Principal Investigator

First Name: Xiaolin
Last Name: Zhong
Degree: M.D.
Primary Affiliation: The Affiliated Hospital of Southwest Medical University
E-mail: zhongxl519@hotmail.com
State or Province: Sichuan
Country: China

General Information

Key Personnel (other than PI):
First Name: Yantong
Last name: Li
Degree: MBBS
Primary Affiliation: The Affiliated Hospital of Southwest Medical University
SCOPUS ID:
Requires Data Access? Yes

Are external grants or funds being used to support this research?: External grants or funds are being used to support this research.
Project Funding Source: Science and Technology Department of Sichuan Province grant 2022YFS0633
How did you learn about the YODA Project?: Other

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2025/06/SV_57KskaKADT3U9Aq-R_4Yo9abspjqkIg1P.pdf
https://yoda.yale.edu/wp-content/uploads/2025/07/SV_57KskaKADT3U9Aq-R_4zxe7zI2kGeoENM.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. NCT01551290 - CR018769; REMICADEUCO3001 - A Phase 3, Multicenter, Randomized, Double-Blind, Placebo-Controlled Study Evaluating the Efficacy and Safety of Infliximab in Chinese Subjects With Active Ulcerative Colitis
2. NCT00036439 - C0168T37 - A Randomized, Placebo-controlled, Double-blind Trial to Evaluate the Safety and Efficacy of Infliximab in Patients With Active Ulcerative Colitis
3. NCT00096655 - C0168T46 - A Randomized, Placebo-controlled, Double-blind Trial to Evaluate the Safety and Efficacy of Infliximab in Patients With Active Ulcerative Colitis
4. NCT01369329 - CNTO1275CRD3001 - A Phase 3, Randomized, Double-blind, Placebo-controlled, Parallel-group, Multicenter Study to Evaluate the Safety and Efficacy of Ustekinumab Induction Therapy in Subjects With Moderately to Severely Active Crohn's Disease Who Have Failed or Are Intolerant to TNF Antagonist Therapy (UNITI-1)
5. NCT02407236 - CNTO1275UCO3001 - A Phase 3, Randomized, Double-blind, Placebo-

controlled, Parallel-group, Multicenter Protocol to Evaluate the Safety and Efficacy of Ustekinumab Induction and Maintenance Therapy in Subjects With Moderately to Severely Active Ulcerative Colitis

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

Research Proposal

Project Title

Inferential interpretability analysis and external validation study on the efficacy of biologics in UC patients

Narrative Summary:

Ulcerative colitis (UC) is a chronic disease causing painful colon inflammation, often treated with biologic therapies. However, predicting which biologic will work best for each patient remains a challenge. This study will develop a machine learning tool to help doctors personalize UC treatment. Using data from past clinical trials--including patient age, symptoms, and lab results--we'll train the tool to predict which biologics are most likely to succeed for individual patients. We'll test its accuracy against traditional statistical methods and identify key factors influencing treatment response. If successful, this tool could reduce trial-and-error in UC care, helping patients achieve faster symptom relief and better long-term outcomes.

Scientific Abstract:

Background: Ulcerative colitis (UC) patients show highly variable responses to biologic therapies, leading to suboptimal treatment selection in clinical practice. Current approaches lack robust methods for predicting individual treatment effects.
Objective: Develop and validate an interpretable machine learning model to estimate personalized treatment effects of three biologic classes in UC, facilitating data-driven therapeutic decisions.
Study Design: Multicenter cohort study using retrospective/prospective data for development, with external validation via real-world electronic medical records.
Participants: Adults with UC receiving biologics (n=XX), with >=12 weeks follow-up and complete outcome data. Exclusion: Prior biologic failure, incomplete records.
Outcomes:
Primary: Clinical remission/response at 12 weeks
Secondary: (1) Adverse events, (2) Endoscopic/lab improvements, (3) Model performance (AUC>=0.75 target)
Statistical Analysis:
Multiple imputation for missing data; Causal forest for individualized treatment effect estimation (CATEs)SHAP for feature interpretationExternal validation of discrimination/calibrationComparison to conventional logistic regression.

Brief Project Background and Statement of Project Significance:

Ulcerative colitis (UC) is a chronic inflammatory bowel disease with significant morbidity, and biologic therapies have become essential for patients with moderate-to-severe disease; however, there is substantial variability in individual response to different biologic classes, making optimal treatment selection challenging. This research aims to apply advanced machine learning methods--specifically, causal forest algorithms combined with SHAP explanations--to develop and validate an interpretable model that predicts personalized efficacy of major biologic agents in UC. The project's significance lies in its potential to improve clinical decision-making, leading to more

effective, individualized treatment for UC patients and reducing trial-and-error drug selection. The resulting decision support tool and validated model will materially enhance generalizable scientific knowledge by advancing the field of precision medicine, and findings can inform treatment guidelines, clinical practice, and future research in inflammatory bowel disease. Prior work has explored predictors of biologic response in UC , but robust, interpretable, and externally validated individualized prediction tools remain limited.

