Principal Investigator

First Name: Gideon
Last Name: Vos
Degree: Doctor of Philosophy
Primary Affiliation: James Cook University
E-mail: gideon.vos@my.jcu.edu.au
State or Province: QLD
Country: Australia

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?: Scientific Publication

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2024/01/SV_57KskaKADT3U9Aq-R_40Z2AObcUFIboa2.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. NCT00518323 - A Randomized, Multicenter, Double-Blind, Weight-Based, Fixed-Dose, Parallel-Group, Placebo-Controlled Study of the Efficacy and Safety of Extended Release Paliperidone for the Treatment of Schizophrenia in Adolescent Subjects, 12 to 17 Years of Age
2. 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
3. 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 Doses of Paliperidone Extended Release in the Treatment of Geriatric Patients With Schizophrenia
4. NCT00083668 - A Randomized, Double-blind, Placebo- and Active-controlled, Parallel-group, Dose-response Study to Evaluate the Efficacy and Safety of 3 Fixed Dosages of Paliperidone Extended Release (ER) Tablets and Olanzapine, With Open-label Extension, in the Treatment of Patients With Schizophrenia
5. NCT00078039 - Trial Evaluating Three Fixed Dosages of Paliperidone Extended-Release (ER) Tablets and Olanzapine in the Treatment of Patients With Schizophrenia

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

Research Proposal

Project Title

Generalizability of clinical prediction models using synthetic data generated through random sampling

Narrative Summary:

Prior studies (https://doi.org/10.1126/science.adg8538) have reported that generalization of machine learning models applied to clinical prediction of Schizophrenia suffer from low reproducibility when applied to new, unseen data collected from trials or study groups not represented in the original training data used when training the machine learning models. Our prior studies using sensor biomarker data obtained from wearable devices for predicting acute stress response have shown promise in dealing with this reproducibility problem by building synthetic datasets constructed from a large collection of individual trials (https://doi.org/10.1016/j.jbi.2023.104556). This study aims to test using the latter approach of synthetic data generation on the clinical data used in the first study (https://doi.org/10.1126/science.adg8538) to validate the synthetic data generation approach beyond that of acute stress prediction. We aim to test our approach that produced good results on time-series sensor biomarker data on this clinical trial data. The prior clinical study publication (https://doi.org/10.1126/science.adg8538) were unable to build a model that can generalize on unseen data, a problem very common in stress prediction using wearable sensor biomarker data. Our approach of synthetic data generation worked well on wearable sensor biomarker data, and we aim to examine whether this method is transferable and usable on clinical trial data. We will use the exact same code, methods and statistical validation methods used in study (https://doi.org/10.1126/science.adg8538) by adding our synthetic data generation method within. We can only test this by using the exact same trial data used by study (https://doi.org/10.1126/science.adg8538).

Scientific Abstract:

Background;
The use of machine learning to measure and predict acute stress response has been gaining increasing attention over the last decade. This has largely been driven by the introduction and availability of low-cost wearable devices geared towards both research and the consumer market, along with a movement towards personal health understanding generally refered to as the quantified-self. A number of stress-related studies have made their design, methods, algorithms and parameters available via scientific publication, and in a number of cases these studies were performed on wearable biomarker datasets that are in the public domain. The generalization ability of these models, when trained on data from single study settings to be able to accurately measure and predict on new, unseen biomarker data remains unproven, and several approaches have been proposed to address this concern. Recent studies have raised similar questions with regards to the generalization of models when predicting treatment outcomes across independent clinical trials.

Objective;
In this study, our objective is to assess the effectiveness and reliability of a synthetic data generation methodology that we have previously proven successful in improving the generalisability of machine learning in predicting acute stress in new unseen stress datasets. We will apply our methodology to the requested YODA project's independent clinical trial. The study seeks to validate our method's performance, generalisability and robustness beyond the time series stress biomarker data types.

Study Design;
Clinical data from five prior trials (NCT00518323, NCT00334126, NCT00085748, NCT00083668, NCT00078039) available through the YODA Project will be sourced to reproduce prior study findings that tested the generalization of machine learning models built to predict treatment outcomes of antipsychotic medication for the treatment of schizophrenia. The clinical trial data will be engineered to enable the generation of synthetic patient outcome data through randomized sampling as

previously applied to time-series data of acute stress response. The existing source code provided publicly for the treatment outcome study will be modified to accept the synthetic data, as applicable. Existing machine learning models provided in the source-code will be re-run, with existing statistical analysis applied as per the existing study to compare predictive accuracy and model generalization, when compared to the models trained on the data from the five prior trials as-is.

