Principal Investigator

First Name: Nicole
Last Name: Cizauskas
Degree: MSc in Health Data Analytics and Machine Learning from Imperial College London
Primary Affiliation: Newcastle University (PhD student)
E-mail: n.cizauskas2@newcastle.ac.uk
State or Province: Newcastle-Upon-Tyne
Country: United Kingdom

General Information

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: MRC-NIHR Trials Methodology Research Partnership (TMRP)
How did you learn about the YODA Project?: Colleague

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2024/01/COI_Form_YODA.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. NCT00638690 - A Phase 3, Randomized, Double-Blind, Placebo-Controlled Study of Abiraterone Acetate (CB7630) Plus Prednisone in Patients With Metastatic Castration-Resistant Prostate Cancer Who Have Failed Docetaxel-Based Chemotherapy
2. NCT00887198 - A Phase 3, Randomized, Double-blind, Placebo-Controlled Study of Abiraterone Acetate (CB7630) Plus Prednisone in Asymptomatic or Mildly Symptomatic Patients With Metastatic Castration-Resistant Prostate Cancer
3. NCT01715285 - A Randomized, Double-blind, Comparative Study of Abiraterone Acetate Plus Low-Dose Prednisone Plus Androgen Deprivation Therapy (ADT) Versus ADT Alone in Newly Diagnosed Subjects With High-Risk, Metastatic Hormone-naive Prostate Cancer (mHNPC)
4. NCT02236637 - A Prospective Registry of Patients With a Confirmed Diagnosis of Adenocarcinoma of the Prostate Presenting With Metastatic Castrate-Resistant Prostate Cancer
5. NCT02065791 - A Randomized, Double-blind, Event-driven, Placebo-controlled, Multicenter Study of the Effects of Canagliflozin on Renal and Cardiovascular Outcomes in Subjects With Type 2 Diabetes Mellitus and Diabetic Nephropathy
6. NCT01989754 - A Randomized, Multicenter, Double-Blind, Parallel, Placebo-Controlled Study of the Effects of Canagliflozin on Renal Endpoints in Adult Subjects With Type 2 Diabetes Mellitus
7. NCT01032629 - A Randomized, Multicenter, Double-Blind, Parallel, Placebo-Controlled Study of the Effects of JNJ-28431754 on Cardiovascular Outcomes in Adult Subjects With Type 2 Diabetes Mellitus
8. NCT01809327 - A Randomized, Double-Blind, 5-Arm, Parallel-Group, 26-Week, Multicenter

Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in Combination With Metformin as Initial Combination Therapy in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control With Diet and Exercise
9. NCT00968812 - A Randomized, Double-Blind, 3-Arm Parallel-Group, 2-Year (104-Week), Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of JNJ-28431754 Compared With Glimepiride in the Treatment of Subjects With Type 2 Diabetes Mellitus Not Optimally Controlled on Metformin Monotherapy
10. NCT01106677 - A Randomized, Double-Blind, Placebo and Active-Controlled, 4-Arm, Parallel Group, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control on Metformin Monotherapy
11. NCT00460512 - An Open-label Prospective Trial to Explore the Tolerability, Safety and Efficacy of Flexibly Dosed Paliperidone ER in Subjects With Schizophrenia
12. NCT00589914 - A Randomized, Double-Blind, Parallel-Group, Comparative Study of Flexible Doses of Paliperidone Palmitate and Flexible Doses of Risperidone Long-Acting Intramuscular Injection in Subjects With Schizophrenia
13. NCT01281527 - A 6-month, Open Label, Prospective, Multicenter, International, Exploratory Study of a Transition to Flexibly Dosed Paliperidone Palmitate in Patients With Schizophrenia Previously Unsuccessfully Treated With Oral or Long-acting Injectable Antipsychotics
14. NCT01515423 - A Randomized, Multicenter, Double-Blind, Non-inferiority Study of Paliperidone Palmitate 3 Month and 1 Month Formulations for the Treatment of Subjects 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

Using Synthetic Controls to Improve Randomised Controlled Trials for Rare Diseases

Narrative Summary:

Treatments for rare diseases often struggle to run the gold standard clinical trial methodology, randomized control trials (RCTs). This is due to a lack of participants available for a well-powered comparison between experimental and control arms together with ethical restrictions assigning patients to control arms (Thorlund et al., 2020).
Synthetic control arms are control groups generated based on real-world patient data with similar attributes to the experimental group (Bouttell et al., 2018). They are typically designed based on previous RCT data, observational study data, or external data (e.g. electronic health records) (Thorlund et al., 2020). Synthetic control arms are particularly useful in clinical trials that would otherwise be restricted by participant recruitment, cost, or ethics. Several previous clinical trial studies have utilized synthetic controls successfully (Berry et al., 2017; Blondeau, Schneider and Ngwa, 2020; Ko et al., 2021). The FDA (U.S. Food and Drug Administration) has approved drug RCTs that utilized synthetic control arms in recent years (Goldsack, 2019; Commissioner, 2020; Gretton, 2021). Using synthetic control arms for rare disease RCTs would address issues of lack of participants and ethical restrictions, however there is often no previous data to help generate synthetic controls from (Prasad, 2021).
The goal of this study is to create a methodology that can be used to generate pools of synthetic controls for rare diseases by training a machine learning algorithm to predict control arms from experimental arms of similar studies.
The first aim of the study looks at the three ways of generating synthetic controls (using previous RCT data, observational study data, or external data) to determine the difference in accuracy between them. Previous work has highlighted likely differences in the quality of these methods (Thorlund et al., 2020), but this has yet to be quantified using real world data. Historical RCTs with

both an experimental and control arm will be used for this analysis. Synthetic control arms will be generated using each of the three methods, and compared to the real control arm used in the studies. The difference in accuracy, correlation, and bias will be calculated for each method. The purpose of this is to understand how much data quality will be lost when using a non-optimal data source; if the data quality loss is minimal, this study will proceed using an expanded criteria for data sources.
The next aim uses previous RCT summary level data to train a machine learning model that predicts control arms from experimental arms. Using the software Kerus (Flynn, 2015), simulation datasets based on summary level data can be generated for the purpose of training (McCartan et al., 2021). The use of simulation data is necessary, as this model will likely require a large amount of data from a variety of RCTs. This model will be validated and tested on with historical RCT data.
Using this model, experimental arms matched by similarity to the rare disease can be fed in to predict an accurate control arm. Similarity in this instance refers to trials matched on the genotype and symptoms of the target disease, and/or characteristics of intended treatment group.
If successful, this model will provide a new method of generating synthetic controls for rare disease RCTs where there is currently no other method of doing so.

Scientific Abstract:

1. Background
It is often difficult to execute a RCT for a rare disease treatment due to lack of participants, ethics, and cost (Thorlund et al., 2020). Synthetic control arms could potentially be used in place of placebo or control arms in these RCTs.

2. Objective
There are two aims of this study: first, to determine which types of historical data are acceptable in generating synthetic controls and what the difference in accuracy is between the methods, and second, to create a machine learning model that can predict control arms given historical experimental arms.

3. Study Design
This is a methdological study with three parts.

4. Participants
For the first aim, historical RCT trials with IPD (individual participant data) will be used. Separate synthetic control arms will be generated based on three different types of historical data (previous RCTs, observational study data, and external data) (external data, observational studies, and previous RCTs) and compared to the real historical control arm. For the second aim, simulated datasets of control arms and experimental arms based on publicly available summary level RCT data will be used to train the machine learning model. Historical RCTs will also be recruited from all available sources for the testing of the model.

5. Primary/Secondary Outcome Measure(s)
This project is focused on developing a methodology. The use of RCT data is to ensure that the methodology works on real RCT data in practice. As such, there is not primary and secondary outcomes. The RCT data can contain any combination of primary or secondary outcomes: what is relevant for the scope of this project is the demographic information (sex, ethnicity, age, e.t.c.) and the arm assigned (control or experimental). Any publications would report summary level statistics of the demographic information and arm assignment.

6. Statistical Analysis
In the first aim, existing packages will be used to generate synthetic controls from previous RCTs, observational studies, and external data. Synthetic controls will be compared to historical control arms and experimental arms using Pearson’s correlation coefficient. In the second aim, a generative adversarial network (GAN) deep learning model will be used to generate control data from input dataset. The generated data will be compared to the test data using Pearson’s correlation

coefficient.

Brief Project Background and Statement of Project Significance:

It is often difficult to execute a RCT for a rare disease treatment due to lack of participants, ethics, and cost (Thorlund et al., 2020). Synthetic control arms could potentially be used in place of placebo or control arms in these RCTs.
Synthetic controls have been successful in historical trials. One phase 1-2 trial looking at establishing novel indicators of survival in acute myeloid leukaemia used a data pool of historical controls from Medidata's archive of trial data (Berry et al., 2017); this study was able to analyse its primary outcomes of complete remission, complete remission without hematologic recovery, and overall survival using their single-arm trial combined with a synthetic control arm.
Another previous study used observational data to generate a synthetic control arm matched to a clinical trial population (Blondeau, Schneider and Ngwa, 2020). Propensity scores were used to match participants in the experimental arm to controls in PharMetrics, a US claims database. A sufficient number of matched participants were selected to allow a well-powered comparison.
A recent study focused on using Electronic Health Records (EHRs) to develop a synthetic control arm (Ko et al., 2021); this study looked at the effect of a lifestyle intervention on cardiovascular health metrics, and use propensity scores to match controls from EHRs to participants in a single-arm trial. The use of EHRs allowed for a follow-up on the same control arm participants 5 years later. Overall, external data from EHRs had mixed reception, with some measurements being more useful than others in analysis. This method is thought to be more prone to bias than using RCT or prospective observational data.
The use of synthetic control arms in RCT is gaining more reception, with recent approvals from the FDA featuring synthetic controls in their analyses (Goldsack, 2019; Commissioner, 2020; Gretton, 2021). However, it is still difficult to provide a synthetic control arm for rare disease RCTs, as there is a lack of previous RCT, observational, and external data (Prasad, 2021). This project aims to fill that gap by developing a methodology for rare disease single-arm trials to generate a synthetic control arm using machine learning algorithms training on predicting control arms from experimental arms in previous RCTs.

