Principal Investigator

First Name: Vivek
Last Name: Rudrapatna
Degree: MD, PhD
Primary Affiliation: University of California, San Francisco
E-mail: vivical@gmail.com
Phone number:
Address:
513 Parnassus Ave, S-357
City: San Francisco
State or Province: CA
Zip or Postal Code: 94158
Country: USA

General Information

Key Personnel (in addition to PI):
First Name: Vivek
Last name: Rudrapatna
Degree: MD, PhD
Primary Affiliation: University of California San Francisco
SCOPUS ID:

First Name: Shan
Last name: Wang
Degree: PhD
Primary Affiliation: University of San Francisco
SCOPUS ID:

First Name: Douglas
Last name: Arneson
Degree: PhD
Primary Affiliation: University of California, San Francisco
SCOPUS ID:

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: Government Funding - NIH NCATS TL1 TR001871
How did you learn about the YODA Project?: Other

Conflict of Interest

https://yoda.yale.edu/system/files/yoda_project_coi_form_for_data_requestors_2019_wang.pdf
https://yoda.yale.edu/system/files/yoda_project_coi_form_for_data_requestors_2019_var_0.pdf
https://yoda.yale.edu/system/files/yoda_project_coi_form_for_data_requestors_2019_dasigned.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. NCT00036439 - C0168T37 - A Randomized, Placebo-controlled, Double-blind Trial to Evaluate the Safety and Efficacy of Infliximab in Patients With Active Ulcerative Colitis
2. NCT00096655 - C0168T46 - A Randomized, Placebo-controlled, Double-blind Trial to Evaluate the Safety and Efficacy of Infliximab in Patients With Active Ulcerative Colitis
3. NCT00487539 - C0524T17 - A Phase 2/3 Multicenter, Randomized, Placebo-controlled, Double blind Study to Evaluate the Safety and Efficacy of Golimumab Induction Therapy, Administered Subcutaneously, in Subjects with Moderately to Severely Active Ulcerative Colitis
4. 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
5. NCT00488631 - C0524T18 - A Phase 3 Multicenter, Randomized, Placebo-controlled, Double-blind Study to Evaluate the Safety and Efficacy of Golimumab Maintenance Therapy, Administered Subcutaneously, in Subjects With Moderately to Severely Active Ulcerative Colitis
6. NCT01863771 - CNTO148UCO3001 - A Safety and Effectiveness Study of Golimumab in Japanese Patients 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

Enhancing inference from real-world data using externally-derived missing data models: a pilot study of Ulcerative Colitis

Narrative Summary:

Real-world evidence is an emerging research area that proposes to use data from non-experimental settings, such as routine clinical care, to guide better decision making. This field has received growing interest in recent years for a variety of reasons, including the realization that randomized controlled trials are too expensive and infeasible to do for every important clinical question. Despite the promise of this field, its progress has been compromised by several major limitations, one of which is the problem of missing data.
[This is a request for data access via Vivli. See attachment for full text.]

Scientific Abstract:

Background: Electronic Health Records (EHR) data are a promising source of information regarding treatment effects in the context of routine clinical care; however, their utility for research has been limited by substantial missing data. Because much of the reason for missing data is related to the availability of other corroborating information about disease activity, and this typically dictates the clinician decision to pursue additional testing and measurement, the 'missing at random' assumption (and therefore, the validity of model-based imputation) appears to be met by EHR data.

Objective: To develop and evaluate a series of missing data models using datasets with substantial completeness -- RCTs of Ulcerative Colitis -- in order to enable less biased estimation from corresponding EHR studies.

Study Design: Post-hoc analysis of individual participant data from randomized, blinded Phase 3 trials of adults with Ulcerative Colitis

Participants: Subjects participating in the above trials

Main Outcome: Outcome variables will include each of the subscores of the Mayo Score of Ulcerative Colitis

activity. We will develop and evaluate several models of missing data, and perform feature selection to identify the most informative variables for prediction.

Statistical Analysis: We will artificially censor observations from a complete data set and test a variety of popular predictive models (logistic regression, random forests, gradient boosted decision trees) according to bias and variance. We will use feature selection to identify highly informative variables.

