Principal Investigator

First Name: Ryan
Last Name: Shewcraft
Degree: PhD
Primary Affiliation: Standard Model Biomedicine
E-mail: ryan@standardmodel.bio
State or Province: CT
Country: USA

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?: Colleague

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2026/07/Shewcraft_YODA_COI.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. NCT01591122 - ABI-PRO-3002 - A Phase 3, Randomized, Double-blind, Placebo-Controlled Study of Abiraterone Acetate (JNJ-212082) Plus Prednisone in Asymptomatic or Mildly Symptomatic Patients With Metastatic Castration-Resistant Prostate Cancer
2. NCT01715285 - 212082PCR3011 - A Randomized, Double-blind, Comparative Study of Abiraterone Acetate Plus Low-Dose Prednisone Plus Androgen Deprivation Therapy (ADT) Versus ADT Alone in Newly Diagnosed Subjects With High-Risk, Metastatic Hormone-naive Prostate Cancer (mHNPC)
3. NCT01867710 - 212082PCR2023 - A Randomized Phase 2 Study Evaluating Abiraterone Acetate With Different Steroid Regimens for Preventing Symptoms Associated With Mineralocorticoid Excess in Asymptomatic, Chemotherapy-naïve and Metastatic Castration-resistant Prostate Cancer (mCRPC) Patients
4. NCT00544440 - COU-AA-BMA - An Observational Study of Continuous Oral Dosing of an Irreversible CYP17 Inhibitor, Abiraterone Acetate (CB7630), in Castration-Resistant Prostate Cancer Patients Evaluating Androgens and Steroid Metabolites in Bone Marrow Plasma
5. NCT01795703 - JNJ-212082-JPN-202 - A Phase II Study of JNJ-212082 (Abiraterone Acetate) in Metastatic Castration-Resistant Prostate Cancer Patients Who Have Received Docetaxel-based Chemotherapy
6. NCT00474383 - COU-AA-003 - A Phase II Open Label Study of CB7630 (Abiraterone Acetate) in Patients With Advanced Prostate Cancer Who Have Failed Androgen Deprivation and Docetaxel-Based Chemotherapy
7. NCT01685983 - 212082PCR2007 - A Phase 2 Open Label Study of Abiraterone Acetate (JNJ-212082) and Prednisolone in Patients With Advanced Prostate Cancer Who Have Failed Androgen Deprivation and Docetaxel-Based Chemotherapy.

