Principal Investigator

First Name: Giovanni
Last Name: Parmigiani
Degree: PhD
Primary Affiliation: Dana Farber Cancer Institute and Harvard University
E-mail: yujiewu@g.harvard.edu
State or Province: Massachusetts
Country: United States

General Information

Key Personnel (other than PI):
First Name: Boyu
Last name: Ren
Degree: PhD
Primary Affiliation: McLean Hospital
SCOPUS ID:
Requires Data Access? No

First Name: Yujie Wu
Last name: Wu
Degree: Master's
Primary Affiliation: Harvard University
SCOPUS ID:
Requires Data Access? Yes

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?: Scientific Publication

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2024/01/COI_FORM_BR.pdf
https://yoda.yale.edu/wp-content/uploads/2024/01/COI_FORM_GP.pdf
https://yoda.yale.edu/wp-content/uploads/2024/01/COI_FORM_YW.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. NCT00334126 - A Randomized, Double-blind, Placebo-controlled, Parallel Group Study to Evaluate the Efficacy and Safety of Paliperidone ER Compared to Quetiapine in Subjects With an Acute Exacerbation of Schizophrenia
2. NCT00085748 - A Randomized, 6-Week Double-Blind, Placebo-Controlled Study With an Optional 24-Week Open-Label Extension to Evaluate the Safety and Tolerability of Flexible

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

Research Proposal

Project Title

Multi-study causal inference for robust clinical outcome predictions

Narrative Summary:

Previous study by Chekroud et al. (2024) that used the data sets from YODA (NCT00518323, NCT00334126, NCT00085748, NCT00078039, and NCT00083668) to predict treatment outcomes in schizophrenia found that even though the machine learning models may have high prediction accuracy on the data set where the model was trained, the out-of-study prediction accuracy decreased drastically, suggesting limited generalizability of the machine learning models. One potential reason for the worse out-of-study prediction accuracy is due to the population heterogeneity, and therefore the machine learning model only learned context-dependent relationship that is sensitive to the shift in study population.

One solution for robust clinical outcome prediction is through causal inference, where previous studies have found that causal relationships are less sensitive to the shift in data distribution across studies, and thus the causal models can lead to more robust predictions on external data sets. In particular, we would like to apply the multi-study R learner for heterogeneous treatment effect estimation that is robust to between-study heterogeneity and achieve more accurate out-of-study predictions.

Scientific Abstract:

Background:
Applying machine learning models in healthcare has become a heated topic, particularly for making clinical predictions. Precise outcome prediction has important clinical benefits as it directly impacts treatment decision-making and benefits patients' health. For predicting models, generalizability is defined as the machine learning model's ability to make accurate predictions on a new/external/independent data sets [3]. If a model does not have the generalizability property, then it would be hard to reproduce results and make predictions for general use.
A recent study has made use of the five trials from YODA: NCT00518323, NCT00334126, NCT00085748, NCT00078039, and NCT00083668, to make predictions for treatment outcomes in schizophrenia. The primary outcome is the Remission in Schizophrenia Working Group criteria (RSWG) and predictors includes all information available at baseline across all trials such as age, race [1]. Their results showed that the machine learning models failed to generalize to unseen, external trials that were not included when training the prediction models. One possible explanation

is the study heterogeneity between different clinical trials where the machine learning models were trained to only extract study-specific information and thus having difficulty to generalize to a new data set, where the data distribution of the new study population might be different.

Objective:
Our project will try to address this issue through causal predictions, where causal relationships between predictors and outcomes are often stable between different studies, and therefore have greater potential to generalize well to an unseen study, making accurate treatment outcome predictions to facilitate clinical treatment decision-making.

Study Design:
We will combine the multiple trials together and perform meta-analysis.

Participants:
Yujie Wu, Ph.D. student in Biostatistics, Department of Biostatistics, Harvard University
Boyu Ren, Ph.D., Instructor in Psychiatry, McLean Hospital, Harvard Medical School
Giovanni Parmigiani, Ph.D., Professor of Biostatistics, Department of Data Science, Dana Farber Cancer Institute

Primary and Secondary Outcome Measure(s):
Following Chekroud et al. (2024), we will use the symptomatic outcomes based on the the Positive and Negative Syndrome Scale (PANSS), which is used to measure treatment outcomes. The primary outcome of interest is the Remission in Schizophrenia Working Group criteria (RSWG), which is a transformation of PANSS.
The secondary outcomes include: 25% symptom reduction (binary), 50% symptom reduction (binary) and Baseline-adjusted percent change in symptoms (continuous).

