

Introduction to Synthetic Control Arm

Venkat Rajagopal and Abhinav Jain, Efficacy Consulting Group

ABSTRACT

There has been a rapid expansion in the use of non-RCT based evidence in the regulatory approval of treatments. A methodology named Synthetic Control Arm (SCA) has emerged to analyze external control data. SCAs help augment or replace randomized controls in many cases. SCAs provide cross-sponsor regulatory grade historical trial data that can be used for comparative analysis in new trials. This helps accelerate development timelines, reduces costs and increases the probability of trial success. It also helps decrease recruitment and reduces patient burden.

Appropriate statistical methodologies need to be used to address the difference between RCT and SCA data. SCA comparisons are based on the complexity associated with matching external data to the study. For a simple comparison, simple mean, median or fixed-effect pooling is enough. Imbalance adjustments are carried out with multivariate regression and propensity scoring. For more complex situations, Bayesian, random forests, and neural networks are used.

INTRODUCTION – SYNTHETIC CONTROL ARM

The pharmaceutical industry has used Randomized Clinical Trial as a tool to conduct trials efficiently and ethically. However, conducting clinical trials increases patient burden. Patients must be recruited and retained for conducting clinical trials efficiently. A key disconnect is that patients are used for the purposes of medical research in clinical trials as “subjects”, but they are patients with needs to get treated and the desire to get healthy. Therefore, any efforts to reduce clinical trial related patient burden is welcome.

On the one hand, there are patient recruitment issues and on the other the sheer volume of clinical trials is increasing year over year. The number of trials is expected to double every few years for the near future (Ref 1). When developing the protocol for a clinical trial, it is important to consider the benefits for patients taking part. Also, it is important to consider the possible burden the trial may place on patients. Scheduling difficulties, transportation issues, and confusing directions can all create a negative experience for patients that can ultimately lead to attrition.

Unfortunately, many clinical trials run into challenges related to enrollment and patient retention. Not only do 80% of trials struggle with enrollment (Ref 2), but some studies estimate that up to 30% of patients who join a clinical trial end up dropping out, including 18% of patients who randomize into a trial. Such numbers do indicate that the burden is high, and we need to think of innovative ways to address this.

Considering health literacy when creating materials for patients, making sure patient-facing materials are simple to follow, and that sites are trained to respond to questions from patients throughout the trial, sharing materials with patients during a feedback session or consult with a trusted partner for an evaluation of readability should all be part of trial design. Simple-to-follow directions on medication can also help improve patient adherence during the trial.

It's critical to offer options to accommodate patients' needs. The COVID-19 pandemic highlighted the need for flexibility regarding clinical trials. Therefore, study designs must be flexible as well. One way to be flexible is to leverage existing information for use in new contexts. For example, if safety data has been generated for a drug class, then that data can be potentially considered across drugs in the same class, given that the biases have been understood. Use of technology not only to improve patient experience but also to leverage other sources of data to be used in trials is a powerful way to reduce patient burden. This publication focuses on the use of technology to obtain data for clinical trials that lead to reduced patient burden and thereby improving the chances of getting better drugs to market faster.

SYNTHETIC DATA FOR CLINICAL TRIALS

There are statistical methods that can be used to synthesize patient data that can be used to replace or augment clinical trial data. The US Food and Drug Administration (FDA) and European Medicines Agency (EMA) have recognized these issues and have taken several initiatives to allow for these novel approaches to external control data. The FDA approved cerliponase alfa for a specific form of Batten disease, based on synthetic control study that compared the data of 22 patients studied in a single-arm trial versus independent external control group data with 42 untreated patients. Across 20 European countries, alectinib, a non-small cell lung cancer treatment, had an expansion of label based on synthetic control study based on an external data set of 67 patients. A kinase inhibitor, palbociclib, also had an expanded indication for men with HR+, HER2-advanced or metastatic breast cancer on the basis of external control data. The use of non-comparative data is not unique to rare diseases alone, as more common chronic diseases such as hepatitis C and previously treated rheumatoid arthritis have had treatments approved based on non-comparative data. Moreover, a recent review of 489 pharmaceutical technologies assessed by the National Institute for Health and Care Excellence (NICE) identified 22 submissions that used external data and synthetic control methods to establish clinical efficacy. Of these, 13 (59%) utilized published RCT data for their external control, and six (27%) utilized observational data. Over half of the applications were made in the last two recent years alone, further confirming the increasing attention paid by both drug manufacturers and health technology assessment agencies on this topic.

From the conventional evidence-based medicine, the use of external data to create synthetic controls for clinical evaluations represents a radical paradigm shift. A healthy degree of skepticism on the use of synthetic controls is thus expected from the scientific community. Nevertheless, it is likely that there will be an increasing number of clinical trials that use external data as a synthetic control, so it is important for researchers to comprehend the validity and reliability of synthetic control studies. Here in this paper, we provide an overview of the synthetic control arm methodology and nuances associated with statistical programming in this context.

SYNTHETIC CONTROL ARM – AN OVERVIEW

A Synthetic Control Arm (SCA) is a type of external control that is generated using external patient-level data to improve the interpretation of uncontrolled trials (Ref 3). Control and treatment data from multiple historical trials for the condition of interest is considered. Only control data from multiple trials is standardized and extracted. Synthetic control arm for the condition is created for the study. Single arm trial for the disease is generated from current clinical trial data. These two arms are compared and analyzed for results and submission. SCA advantages include reduced uncertainty, large control group, exploratory subgroup analysis, arms are matched at subject level, bias in time/site is minimized and helps design upcoming trials.

