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

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

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: 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_mosenia.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

Missing data models for Ulcerative Colitis

Narrative Summary:

Achieving the goals of precision medicine fundamentally requires clinical datasets that are large, granular, and high-quality. Unfortunately, these characteristics are rarely found all together in any single dataset. To overcome this limitation, we propose to combine the complementary properties of datasets which are detailed and complete but covering fewer patients (randomized controlled trials; RCTs), vs datasets involving many patients but with substantial missing data (electronic health records; EHR). We will develop and evaluate missing data models using RCTs from Ulcerative Colitis and apply these on EHR data to uncover important but obscured signals about treatment efficacy/safety.

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:

Achieving the goals of precision medicine fundamentally requires clinical datasets that are large, granular, and high-quality. Unfortunately, these characteristics are rarely found all together in any single dataset. For example, randomized controlled trial (RCT) data tends to be high-quality, and well-measured with limited missingness, but often characterize small cohorts that may not reflect real-world practice. By contrast, electronic health records (EHR) data can capture large and highly relevant patient cohorts but at the expense of several limitations to the underlying data quality, especially substantial data missingness [1].

Importantly, these data are typically not 'missing completely at random': the absence of a measurement is commonly dictated by the clinical circumstances which are directly tied to outcomes of interest, therefore the imputation of the missing information becomes crucial. Obtaining unbiased estimates in the face of missing data that accords with the 'missing at random' (MAR) assumption requires the creation of a model incorporates sufficient auxiliary variables such that the unmeasured value becomes conditionally independent of the presence of missingness in the first place. Because the existence of other corroborating information about disease activity typically dictates the clinician decision to pursue additional testing and measurement, the MAR assumption (and the validity of model-based imputation) appears to be met by EHR data [2]. Considering this, it would be particularly helpful to develop a general modeling framework and to identify significant predictors for the missing information.

Following this need, to fill in the gaps in EHR data and enhance its overall utility for all secondary uses including research, we propose to use RCT data to develop and evaluate a series of missing data models that predict disease activity scores using auxiliary variables commonly present in EHR data. The completeness of RCT data will help with model selection and bias estimation. More accurate and less biased models identified from this effort may help unlock important latent information hidden in EHR data and reveal insights into real-world treatment effectiveness. We propose to pilot this concept in the setting of Ulcerative Colitis, a major subtype of Inflammatory Bowel Disease.

Specific Aims of the Project:

The primary objective of this research is to develop several models on RCT to predict missing disease activity subscores using auxiliary variables commonly available in EHR data. The disease activity score used for the assessment of Ulcerative Colitis is the Mayo Score, which consists of 4 components, of which one or two components may be missing from EHR data at any given point in time.

We will evaluate and report different model approaches to impute each Mayo subscore assuming the availability of 2-3 other Mayo subscores and biomarkers. We will compare ordinal logistic modeling, random forest for classification, gradient boosted decision trees for the prediction of the Mayo scores. We will perform feature selection for each model approach to include significant predictors, and report model accuracy by evaluating mean square errors (MSE). Selected models may help unlock important latent information hidden in EHR data and reveal insights into real-world treatment effectiveness. Therefore, to maximize the utility of these models for future studies using EHR data, we will publish them in full.

What is the purpose of the analysis being proposed? Please select all that apply.
Develop or refine statistical methods

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:

Our study will encompass models for each of the 4 outcome variables corresponding to different components of the Mayo Score: stool frequency, rectal bleeding, physician global assessment, and endoscopic subscore. Each of these variables are ordinal variables graded on a 0-3 score.

.misc-fixes { display: none; } #admin-region { z-index: 9999999; } #admin-menu { z-index: 99999999; }

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

For each mayo subscore, predictor variables will include each of the other mayo subscores, c-reactive protein, fecal calprotectin, erythrocyte sedimentation rate, white blood cell count, hemoglobin, albumin, platelet count, ferritin, history of prior medication failure, and concurrent oral steroid use.

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

All key auxiliary variables are as listed above. We will additionally explore other variables such as gender, age, and ethnicity and determine their utility for Mayo score imputation.

Statistical Analysis Plan:

Hypothesis testing: This study does not intend to test statistical hypotheses; however the underlying motivation for this study is it is possible to build models using auxiliary variables shared between EHR and RCT datasets that are characterized by minimal bias and acceptable variance to enable downstream analyses.

Sample size considerations: None planned, since this is a post-hoc analysis of published trial data.

Development of predictive models: We will separately use data elements collected/reported at baseline and at follow-up to guide model development. Using complete case data, we will censor data elements according to a missing at random mechanism and then explore modeling approaches to impute these elements. The variables to be estimated are the four components of the Mayo Score. Each variable is an ordinal variable which may be estimated by a classification modeling approach. Several popular models for classification, such as ordinal logistic modeling, random forest for classification, and gradient boosted decision tree, will be evaluated. Models that are sufficiently accurate will be published in order to enable future analyses done using EHR data.

We will perform a similar process at the time of follow-up when most of the included trials include the full Mayo Score (which requires endoscopy). This may be at week 24, possibly also week 12. We anticipate some missing data at this time related to subject dropout and/or adverse events. We will repeat the above steps using a complete case analysis. For these models, we repeat the process using placebo or individual drugs separately as well as taking all data together.

Exploratory models will include several deep neural network architectures including multilayer perceptrons and recurrent neural networks.

Model assessment: Model accuracy, bias, and variance will be measured against the underlying censored values. We will report both accuracy and mean squared error.

We will compare differences in predictions between models developed from baseline data with those at the time of follow-up (i.e. assess the stability of variable relationships at different times following treatment exposure). As an exploratory analysis we will also assess the stability of these models at follow-up and determine whether or not the grouping of placebo/active arms is justified (i.e. whether or not the presence of a drug affects the relationships between variables).

Following model evaluation, selected models will be compressed and exported from the computing environment to be made publicly available at the time of publication. This will be necessary as many of the ‘black box’ models being used here (e.g. random forests) cannot be fully specified in a written form nor can they be re-used by the general research community unless they are provided as software files.

Quality Control: Exploratory data analysis will be performed to confirm accuracy of data entry and identify outlying observations.

Missing data: For the analysis that uses baseline covariates we will assume near complete data availability and remove any rare data points with any missing observations at baseline. These data will then be censored according to the above mechanism prior to the development and evaluation of the missing data.

Software: To maximize generalizability and re-use by the clinical research community, we will fit models in both R

and Python computing environments and export model software objects from both.
Software Used:
R

Project Timeline:

Start date: 10/2020
Completion date: 6/2021
Manuscript completion date: 8/2021
Results posted to YODA project: 8/2021

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