8. NCT00485303 - COU-AA-004 - A Phase II Open Label Study of CB7630 (Abiraterone Acetate) and Prednisone in Patients With Advanced Prostate Cancer Who Have Failed Androgen Deprivation and Docetaxel-Based Chemotherapy
9. NCT01695135 - ABI-PRO-3001 - A Phase 3, Randomized, Double-blind, Placebo-Controlled Study of Abiraterone Acetate (JNJ-212082) Plus Prednisone in Patients With Metastatic Castration-Resistant Prostate Cancer Who Have Failed Docetaxel-Based Chemotherapy
10. NCT01314118 - 212082PCR2005 - A Multicenter, Open-label, Single-arm, Phase 2 Study of Abiraterone Acetate Plus Prednisone in Subjects With Advanced Prostate Cancer Without Radiographic Evidence of Metastatic Disease
11. NCT00887198 - COU-AA-302 - A Phase 3, Randomized, Double-blind, Placebo-Controlled Study of Abiraterone Acetate (CB7630) Plus Prednisone in Asymptomatic or Mildly Symptomatic Patients With Metastatic Castration-Resistant Prostate Cancer
12. NCT00638690 - COU-AA-301 - A Phase 3, Randomized, Double-Blind, Placebo-Controlled Study of Abiraterone Acetate (CB7630) Plus Prednisone in Patients With Metastatic Castration-Resistant Prostate Cancer Who Have Failed Docetaxel-Based Chemotherapy
13. NCT01381874 - 212082BCA2001 - Randomized, Open-Label Study of Abiraterone Acetate (JNJ-212082) Plus Prednisone With or Without Exemestane in Postmenopausal Women With ER+ Metastatic Breast Cancer Progressing After Letrozole or Anastrozole Therapy
14. NCT02489318 - 56021927PCR3002 - A Phase 3 Randomized, Placebo-controlled, Double-blind Study of Apalutamide Plus Androgen Deprivation Therapy (ADT) Versus ADT in Subjects With Metastatic Hormone-sensitive Prostate Cancer (mHSPC)
15. NCT01946204 - ARN-509-003 - A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase III Study of ARN-509 in Men With Non-Metastatic (M0) Castration-Resistant Prostate Cancer
16. NCT02257736 - 56021927PCR3001 - A Phase 3 Randomized, Placebo-controlled Double-blind Study of JNJ-56021927 in Combination With Abiraterone Acetate and Prednisone Versus Abiraterone Acetate and Prednisone in Subjects With Chemotherapy-naive Metastatic Castration-resistant Prostate Cancer (mCRPC)
17. NCT01790126 - ARN-509-002 - The Role of Highly Selective Androgen Receptor (AR) Targeted Therapy in Men With Biochemically Relapsed Hormone Sensitive Prostate Cancer
18. NCT03390504 - 42756493BLC3001 - A Phase 3 Study of Erdafitinib Compared With Vinflunine or Docetaxel or Pembrolizumab in Subjects With Advanced Urothelial Cancer and Selected FGFR Gene Aberrations
19. NCT02365597 - 42756493BLC2001 - An Efficacy and Safety Study of Erdafitinib (JNJ-42756493) in Participants With Urothelial Cancer
20. NCT00653952 - 30-57 - A Phase 3, Randomized, Open-Label, Comparative Study of CAELYX(R) versus Paclitaxel HCl in Patients with Epithelial Ovarian Carcinoma Following Failure of First-Line, Platinum-Based Chemotherapy
21. - 30-49 - A Phase 3, Randomized, Open-Label, Comparative Study of DOXIL/CAELYX(R) versus Topotecan HCl in Patients with Epithelial Ovarian Carcinoma Following Failure of First-Line, Platinum-Based Chemotherapy

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

Research Proposal

Project Title

An Open, Tool-Agnostic Benchmark for Evaluating Trial-Design Efficiency Methods Using Completed Oncology Trials

Narrative Summary:

Trial-efficiency methods (covariate adjustment, prognostic/predictive enrichment, etc.) are usually validated on bespoke simulations, so cross-method comparison is unreliable and claims hard to

reproduce. We propose an open, reusable benchmark quantifying, on completed oncology trials with known outcomes, how much each method could have improved efficiency (power, sample size, duration, precision) without changing each trial's conclusion. We will (1) reproduce each trial's primary analysis as a reference; (2) define benchmark tasks with pre-registered estimands, standardized metrics, and type-I-error and sign/coverage guardrails; and (3) return the tasks, evaluation harness, and leaderboard to YODA for shared use. This lets methods be compared on identical real-data tasks, advancing generalizable knowledge for more efficient trials.

Scientific Abstract:

Background: Methods that improve randomized-trial efficiency (covariate adjustment, prognostic-covariate adjustment, prognostic and predictive enrichment) are typically validated by their developers on bespoke simulations. Because each is assessed on data-generating processes its proponents choose, cross-method comparison is unreliable and efficiency claims are hard to reproduce or generalize.

Objective: Build an open, reusable benchmark quantifying, on completed oncology trials with known outcomes, how much each design method could have improved efficiency while preserving each trial's original inferential conclusion.

Study Design: Methodological benchmarking study re-emulating completed trials, releasing tasks, an evaluation harness, and a leaderboard to YODA for shared use.

Participants: De-identified participant-level data from completed Janssen oncology trials.

Primary and Secondary Outcome Measures: Efficiency metrics: relative efficiency and effective sample-size reduction, subsampling-based power curves, and time-to-decision, each constrained by type-I-error and validity guardrails, including preservation of the treatment-effect sign and confidence-interval coverage of the full-data estimate.

Statistical Analysis: Reproduce each trial's pre-specified primary analysis as a reference, then score candidate methods on pre-registered estimands and standardized metrics, enabling academic and industry methods to be compared on identical real-data tasks, as shared benchmarks have done for machine learning.