Specific Aims of the Project:

Aim 1: Determine which of current three synthetic control arm methods, using previous RCT data, observational data, or external data, are viable for creating RCT controls
Objectives:
• Assess which of the three methods result in statistically indistinguishable controls
• Compare the bias of any successful methods
• Compare the accessibility and ease of access of any successful methods
Hypothesis:
• Synthetic controls created from previous RCT data will have significantly higher accuracy compared to the real control arm than the other methods.
Aim 2: Generate control arms using ML/DL methods by training on previous, related RCT data
Objectives:
• Create a ML model that predicts control arms from experimental groups
• Test and validate the model
• Use this model to create RCT control libraries for rare diseases
Hypothesis:
• A ML model can generate statistically indistinguishable control arms from real control arms.

Study Design:

Methodological research

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

Develop or refine statistical methods

Research on clinical trial methods

Research Methods

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

This project is focused on developing a methodology. The use of RCT data is to ensure that the methodology works on real RCT data in practice. As such, there is not explicit inclusion/exclusion criteria at the patient level. However, there is such at the dataset level. The below criteria is therefore measured at the dataset level, not the participant level.
Inclusion: RCT datasets with separate experimental and control arms; RCT datasets related to drugs/diseases that have multiple available RCT study datasets to compare to
Exclusion: Patients who dropped out partway (only complete cases are taken).

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

This project is focused on developing a methodology. The use of RCT data is to ensure that the methodology works on real RCT data in practice. As such, there is not primary and secondary outcomes. The RCT data can contain any combination of primary or secondary outcomes: what is relevant for the scope of this project is the demographic information (sex, ethnicity, age, e.t.c.) and the arm assigned (control or experimental). Any publications would report summary level statistics of the demographic information and arm assignment.

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

Aim 1: Comparing the accuracy of different methods for generating synthetic controls
• Independent variable: type of data used in synthetic control generation
• Dependent variable: accuracy and correlation of the synthetic control to the actual control
Aim 2: Creating a ML model to predict control arms from experimental arms
• Independent variable: training data used/model specifications such as epoch length, batch size, loss function
• Dependent variable: ML model’s accuracy/bias

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

The control arms will be compared based on a number of common demographic information, such as sex, ethnicity, age, and previous health conditions where appliable.
Relative change (as a percentage) from baseline will be used to access the outcome. This measurement was chosen as it can reflect magnitude without units, and so the model can accurately understand data from studies with different outcome measurements.

Statistical Analysis Plan:

The requested clinical data from YODA will be used in both aim 1 and aim 2 of this study. The programming languages R and Python will be used to complete this analysis, likely in the programs R Studio and Jupyter Lab respectively.
In aim 1, key characteristics (sex, ethnicity, age) in selected YODA trials will be used to generate synthetic controls. Synthetic controls will be made using the R packages Tidysynth (Dunford, 2023) and Scpi (Cattaneo et al., 2022). Across all trials, the average outcomes in synthetic controls and actual control arms will be compared using Pearson’s Correlation Coefficient. For this analysis, RCT studies need to be paired with other RCTs, observational studies, and external data for similar drugs/diseases.

In aim 2, RCT data from YODA will be used in the testing of a GAN (generative adversarial network) deep learning model training on a combination of simulated datasets and real world RCT datasets. GAN models are made of two components: a generator and a discriminator (Goodfellow et al., 2020; Arora and Arora, 2022). They work by generating a random sequence and using the discriminator to compare this to the chosen input. The model compares the generated data to the input data and then repeats the process while trying to improve the similarity. This can run for any epochs until the model can successfully create generated data that is similar in key properties to the input data. In the test set, experimental or single arms will be used. This synthetic control arm will be compared to the trial control or placebo arm and the experimental arm using Pearson's correlation coefficient. In order to input the data into the GAN model, the data must be cleaned and one-hot-encoded to appear in numerical form. Demographic data and relative change from baseline are the two key categories of data being looked at.
Examples of the data cleaning process and GAN modelling can be found on an example dataset in the Github repository linked below:
(link to data cleaning procedure and example GAN model:
https://github.com/N-cizauskas/GAN-for-Synthetic-Controls)

Rare disease data will be obtained later in the project and tested on an improved version of the model.