Synthetic control arms provide a way to safely and cost effectively leverage existing data instead of collecting data from patients recruited for a trial who have been assigned to the control or standard-of-care arm, synthetic control arms model those comparators using real-world data that has previously been collected from sources such as health data generated during routine care, including electronic health records, administrative claims data, patient-generated data from fitness trackers or home medical equipment, disease registries and historical clinical trial data.

The benefits to the pharmaceutical industry are clear. By reducing or eliminating the need to enroll control participants, a synthetic control arm can increase efficiency, reduce delays, lower trial costs, and speed lifesaving therapies to market. Imagine a trial that needs to have 500 participants in the treatment arm in order to demonstrate the effectiveness of a new therapy. Instead of having to recruit 1,000 patients — 500 for the active arm, 500 for the control arm — only 500 participants need to be recruited when a synthetic control arm is employed.

Fear of being assigned to placebo is one of the top reasons patients choose not to participate in clinical trials. This concern is amplified when an individual's prognosis is poor, or the current standard of care has limited effectiveness. Using a synthetic control arm instead of a standard control arm ensures that all

participants receive the active treatment, eliminating concerns about treatment assignment. This addresses an important participant concern and also removes an important barrier to recruitment.

The use of synthetic control arms can also eliminate the risk of unblinding when patients lean on their disease support social networks, posting details of their treatment, progress, and side effects that could harm the integrity of the trial.

APPLYING CDISC TO SCA

Source for generating SCA datasets can either be from an already CDISC standardized dataset or from other non-standardized sources such as EHR, hospital and claims data. The source dataset could be in standardized format in which case these datasets can be combined to generate the unique control. Otherwise, the existing data has to be converted to CDISC format before combining them. The combined SDTM dataset will form the basis of analysis datasets.

Data for SCA can be obtained either from historical clinical data or through algorithms with defined parameters allowing us to generate synthetic data. Synthetic Control Arm should represent an actual control arm in all ways possible. The data sources for SCA are multiple and disparate. They typically come from different clinical trials and can then be combined with claims and EHR data, if needed.

SCA datasets have to go through the rigor of CDISC before they can be used as a comparator to the clinical arm. For this, each of the variables in the original study (in case historical data is used) has to be mapped based on CDISC principles.

DATA SYNTHESIS METHOD

Historical clinical trial data is one source of for SCA data. Another source of SCA data is based on data synthesis using a statistical and algorithmic approach. The data synthesis process takes a real dataset as input, trains a generative model from it then generates synthetic data using the model. Multiple statistical or machine learning methods can be used to create a generative model.

Sequential decision trees can be used for data synthesis to fit a generative model (Ref 4). Sequential decision trees are used quite extensively in the health and social sciences for the generation of synthetic data. In these models, data is synthesized based on historical distribution and parameter assumptions. Approaches used can include conditional trees, parametric or tree algorithms. Methods such as deep learning have also been proposed for the synthesis of patient data. However, compared with deep learning synthesis methods, sequential decision trees have the advantage of not requiring a large input dataset that is used for training. It is therefore suitable for creating synthetic variants of clinical trial data that typically have a relatively small number of participants.

STATISTICS AND ANALYSIS

Given that the RCT approach is not taken, we will have to rely on statistical methodologies to help us reduce bias. Models can be chosen to be simple, or complex based on the scope of the study. For example, a naïve, simple or fixed-effect pooling approach can be easy to perform and interpret. However, this approach may be valid only for a small set of sub-group populations, thereby may not lead to precise results.

In the next level of complexity, imbalance adjustments can be carried out using multivariate regression and propensity scoring (Ref 5). This approach adjusts for imbalance to the extent explanatory factors are available in data. These are relatively easy to perform and relatively easy to interpret.

We can also leverage highly complex methodologies such as Bayesian mixed models, random forests, neural networks and cluster analysis. Although these methods are generally considered valid with good data and sufficient plausible confounding variables, they can be very complex or relatively time consuming to implement and test.

The statistical methodology should be chosen by the biostatistics team that is fully aware of the scope of the project and nuances of the data needed to carry out the analyses.

CONCLUSION

Synthetic Control Arm is an approach that can be leveraged to reduce patient burden in clinical trials. SCAs provide the flexibility of leveraging historical data or synthesized data to either replace or complement control arms in clinical trials. Advantages of SCAs include increased sample size leading to better sub-group analyses, reduces overall costs and increases the probability of trial success.

REFERENCES

1. Clinical Trial Trends - Trends, Charts, and Maps - ClinicalTrials.gov
2. Patient Enrolment - How To Improve Patient Recruitment In Clinical Trials? | Credevo Articles
3. Clin Epidemiol. 2020; 12: 457–467, Synthetic and External Controls in Clinical Trials – A Primer for Researchers, Kristian Thorlund,1,2 Louis Dron,2 Jay J H Park,2,3 and Edward J Mills1,2
4. Health informatics Original research - Can synthetic data be a proxy for real clinical trial data? A validation study, Zahra Azizi1, Chaoyi Zheng2, Lucy Mosquera2, GOING-FWD Collaborators, Correspondence to Dr Khaled El Emam; kelemam@ehealthinformation.ca
5. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. Stat Methods Med Res. 2012;21(3):273–293.

ACKNOWLEDGEMENTS

Ephicity Consulting Group colleagues

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Venkat Rajagopal
Ephicity Consulting Group
venkat.rajagopal@ephicity.com