Brief Project Background and Statement of Project Significance:

The clinical-trials enterprise is under sustained pressure to answer efficacy and safety questions with fewer participants, shorter timelines, and greater precision. A large methodological literature offers levers toward this goal: covariate adjustment (endorsed for confirmatory use by FDA's 2023 covariate-adjustment guidance), prognostic-covariate adjustment that borrows information from historical controls, enrichment and predictive-biomarker targeting, group-sequential and adaptive designs that permit early stopping for efficacy or futility, and Bayesian dynamic borrowing from external data. Each can, in principle, increase power at fixed sample size or reduce the sample size and duration needed to reach a decision.

The field lacks a common yardstick. New efficiency methods are almost always validated on simulations whose data-generating processes are chosen by the method's own developers. This creates two problems: optimism bias, because a method is evaluated on the very conditions its authors judged favorable; and non-comparability, because two methods are rarely tested on the same tasks. As a result, reported efficiency gains are hard to reproduce and harder still to rank across methods. This situation slows adoption of genuinely useful designs and permits overstated claims to persist.

Completed randomized trials provide something simulations cannot: real data with a known, pre-specified answer. Re-emulating a completed trial and asking "how much more efficiently could this

question have been answered, without changing the conclusion?" yields a ground-truth, reproducible measure of a design method's value. Building this into a shared benchmark with fixed tasks, standardized metrics, and a public leaderboard would let any group, academic or industry, evaluate its method on identical real-data problems. This is the model that shared benchmarks provided for machine learning, and it is squarely within the YODA Project's accepted purposes of developing and refining statistical methods and conducting research on clinical-trial methods.

The significance is threefold. Scientifically, the benchmark converts a fragmented, self-graded literature into a comparable, reproducible evidence base on trial-design efficiency. For public health, methods that genuinely reduce sample size and duration lower patient burden and accelerate access to effective therapies; a trustworthy benchmark helps identify which methods deliver. For open science, all task specifications, reference results, and derived methodological artifacts will be free to use by any YODA-approved applicant.

Specific Aims of the Project:

Aim 1: Reference reproductions. Construct harmonized analytic datasets for a panel of completed oncology trials and reproduce each trial's pre-specified primary analysis, establishing reference results and a documented "task card" (estimand, population, endpoint, stratification, and any deviations) for each trial.
Aim 2: Efficiency-benchmark tasks and metrics. Define and validate a suite of benchmark tasks spanning design families: (a) covariate and prognostic-covariate adjustment, (b) prognostic and predictive enrichment, (c) open-ended novel designs (e.g. group-sequential or adaptive designs). Proposed designs will be scored by standardized efficiency metrics under explicit type-I-error and validity guardrails.
Aim 3: Benchmark release. Give the benchmark back to YODA: task specifications, an evaluation harness (code), reference results, and a leaderboard, enabling equal-footing assessment of academic and industry methods.

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 on clinical prediction or risk prediction

Research Methods

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

No additional inclusion/exclusion criteria.

The goal is to replicate trial outcomes and see if other statistical methods would give better operating characteristics using the original trial cohort. As such, the data source is the trial data from YODA and nothing else.

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

Primary outcome measures. As a methodological benchmarking study, our outcome measures are properties of each candidate design method scored on the benchmark tasks, not patient endpoints.

The source trials' clinical endpoints (overall survival, radiographic PFS, metastasis-free survival, PFS, each as pre-specified in that trial's SAP) are the substrate on which metrics are computed. Primary measures:

1. Relative efficiency / effective sample-size reduction: ratio of the reference (as-conducted) estimator variance to the candidate method's variance for the trial's primary treatment-effect estimand, expressed as equivalent % reduction in N at fixed power.
2. Sample size at target power: N needed for 80% and 90% power under the candidate method, from subsampling-based power curves, versus actual enrolled N.
3. Validity (decision preservation): paired constraints scored for every efficiency claim: (a) empirical type-I error under the null (permutation), required <= nominal; and (b) preservation of treatment-effect sign plus CI coverage of the full-data estimate.

Efficiency is credited only when validity criteria are met; a method that gains efficiency while inflating type-I error or reversing the conclusion fails that task.

Secondary outcome measures.

1. Time-to-decision / expected duration: for group-sequential/adaptive tasks, information time (and implied duration) to cross an efficacy/futility boundary, and expected N under null and observed alternative.
2. Estimation precision: CI width for the primary estimand versus the reference.
3. Effective borrowed sample size: for borrowing tasks, effective borrowed control observations, with bias--variance and prior-data-conflict diagnostics.