Participants;
Treatment data from five international, multisite RCTs (NCT00518323, NCT00334126, NCT00085748, NCT00078039, and NCT00083668) will be requested through the YODA Project (https://yoda.yale.edu/). This data was sourced between 29 March 2004 to 30 March 2009, providing a total of 1962 patients aged 12 to 81 years across five randomized controlled trials.

Primary and Secondary Outcome Measure(s);
While the primary outcome within the trial data is the Remission in Schizophrenia Working Group criteria (RSWG) score, the primary outcome in this particular study will be the validation of the use of synthetically generated data for training machine learning models to generalize on new, unseen data generated from clinical trials not included in the original training data. A secondary outcome measure will be to test whether this approach provide higher predictive accuracy across the same statistical measures previously reported.

Statistical Analysis
Balanced accuracy [(sensitivity + specificity) / 2] will be the primary statistical measure, and this will be applied across within-trial with no validation, within-trial with cross-validation, paired-trial and leave one trial out test sets.

Brief Project Background and Statement of Project Significance:

Machine learning model generalization and the reproduction of result findings is of great importance to research within the area of artificial intelligence. A number of prior studies have reported high predictive accuracy of machine learning models; however, these were often tested on single study datasets with small sample sizes [2]. To address the problem of generalization, a number of techniques have been proposed including synthetic data generation to artificially enlarge training datasets and the use of ensemble methods [3]. These techniques were applied to sensor biomarker data obtained using personal wearable devices for acute stress prediction, and whether these same techniques can address generalization in other health care studies remains outstanding. This study aims to apply these techniques on datasets previously used in studies [3] where generalization was problematic in order to study whether those techniques can be applied in a wider range of experimental conditions and clinical outcomes.

Specific Aims of the Project:

The aim of this project is to address the growing concern around machine learning study results and reproducibility of those results. Prior publications have addressed this concern for specific domains (acute stress) using data collected using wearable electronic monitors (Empatica E4), however it would be of great interest to researchers if these same techniques can be applied on other clinical studies, to address the same concerns with regards to machine learning generalization.

Study Design:

Individual trial analysis

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

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:

Data from a prior study will be utilized [1]:

Data accession numbers for the
five trials analyzed here are: R076477-PSZ-3001 (Teens trial),
R076477-SCH-3015 (Adults First Episode), R076477-SCH-302
(Older Adults), R076477-SCH-305 (Adults Chronic #1), R076477-
SCH-303 (Adults Chronic #2).

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

The primary outcome is the Remission in Schizophrenia Working Group criteria (RSWG), secondary is treatment assignment: 'paliperidone vs. placebo', to match prior study [1].

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

Remission in Schizophrenia Working Group criteria will be the main predictor variable, as per prior study.

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

In order to reproduce the prior study [1], 217 variables will be utilized including basic demographic features, psychiatric history (DSM-IV diagnosis category, age of diagnosis, psychiatric hospitalizations), clinical data (PANSS, Clinical Global Impression) (17), extrapyramidal symptom scales (Abnormal Involuntary Movement Scale) (19) and Simpson Angus Scale (20), biometric data (blood chemistry panel, hematology, urinalysis), and treatment randomization.

Statistical Analysis Plan:

In line with prior study [1], within-trial cross validation, leave one trial out and paired trial cross-validation will be utilized to measure balanced predictive accuracy.

Software Used:

R

Project Timeline:

Study will commence 1 February 2024, pending available of the data we are requesting, and aims to complete by July 2024 (analysis completion). Manuscript will then be drafted with the aim of first submission by October 2024, with results reported back to YODA upon manuscript acceptance.

Dissemination Plan:

If the study is successful, a manuscript will be prepared for submission to Science as a follow up to prior study [1].

Bibliography:

[1] Frederike H. Petzschner,Practical challenges for precision medicine, Science, 383, 6679, (149-150), (2024). /doi/10.1126/science.adm9218

[2] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, Mostafa Rahimi Azghadi,Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review,International Journal of Medical Informatics,Volume 173,2023,105026,ISSN 1386-5056,https://doi.org/10.1016/j.ijmedinf.2023.105026.

[3] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, Mostafa Rahimi Azghadi, Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices, Journal of Biomedical Informatics, Volume 148,2023,104556,ISSN 1532-0464,https://doi.org/10.1016/j.jbi.2023.104556.