Brief Project Background and Statement of Project Significance:

Following a search of clinicaltrials.gov, we have identified all completed, phase 2-3, randomized controlled trials of FDA-approved therapeutics for Ulcerative colitis in adults. We are requesting participant-level data corresponding to these trials.

For the primary analysis of this study, we use the Total Mayo Score as the outcome variable. This will be predicted as a function of different combinations of available Mayo subscores and auxiliary variables. We will use nested cross-validation to estimate model accuracy and variance. We will use feature-selection methods to identify reduced models that maintain high predictive accuracy, prioritizing those features that are more convenient to obtain in practice (patient- and physician-reported outcomes > blood tests > stool tests).

Finalized models in the form of as software files will be published at the end of the study; the rationale for this is that ensemble machine learning models have a complex parameterization and thus may not be easily conveyed or transported for real-world use in any other form. In addition, we will also publish the list of features found to be most informative by feature selection.

Strengths of the proposed study include its use of high-quality and complete data to address several important research problems in the field of real-world evidence. Limitations include the possibility that these models may fail to generalize to real-world contexts due to substantive reasons or other modelling problems (model misspecification, overfitting, lack of robustness, undersampling of rare strata in the included data).

Specific Aims of the Project:

There are two specific aims of this proposed research. 1) Derive and internally validate a series of models for predicting Ulcerative colitis disease activity (Total Mayo Score) in the presence of missing subscore data, and 2) identify combinations of features that are most informative for predicting the Total Mayo Score in the presence of different patterns missing subscores, with an emphasis on a missing endoscopic subscore. Exploratory aims include deriving a new composite disease activity score.
[See attached for full text]

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
Confirm or validate previously conducted research on treatment effectiveness
Preliminary research to be used as part of a grant proposal
Participant-level data meta-analysis
Participant-level data meta-analysis pooling data from YODA Project with other additional data sources

Research Methods

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

We are requesting individual participant-level data from all completed RCTs of Ulcerative Colitis in adults based on a search of clinicaltrials.gov. This decision was made in order to maximize the generalizability of these findings to that of routine clinical practice.

Main Outcome Measure and how it will be categorized/defined for your study:

For the primary analysis, the outcome element is the Total Mayo Score, an ordinal variable on a 0-12 scale. This variable has been used to define the primary outcome (typically, a binarization of this variable) of all included trials.

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

The main independent variables will be different combinations of mayo subscores. Each of these are ordinal variables on a 0-3 scale. Most of the derived models will exclude the mayo endoscopic subscore as a predictor, as this variable is least available in real-world data due to its cost and inconvenience.

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

Other variables of interest fall into the following 4 categories 1) demographic variables (gender, age, race, ethnicity), 2) disease characteristics (disease duration, disease location, current steroid use, assignment to an active arm or placebo, prior treatment failure, presence of extraintestinal manifestations where available, history of other autoimmune diagnoses where available), 3) biochemistries (hemoglobin, white count, albumin, c-reactive protein, erythrocyte sedimentation rate, fecal calprotectin), and 4) other patient reported outcomes data as available (e.g. IBDQ, SF-36). These variables will each be categorized in their native form (binary, categorical, continuous) for the purposes of modeling. Other variables that we will assess during modeling include 1) an indicator variable for trials that use central reading of endoscopy vs not, 2) an indicator variable for the trial of origin corresponding to each included data point, 3) a variable corresponding to the patient identifier (for models that allow for multiple observations per subject)

Statistical Analysis Plan:

See attachment
Software Used:
R
Project Timeline:

Start date: 4/2021
Completion date: 1/2022
Manuscript completion date: 3/2022
Results posted to YODA project: 4/2022

Dissemination Plan:

This work will be presented at national Gastroenterology meetings and will be submitted to journals of interest both to the IBD and Gastroenterology community as well as the general clinical research community: JAMA network journals, BMJ, Gastroenterology, American Journal of Gastroenterology, Inflammatory Bowel Diseases

Bibliography:

1. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest. 2020;130(2):565?574. doi:10.1172/JCI129197
2. S. van Buuren (2018). Flexible Imputation of Missing Data. Second Edition. CRC/Chapman & Hall, FL: Boca Raton

Supplementary Material:

https://yoda.yale.edu/sites/default/files/yoda_supplemental_text.docx