Specific Aims of the Project:

The specific aims of this project are to develop and validate an interpretable machine learning model using causal forest and SHAP to predict individualized treatment effects of three major biologic classes in ulcerative colitis, to evaluate the model's accuracy and generalizability using multicenter retrospective and prospective cohorts as well as electronic medical record data, and to create a clinical decision support tool for optimizing biologic selection in UC patients; the primary hypothesis is that the model will accurately identify patients most likely to benefit from each biologic class, thereby improving personalized treatment and clinical outcomes.

Study Design:

Individual trial analysis

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

New research question to examine treatment effectiveness on secondary endpoints and/or within subgroup populations

Research Methods

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

No exclusion criteria.
Additional Data Sources and Analysis Plan:
We plan to supplement the YODA project data sourced through the Vivli platform. These studies were selected because they: Investigate the same biologic classes in UC, Include comparable outcome measures (e.g., clinical remission at 12 weeks), and Share key baseline variables (e.g., disease severity, prior treatments).
Data Integration Approach: All analyses will be performed within the Vivli platform's secure research environment to ensure compliance with data use agreements. IPD from YODA and Vivli will be harmonized and merged using common data elements (e.g., standardized variable names for outcomes, demographics).
No external data will be exported; all statistical modeling (causal forest/SHAP) will be executed within Vivli.
Justification for Merging:
Combining datasets will increase sample size and diversity, improving the generalizability of our prediction model. We will account for between-study heterogeneity through: Stratified analyses by data source, Covariate adjustment.

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

Primary Outcome Measure(s):
Clinical Remission at Week 8 (or end of induction period): Defined as a total Mayo score <=2, with no individual subscore >1 and rectal bleeding subscore of 0.
Clinical Response at Week 8: Defined as a reduction from baseline in the total Mayo score of >=3 points and >=30%, with an accompanying decrease in the rectal bleeding subscore of >=1 or an absolute rectal bleeding subscore of 0 or 1.

Secondary Outcome Measure(s):
Endoscopic Remission at Week 8: Defined as a Mayo endoscopic subscore of 0 or 1.
Steroid-free Clinical Remission at Week 8: As defined above, achieved without the use of corticosteroids.
Sustained Clinical Remission through Week 52 (maintenance period): Defined by the same criteria as above, maintained at all subsequent assessments up to Week 52.
Change in Inflammatory Biomarkers (e.g., CRP, fecal calprotectin): Measured as absolute or relative change from baseline to Week 8 and/or Week 52.
Adverse Events: Incidence of serious and non-serious adverse events during the study period.

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

Main Independent Variable(s):
The main independent variable(s) of this study will be:
Treatment Group Assignment: Defined as the allocation of participants to either the intervention group or to the control group (e.g., placebo or standard of care), as determined by the original study protocol.
Example: "Participants randomized to receive [Drug X, 10 mg once daily] compared with those randomized to placebo."
Duration of Exposure: The length of time participants receive the assigned intervention (e.g., number of weeks on study drug).
Baseline Disease Activity: When relevant, baseline disease severity or activity (e.g., total Mayo score, baseline biomarker levels) may be considered as an additional independent variable for subgroup or adjusted analyses.
All independent variables are defined as per the original trial protocol and dataset documentation, ensuring clarity and reproducibility. Variable codings and definitions will match those described in the publication of the primary trial results, to allow for direct comparison with final analyses.

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

Other Variables Used in Analysis

To characterize the study sample and for purposes of multivariable risk adjustment, the following variables will be included:

1. Demographic Variables:
Age: Defined as years at baseline/randomization, analyzed as a continuous variable and/or categorized (e.g., 65).
Sex: Male or Female, as recorded at baseline.
Race/Ethnicity: As reported in the original dataset (e.g., White, Black, Asian, Hispanic, Other).
2. Clinical Characteristics:
Baseline Disease Duration: Time since initial diagnosis (in years).
Baseline Disease Severity/Activity: Defined by validated scores (e.g., total Mayo score or other disease-specific scale at baseline).
Comorbidities: Presence of relevant comorbid conditions (e.g., diabetes, cardiovascular disease), as defined in the original study data.
3. Concomitant Medications: Use of Other Medications: Such as corticosteroids, immunomodulators, or biologics, documented at baseline and during the study period (coded as Yes/No).
4. Laboratory and Biomarker Data: Baseline Biomarker Levels: Such as C-reactive protein (CRP), fecal calprotectin, or other disease-specific markers, measured at baseline and specified in standard units.
5. Geographic Region/Site: Study Site or Geographic Region: As coded in the dataset, if applicable.
6. Prior Treatment History: Previous Use of Study Drug/Class: Documented as Yes/No, based on participant's medical history at baseline.