Software Used:

Python

Project Timeline:

I plan to complete Aim 1 within 12 months of receiving the data. The duration of Aim 2 is dependent on how long it takes to acquire the other necessary datasets, specifically those for training the machine learning algorithm. This may exceed 12 months.

Dissemination Plan:

Project results will be disseminated via journal publication and at conference presentations. Potential journals include:
• Potential conferences include: PSI (Statisticians in the Pharmaceutical Industry), ICTMC (International Clinical Trials Methodology Conference), and others
• Potential journals include: Pharmaceutical Statistics Journal, Contemporary Clinical Trials, International Journal of Clinical Trials, and others

Bibliography:

Arora, Anmol and Arora, Ananya (2022) 'Generative adversarial networks and synthetic patient data: current challenges and future perspectives', *Future Healthc J*, 9(2), pp. 190–193. Available at: https://doi.org/10.7861/fhj.2022-0013.

Berry, D.A. *et al.* (2017) 'Creating a synthetic control arm from previous clinical trials: Application to establishing early end points as indicators of overall survival in acute myeloid leukemia (AML).', *Journal of Clinical Oncology*, 35(15_suppl), pp. 7021–7021. Available at: https://doi.org/10.1200/JCO.2017.35.15_suppl.7021.

Blondeau, K., Schneider, A. and Ngwa, I. (2020) 'A synthetic control arm from observational data to estimate the background incidence rate of an adverse event in patients with Alzheimer's disease matched to a clinical trial population', *Alzheimer's & Dementia*, 16(S10), p. e043657. Available at: https://doi.org/10.1002/alz.043657.

Bouttell, J. *et al.* (2018) 'Synthetic control methodology as a tool for evaluating population-level health interventions', *J Epidemiol Community Health*, 72(8), pp. 673–678. Available at: https://doi.org/10.1136/jech-2017-210106.

Cattaneo, M.D. *et al.* (2022) 'scpi: Uncertainty Quantification for Synthetic Control Methods'. arXiv. Available at: https://doi.org/10.48550/arXiv.2202.05984.

Commissioner, O. of the (2020) *Statement from FDA Commissioner Scott Gottlieb, M.D., on FDA's new strategic framework to advance use of real-world evidence to support development of drugs and biologics*, *FDA*. FDA. Available at: https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-fdas-new-strategic-framework-advance-use-real-world (Accessed: 2 October 2023).

Dunford, E. (2023) 'tidysynth: A Tidy Implementation of the Synthetic Control Method'. Available at: https://cran.r-project.org/web/packages/tidysynth/index.html (Accessed: 19 October 2023).

Flynn, A. (2015) 'Exploristics', *Personalized Medicine*, 12(6), pp. 537–540. Available at: https://doi.org/10.2217/pme.15.32.

Goldsack, J. (2019) 'Synthetic control arms can save time and money in clinical trials', *STAT*, 5 February. Available at: https://www.statnews.com/2019/02/05/synthetic-control-arms-clinical-trials/ (Accessed: 2 October 2023).

Goodfellow, I. *et al.* (2020) 'Generative adversarial networks', *Communications of the ACM*, 63(11), pp. 139–144. Available at: https://doi.org/10.1145/3422622.

Gretton, C. (2021) *Synthetic Control Arms: A Broader Clinical reach*, *PharmaVoice*. Available at: https://www.pharmavoice.com/news/2021-10-synthetic-control-arms-a-broader-clinical-reach/612010/ (Accessed: 2 October 2023).

Ko, Y.-A. *et al.* (2021) 'Developing a synthetic control group using electronic health records: Application to a single-arm lifestyle intervention study', *Preventive Medicine Reports*, 24, p. 101572. Available at: https://doi.org/10.1016/j.pmedr.2021.101572.

McCartan, S. *et al.* (2021) 'A comparison of individual participant data (IPD) and summary data to inform clinical trial simulation'.

Prasad, V. (2021) 'Reliable, cheap, fast and few: What is the best study for assessing medical practices? Randomized controlled trials or synthetic control arms?', *European Journal of Clinical Investigation*, 51(8), p. e13580. Available at: https://doi.org/10.1111/eci.13580.

Thorlund, K. *et al.* (2020) 'Synthetic and External Controls in Clinical Trials – A Primer for Researchers', *Clinical Epidemiology*, 12, pp. 457–467. Available at: https://doi.org/10.2147/CLEP.S242097.

Supplementary Material:

https://yoda.yale.edu/wp-content/uploads/2024/01/First_Year_Project_Proposal_V3.pdf