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

Each proposed trial design method is applied to the same source-trial data and the same primary estimand, so its levels are directly comparable and the reference (as-conducted) analysis serves as the common baseline against which every method's independent effect on the outcome measures is quantified.

Two secondary independent factors are held fixed or varied systematically and reported alongside the main variable: the source trial (each requested Janssen oncology trial, treated as a repeated benchmark instance) and, for sample-size and power tasks, the sampling fraction of the subsampling framework (the grid of fractions used to build empirical power curves).

Every level of the main independent variable, and the values of the secondary factors, are pre-registered before analysis and documented in each trial's task card, so the specification can be readily compared to the methods and results reported in the resulting publication and leaderboard. No other independent variables will be tested.

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

Other variables include the baseline participant characteristics and trial-structure variables drawn from each source trial's de-identified data, used to (a) describe the benchmark sample and (b) serve as adjustment covariates in the covariate-adjustment, prognostic-score, and enrichment tasks.

Sample characterization variables (reported per trial to document each benchmark instance): demographics (age (years), sex, race/ethnicity, and geographic region; baseline functional status, etc); ECOG performance status; treatment-arm assignment; analysis-population flags as defined in the trial's SAP; etc.

Multivariable risk-adjustment / prognostic variables (used as inputs to adjusted estimators and prognostic-score models): the trial's pre-specified randomization stratification factors, plus standard baseline prognostic covariates available for the indication. For prostate-cancer trials these include

disease burden and metastatic status (e.g., visceral vs. bone-only disease), baseline PSA, Gleason score, and prior therapy exposure; for multiple-myeloma trials, ISS stage, cytogenetic risk category, and number of prior lines of therapy. Baseline laboratory values (e.g., hemoglobin, LDH, alkaline phosphatase) will be included where available and clinically relevant to the endpoint.

Time-to-event structure variables: follow-up time and the event/censoring indicator for each endpoint, defined consistent with the source trial's SAP (including its censoring rules), since these define the outcome and enter every efficiency computation.

Statistical Analysis Plan:

Reference reproduction. For each trial we first reproduce the pre-specified primary analysis to establish reference results. For time-to-event endpoints (OS, rPFS, MFS, PFS) this may be a stratified log-rank test and a Cox proportional-hazards model using the trial's SAP-defined stratification, censoring rules, and analysis population, with the treatment-effect estimand (hazard ratio and 95% CI) confirmed against the CSR/published result within tolerance. Proportional-hazards assumptions are checked via scaled Schoenfeld residuals; where PH is violated, restricted mean survival time (RMST) differences are computed as a complementary reference.
Descriptive analyses. Baseline characteristics are summarized per trial and per arm: continuous variables as mean (SD) and median (IQR), categorical variables as counts and percentages, with missingness tabulated for every variable. Follow-up is summarized by median follow-up (reverse Kaplan--Meier) and event counts. These characterize each benchmark instance and are reported in the task card.
Bivariate analyses. Unadjusted associations between each candidate covariate and the endpoint are estimated (univariable Cox HRs with 95% CIs; Kaplan--Meier curves with log-rank tests across strata) to document prognostic strength and inform prognostic-score construction. Baseline balance between arms is assessed by standardized mean differences rather than significance tests.
Multivariable analyses. Adjusted treatment effects are estimated with multivariable Cox models including the pre-specified stratification factors and baseline prognostic covariates defined in the task card. For the covariate-adjustment benchmark family we compare the unadjusted reference to covariate-adjusted estimators and to prognostic-score adjustment (a PROCOVA-style approach in which a control-outcome model, fit on external or held-out control data, produces a prognostic score entered as a single covariate). Efficiency is quantified as the ratio of reference to adjusted estimator variance, expressed as effective sample-size reduction. Cross-fitting is used wherever a fitted model (e.g., the prognostic score, or a predictive-enrichment rule) is applied to the same data, to prevent optimistic bias.
Advanced analyses.
Subsampling power engine. Sample-size and power outcomes are estimated non-parametrically by nested random subsampling: at a grid of sampling fractions we draw repeated stratified subsamples (preserving arm ratio and stratification), re-estimate the treatment effect under each candidate method, and construct empirical power curves (proportion of subsamples rejecting the null vs. N). The N required for 80% and 90% power is read from these curves and compared to the trial's enrolled N. All metrics carry resampling-based uncertainty intervals.
Group-sequential / adaptive analyses. We reconstruct the accruing information order and apply alpha-spending boundaries (O'Brien--Fleming, Pocock) with futility rules, estimating the information time (and implied duration) to cross an efficacy or futility boundary and the expected sample size under both the null and the observed alternative.
Enrichment. Prognostic- and predictive-enrichment tasks compare power and effect estimation in enriched versus all-comers populations, with subgroups defined on pre-specified biomarkers/risk strata, cross-fitted rule estimation, and appropriate multiplicity control.
Validity guardrails. Every efficiency claim is paired with (a) an empirical type-I-error check under the null, obtained by permutation/label-shuffling of the treatment assignment, required to remain at or below the nominal level; and (b) a decision-preservation check requiring the reduced-design analysis to retain the sign of the treatment effect and to yield a CI covering the full-data point estimate. Tasks failing either check are scored as failures, not efficiency gains. Where distributional assumptions are questionable, non-parametric alternatives (log-rank, permutation inference, RMST) are used in place of parametric tests.
Software and reproducibility. Pipelines are versioned and seeded; every estimand, covariate set, and