Statistical Analysis Plan:

Data Preparation and Management All clinical study data will be accessed and analyzed exclusively within the secure, password-protected analytic workspace provided by the Vivli platform. Data cleaning procedures will be performed within this secure environment, including checks for completeness, identification of outliers, and handling of missing data according to pre-specified rules (e.g., multiple imputation, last observation carried forward, or as per protocol). Variables will be coded and labeled consistently, and a data dictionary will be developed within the secure workspace. Handling Differences Among Studies and Maintaining Study Independence To ensure the integrity and independence of each study and to account for differences among studies (e.g., study design, patient population, intervention, follow-up time), the following strategies will be used: Separate Analyses: All analyses will first be conducted separately for each individual study. Study-level results (e.g., effect sizes, standard errors) will be calculated independently to maintain the structure of each dataset. Meta-Analysis Approach: Where appropriate, study-specific results will be combined using meta-analytic techniques (such as random-effects or fixed-effects meta-analysis models), which explicitly account for between-study heterogeneity and ensure independence of study results. Heterogeneity will be assessed using the I^2 statistic and Cochran's Q test. Pooled Analysis with Study as a Covariate: For pooled individual participant data (IPD) analyses, study ID will be included as a stratification factor or fixed effect in all regression or time-to-event models, thus adjusting for inter-study differences and preserving the independence of each study population. Sensitivity Analyses: Sensitivity analyses will be conducted to explore the impact of different study characteristics (e.g., study design, region, inclusion criteria) on the overall results. Data for Each Stage of Model Development Data Cleaning and Exploration: All available data from each individual study will be used to assess data quality and distribution of key variables. Data Splitting for Machine Learning: For machine learning analyses, datasets will be split into distinct training (e.g., 70--80%) and validation/test (e.g., 20--30%) sets, either within each study or across pooled datasets, depending on the specific analysis. Care will be taken to avoid data leakage and maintain independence between training and validation sets. Model Development and Feature Selection: Prediction models will be developed using the training dataset. Feature selection may be performed using univariate analyses or regularization techniques (e.g., LASSO, elastic net) as appropriate. Model Validation: Model performance will be assessed using the validation/test dataset, with metrics including accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and calibration. Cross-validation (e.g., k-fold) may be used for internal validation where feasible. For multi-study analyses, external validation may be performed by training on one study and validating on another. Comparison to Logistic Regression: As a benchmark, standard multivariable logistic regression models will be fit on the same training data and evaluated on the same validation/test data as the machine learning models. Performance metrics will be directly compared to evaluate added value of machine learning approaches. Descriptive Statistics Baseline characteristics (e.g., age, gender, disease duration) will be summarized for each study using means and standard deviations for continuous variables, and frequencies and percentages for categorical variables. Group comparisons at baseline will be assessed within each study. Primary and Secondary Outcome Analyses The primary endpoint will be analyzed according to the intention-to-treat (ITT) principle within each study. Study-level primary and secondary outcome results will then be synthesized by meta-analysis as needed. For continuous outcomes, comparisons between groups will be performed using independent t-tests or Analysis of Covariance (ANCOVA) adjusting for baseline values. For binary outcomes, chi-square tests or logistic regression models will be applied. For time-to-event data, Kaplan-Meier survival analysis and log-rank tests will be used, with Cox proportional hazards models applied for adjusted analyses as appropriate. Subgroup and Sensitivity Analyses Pre-specified subgroup analyses (e.g., by age group, disease severity, geographic region) will be conducted within and across studies as appropriate, exploring consistency of findings. Sensitivity analyses may be performed to assess robustness to different analytic assumptions or missing data methods. (See supplementary doc for more)

Software Used:

R, RStudio, STATA

Project Timeline:

Anticipated Project Start Date:
July 17, 2025 (Upon data request approval and Data Use Agreement execution)
Data Acquisition and Cleaning Completed By:
August 20, 2025
Data Analysis Completion Date:
September 20, 2025
Manuscript Drafted:
September 30. 2025
First Manuscript Submission for Publication:
October 1, 2025
Results Reported Back to the YODA Project:
By October 15, 2025 (within 2 weeks of manuscript submission)

Dissemination Plan:

Anticipated Products and Target Audience(s):
The primary product of this study will be a peer-reviewed manuscript reporting the main findings of the research. Additional products may include abstract submissions to relevant scientific conferences and summary reports for data-sharing partners and stakeholders.
Target Audiences:
Clinical researchers and scientists in the relevant medical/health field
Healthcare professionals and practitioners seeking to improve patient care
Policymakers interested in evidence-based healthcare improvements
Patients, patient advocacy groups, and the general public (for lay summaries and knowledge translation materials)
Expectation for Study Manuscript(s):
The core expectation is to prepare and submit at least one full-length manuscript for publication in a high-impact, peer-reviewed journal. Secondary analyses, if relevant, may also result in additional manuscript submissions or brief reports.
Potentially Suitable Journals:
Depending on the specifics and findings of the study, suitable journals for submission may include (please tailor the list to your precise research area):
The New England Journal of Medicine
The Lancet
Inflammatory Bowel Diseases
Journal of Crohn's and Colitis
Gastroenterology
Alimentary Pharmacology & Therapeutics
Digestive Diseases and Sciences
Gut

Supplementary Material:

https://yoda.yale.edu/wp-content/uploads/2025/06/2025-0444-Supplementary-Material_SAP_25-09-17.docx