Statistical Analysis:
We will first impute the missing data for each predictor variable by the median values, and run the descriptive analysis on the baseline covariates. Mean will be reported for continuous variables and percentage will be reported for categorical variables. Standard deviation will be reported as a measure of variability of the variables.
We then will replicate the results presented in Chekroud et al. (2024) by using the elastic net algorithm and the tuning parameters will be selected through the 10-fold cross-validation.
We will apply the multi-study R-learner on the trials to estimate the robust causal relationships and make comparisons on the prediction accuracy with the elastic net algorithm.

Brief Project Background and Statement of Project Significance:

Applying machine learning models in healthcare has become a heated topic, particularly for making clinical predictions. Precise outcome prediction has important clinical benefits as it directly impacts treatment decision-making and benefits patients' health. For predicting models, generalizability is defined as the machine learning model's ability to make accurate predictions on a new/external/independent data sets [3]. If a model does not have the generalizability property, then it would be hard to reproduce results and make predictions for general use.

A recent study has made use of the five trials from YODA: NCT00518323, NCT00334126, NCT00085748, NCT00078039, and NCT00083668, to make predictions for treatment outcomes in schizophrenia. The primary outcome is the Remission in Schizophrenia Working Group criteria (RSWG) and predictors includes all information available at baseline across all trials such as age, race [1]. Their results showed that the machine learning models failed to generalize to unseen, external trials that were not included when training the prediction models. One possible explanation is the study heterogeneity between different clinical trials where the machine learning models were trained to only extract study-specific information and thus having difficulty to generalize to a new data set, where the data distribution of the new study population might be different.

Our project will try to address this issue through causal predictions, where causal relationships between predictors and outcomes are often stable between different studies, and therefore have

greater potential to generalize well to an unseen study, making accurate treatment outcome predictions to facilitate clinical treatment decision-making.

Specific Aims of the Project:

In this project, we apply the multi-study R learner for heterogeneous treatment effect estimation that is robust to between-study heterogeneity [2] to make predictions for treatment outcomes in schizophrenia. We will compare prediction accuracy from our model with traditional machine learning models used by Chekroud et al. (2024).

Study Design:

Meta-analysis (analysis of multiple trials together)

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

Participant-level data meta-analysis

Meta-analysis using only data from the YODA Project

Develop or refine statistical 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:

The data Source and inclusion/exclusion criteria follow the paper by Chekroud et al. (2024). We will use treatment data from five international, multisite RCTs (NCT00518323, NCT00334126, NCT00085748, NCT00078039, and NCT00083668). We will exclude participants that did not have a follow-up at the 4-week point after study enrollment.

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

Following Chekroud et al. (2024), we will use the symptomatic outcomes based on the the Positive and Negative Syndrome Scale (PANSS), which is used to measure treatment outcomes. The primary outcome of interest is the Remission in Schizophrenia Working Group criteria (RSWG), which is a transformation of PANSS.

The secondary outcomes include: 25% symptom reduction (binary), 50% symptom reduction (binary) and Baseline-adjusted percent change in symptoms (continuous).

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

The main predictor variable is the treatment randomization indicating whether the patients were randomized to an antipsychotic medication or placebo.

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

The other predictors in the model include basic demographic features, psychiatric history (DSM-IV diagnosis category, age of diagnosis, psychiatric hospitalizations), clinical data (PANSS, Clinical Global Impression), extrapyramidal symptom scales (Abnormal Involuntary Movement Scale) and

Simpson Angus Scale, biometric data (blood chemistry panel, hematology, urinalysis).

Statistical Analysis Plan:

We will first impute the missing data for each predictor variable by the median values, and run the descriptive analysis on the baseline covariates. Mean will be reported for continuous variables and percentage will be reported for categorical variables. Standard deviation will be reported as a measure of variability of the variables.

We then will replicate the results presented in Chekroud et al. (2024) by using the elastic net algorithm and the tuning parameters will be selected through the 10-fold cross-validation.

We will apply the multi-study R-learner on the trials to estimate the robust causal relationships and make comparisons on the prediction accuracy with the elastic net algorithm.

Software Used:

RStudio

Project Timeline:

We will begin analyzing the data upon approval of the data sets. The expected analysis completion date will be the end of July 2024, and the manuscript will be drafted and submitted in August 2024. The date when the results are reported back to the YODA Project is expected to be by December 2024.

Dissemination Plan:

We will develop an R package for public use. We will target our project to researchers who work on applying machine learning models in health care, particularly for those with multiple data sets, informing researchers the importance of generalizability of machine learning models. Suitable journals for submission include Biostatistics, Statistics in Medicine, Bioinformatics.

Bibliography:

[1] Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., ... & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science*, *383*(6679), 164-167.

[2] Shyr, C., Ren, B., Patil, P., & Parmigiani, G. (2023). Multi-study R-learner for Heterogeneous Treatment Effect Estimation. *arXiv preprint arXiv:2306.01086*.

[3] Yang, J., Soltan, A. A., & Clifton, D. A. (2022). Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *npj Digital Medicine*, *5*(1), 69.