metric is pre-registered before methods are applied and documented in each trial's task card, so the analysis specification can be compared directly to the results reported in the publication and leaderboard.

Software Used:

Python

Project Timeline:

Months 1--3
Secure-platform onboarding; DUA execution; data harmonization; construction of analytic datasets.
Months 3--5
Reproduce each trial's primary analysis; establish reference results and task cards (Aim 1).
Months 5--10
Implement the four design-lever method families; build and validate benchmark tasks and metrics; calibrate type-I-error guardrails (Aim 2).
Months 10--12
Build the evaluation harness and leaderboard; define the third-party submission protocol; internal reproducibility audit (Aim 3).
Months 12--15
Manuscript preparation and submission; release of code and benchmark; results reported to the YODA Project and Data Partner.

Dissemination Plan:

Peer-reviewed publication in a methods or applied venue.
Preprint posted to arXiv and/or medRxiv.
Presentation at a scientific meeting.
Release of the benchmark task specifications and evaluation code through YODA.
Compliance with YODA policy: the unique application identifier and required acknowledgment language will appear in all outputs; the abstract will be shared with the Data Partner; no participant-level data will be redistributed; and any unexpected or serious safety findings will be reported to the Data Partner immediately.

Bibliography:

1. U.S. Food and Drug Administration. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. Guidance for Industry, 2023.
2. U.S. Food and Drug Administration. Enrichment Strategies for Clinical Trials to Support Demonstration of Effectiveness of Human Drugs and Biological Products. Guidance for Industry, 2019.
3. International Council for Harmonisation. E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials, 2019.
4. Benkeser D, Díaz I, Luedtke A, et al. Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. Biometrics, 2021.
5. Schuler A, Walsh D, Hall D, et al. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. International Journal of Biostatistics, 2022.
6. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics, 2011.
7. de Bono JS, Logothetis CJ, Molina A, et al. Abiraterone and increased survival in metastatic prostate cancer . New England Journal of Medicine, 2011.
8. Ryan CJ, Smith MR, de Bono JS, et al. Abiraterone in metastatic prostate cancer without previous chemotherapy. New England Journal of Medicine, 2013.
9. Smith MR, Saad F, Chowdhury S, et al. Apalutamide treatment and metastasis-free survival in prostate cancer. New England Journal of Medicine, 2018.
10. Chi KN, Agarwal N, Bjartell A, et al. Apalutamide for metastatic, castration-sensitive prostate

cancer. New England Journal of Medicine, 2019.
11. Palumbo A, Chanan-Khan A, Weisel K, et al. Daratumumab, bortezomib, and dexamethasone for multiple myeloma. New England Journal of Medicine, 2016.
12. Dimopoulos MA, Oriol A, Nahi H, et al. Daratumumab, lenalidomide, and dexamethasone for multiple myeloma. New England Journal of Medicine, 2016.
13. Facon T, Kumar S, Plesner T, et al. Daratumumab plus lenalidomide and dexamethasone for untreated myeloma. New England Journal of Medicine, 2019.