a raw difference in means comparison with an adjusted analogue, and measure how much lower variance the adjusted estimates are.

Participants: all original participants in the original trials will be used in the analysis.

Primary and Secondary Outcome Measures: The primary outcome will as closely as possible follow what was done in the original studies.

Statistical Analysis: A raw mean comparison of baseline/endline differences and an ANCOVA estimate of the same difference, using the baseline as a covariate.

Brief Project Background and Statement of Project Significance:

This project is motivated by the observation that 1) statistical theory is unequivocal about the power benefits from including covariates in adjustment when analyzing data from randomized trials, while 2) practitioners often fail to do so. In test-retest designs, the fact that practice seems to lag behind theory is especially puzzling: the baseline is almost by construction typically a good predictor variable to use. In this study, I aim to quantify how substantial are the economic losses from not making the adjustments that theory recommends. I will do this by comparing standard errors across statistical analyses that do and do not adjust for the baseline.

Specific Aims of the Project:

In this project, I will replicate treatment effect estimators in three studies, without and without adjusting for the baseline measure as a covariate. I will quantify how large the variance of the resulting treatment effect estimator is, depending on if the baseline is used or not.

Study Design:

Methodological research

What is the purpose of the analysis being proposed? Please select all that apply.

Develop or refine statistical methods

Research on clinical trial methods

Research Methods

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study:

There are no exclusion criterion. In other words, no demographic information will be used to decide who does or does not end up in the sample. All individuals in the original dataset who were assigned treatment will be included for analysis.

The three trials I have chosen for this project were selected because they represent three examples of the test-retest structure, where the primary endpoint is a change in a measure. They were chosen to also correspond to different underlying conditions, to demonstrate the the basic statistical insight applies across contexts.

Primary and Secondary Outcome Measure(s) and how they will be categorized/defined for your study:

I will use the same primary endpoints as in the original publications. There are three studies, so I will describe each in detail sequentially.

The first trial is NCT00091715: Efficacy and Safety of Oral Bosentan in Pulmonary Arterial Hypertension Class II. In the publication corresponding to this trial, the main outcomes analyzed were pulmonary vascular resistance = (mean pulmonary arterial pressure - pulmonary capilary wedge pressure) * 80 / (cardiac output), as well as exercise capacity, as measured by 6-minute walking distance.

The second trial is NCT01323192: An Efficacy and Safety Study for JNS001 in Adults With Attention-Deficit Hyperactivity Disorder. The primary outcome reported in clincaltrials.gov is Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) Total Attention Deficit-Hyperactivity Disorder (ADHD) Symptoms Scores of Conners' Adult ADHD Rating Scale - Observer Screening Version (CAARS-O: SV), so I will take this as the primary outcome in my study as well.

The third trial is NCT00236665: A Study of Efficacy and Safety of Topiramate in the Treatment of Obese Patients With Mild to Moderate Essential Hypertension. The primary outcome measures reported in clinicaltrials.gov for this trial are body weight and sitting diastolic blood pressure, which once again, I will also take as my primary outcome.

Main Predictor/Independent Variable and how it will be categorized/defined for your study:

The main predictor variable in this study is simply treatment assignment in the respective studies. In study 1, this is an indicator for whether the drug bosentan was taken (instead of a placebo in control). In study 2, this is an indicator for whether the drug JNS001 was taken. Finally, in study 3, the predictor variable is whether or not the drug topiramate was taken.

Other Variables of Interest that will be used in your analysis and how they will be categorized/defined for your study:

In each of the studies I plan on studying, the primary endpoint of the original study is a mean difference of the relevant measure(s), comparing baseline and endline.

Take study 2 as an example, which is easiest to describe, since it has a single outcome measure, a score on the diagnostic test: CAARS-O: SV. This score is measured twice, once at the start of the trial (baseline) and once at the end (endline). The primary _outcome_ is the difference, (endline - baseline). In addition to this difference, I will also separately use the baseline variable itself. The purpose here is to use the baseline in a covariate adjustment analysis to show how much doing so reduces variance.

In studies 1 and 3, I will do the analogous analysis, but with their respective outcome variables, as defined earlier.

Statistical Analysis Plan:

I will run two tests.

One is a simple two-sample t-test, comparing treatment and control. I will compute means of the outcome in treatment and control, compute standard errors of each respective mean, and compute the difference in means, as well as the standard error in this difference.

A second test is based on an analysis of covariance. The procedure can be described as follows. I will first run two OLS regressions, one using only treated observations and one using only control observations. The regression will be of the change, which is the primary endpoint, regressed on the baseline. Let beta_t, beta_c be the slope from these regressions. I will then perform a 2-sample t-test on the constructed quantity [change - (beta_t+beta_c)/2 * baseline], as I did with the raw outcomes. In the accompanying paper I am trying to write, I show 1) that this is also a valid estimate of the average treatment effect, 2) the standard errors from doing the 2-sample t-test are asymptotically correct, and 3) the variance of this procedure is lower. I aim to quantify how much lower.

Software Used:

R, RStudio

Project Timeline:

Start date - November 2025
Analysis completion date - January 2026
Manuscript draft - January 2026
Submitted for publication - February 2027

Dissemination Plan:

I plan to submit the resulting analysis as part of a paper for the American Journal of Epidemiology.