Project Execution Plan:
Calibrating a Causal Transformer with Aggregate RCT Data

MyWay Digital Health

June 8, 2025

Abstract

This document outlines the step-by-step instructions for implementing the Causal Transformer calibration project. The objective is to develop a system that produces personalized treatment effect estimates that are calibrated against aggregate evidence from published Randomized Controlled Trials (RCTs). The methodology combines a Causal Transformer as a base predictor with a Gaussian Process (GP) as a calibration model. This revision incorporates the use of a dedicated Synthetic Data Generation (SDG) model for creating patient cohorts.

1 Phase 1: Base Model Preparation and Training

The first phase involves defining the data structure, preparing the observational dataset from SCI-Diabetes, and training the base Causal Transformer model.

1.1 Understanding the Causal Transformer Architecture

The Causal Transformer is a sequence model adapted for causal inference. It processes patient data as a series of temporal "tokens". For our purposes, we will model a simplified, static version where each patient has three corresponding input streams:

- **Covariate Stream (X):** A vector of baseline patient characteristics (e.g., age, sex, comorbidities, baseline biomarkers). $\mathbf{X} \in \mathbb{R}^{d_x}$.
- **Treatment Stream (A):** A vector indicating the treatment(s) assigned. This will be one-hot encoded. $\mathbf{A} \in \mathbb{R}^{d_a}$.
- **Outcome Stream (Y):** A vector of the outcomes observed for the patient under treatment $\mathbf{A}$. $\mathbf{Y} \in \mathbb{R}^{d_y}$.

The model is trained to predict the outcome stream $\mathbf{Y}$ given the history of covariates $\mathbf{X}$ and treatments $\mathbf{A}$. It can then be used to generate counterfactual predictions by changing the input treatment vector $\mathbf{A}$.

1.2 Step 1.1: Define Data Streams (A, X, Y)

We will define the specific variables for each stream.

- **Treatments (A):** Based on our target RCTs, we will initially focus on SGLT2 inhibitors and GLP-1 receptor agonists. The vector $\mathbf{A}$ should encode treatment type and potentially dose level, e.g., [`SGLT2i_low`, `SGLT2i_high`, `GLP1a_low`, `GLP1a_high`, `Placebo`].

- **Outcomes (Y):** The outcome vector **Y** should contain outcomes that are both clinically important and commonly reported in our set of RCTs. A task is to **survey all N=25 RCTs and identify the set of common primary and key secondary outcomes**. A proposed initial set is:
  - Change in HbA_{1c} from baseline.
  - Time-to-first Major Adverse Cardiovascular Event (MACE).
  - All-cause mortality.
- **Covariates (X):** See next step.

1.3 Step 1.2: Covariate Selection (The X vector)

This is a critical data curation task. The set of covariates **X** must be features that are available in **both** the SCI-Diabetes observational dataset and reported as baseline characteristics in our N=25 grounding RCTs.

1. Create a master spreadsheet of all baseline characteristics reported in the "Table 1" of each of the 25 RCTs.
2. Create a list of all available, high-quality features in the SCI-Diabetes dataset.
3. The final covariate vector **X** will be the **intersection of these two sets**. This ensures that we can fully describe each RCT population using features the Transformer understands.

Note: For simplicity and robustness, it is advisable to start with a parsimonious set of universally available covariates (e.g., age, sex, BMI, baseline HbA1c, baseline eGFR, duration of diabetes, history of cardiovascular disease) and expand later if necessary.

1.4 Step 1.3: Create the Master Dataset

With **A**, **X**, and **Y** defined, extract the relevant data from the SCI-Diabetes database.

1. For each patient in SCI-Diabetes, create a record containing their baseline covariate vector **X**, the treatment they received **A**, and the outcomes they experienced **Y**.
2. Perform necessary cleaning, imputation (e.g., MICE), and normalization (e.g., standard scaling for continuous features).
3. The final output should be a single, clean dataset (e.g., a `.csv` or `.parquet` file) ready for model training.

1.5 Step 1.4: Train the Causal Transformer

1. Download a public implementation of the Causal Transformer (e.g., from the original authors' GitHub repository).
2. Adapt the data loader to read our master dataset format.
3. Configure the model architecture and training hyperparameters.
4. Train the Causal Transformer model on the entire SCI-Diabetes dataset. This is a one-time, computationally intensive step.
5. Save the trained model weights. This is now your **Base Predictor**.

2 Phase 2: Calibration Model Development

This phase involves using the Base Predictor to generate predictions for our RCTs and preparing the data to train the calibration model.

2.1 Step 2.1: Simulate RCT Cohorts

For each of the $i \in \{1, ..., 25\}$ RCTs:

1. **Generate a large synthetic patient pool:** Utilize the project's dedicated **synthetic data generation models** to produce a large pool of virtual patients ($M \approx 1,000,000$). Each patient j must have a complete covariate vector $\mathbf{X}_j$ that is statistically representative of the SCI-Diabetes population.

2. **Apply Exclusion Criteria:** Apply the pragmatic filters identified in the project planning phase (e.g., age range, baseline values) to this pool to create a candidate set of patients.

3. **Match Population Characteristics via Weighting:** To avoid the curse of dimensionality, use a weighting technique like **Iterative Proportional Fitting (Raking)** to assign a weight w_j to each patient in the candidate set. Adjust the weights w_j such that the weighted summary statistics of the synthetic cohort match the target statistics from the RCT's Table 1.

$$\sum_{j=1}^{M'} w_j \cdot \text{age}_j = \text{mean\_age}_{\text{RCT}_i}$$

$$\sum_{j=1}^{M'} w_j \cdot \text{is\_male}_j = \text{proportion\_male}_{\text{RCT}_i}$$

$$\vdots$$

   The result is a weighted set of virtual patients that statistically replicates the population of RCT_i.

2.2 Step 2.2: Generate Transformer Predictions

For each simulated, weighted RCT cohort i:

1. Feed the cohort's covariate vectors $\mathbf{X}$ through the Base Predictor twice:
   - Once with the treatment vector $\mathbf{A}_{\text{treat}}$ to get predicted outcomes $\hat{\mathbf{Y}}_{\text{treat}}$.
   - Once with the treatment vector $\mathbf{A}_{\text{placebo}}$ to get predicted outcomes $\hat{\mathbf{Y}}_{\text{placebo}}$.

2. Calculate the Individual Treatment Effect (ITE) for each patient j: $\text{ITE}_j = \hat{\mathbf{Y}}_{\text{treat},j} - \hat{\mathbf{Y}}_{\text{placebo},j}$.

3. Calculate the Transformer's predicted Average Treatment Effect (ATE) for that cohort by taking the weighted average of the ITEs:

$$\text{ATE}_{\text{transformer},i} = \sum_{j=1}^{M'} w_j \cdot \text{ITE}_j \tag{1}$$

2.3 Step 2.3: Feature Engineering & Error Calculation

For each RCT i:

1. **Compute $\texttt{Population\_Vector}_i$:** This is the vector of summary statistics from the RCT's Table 1 that you used as targets for raking. It is your feature vector for the calibration model.
$$\texttt{Population\_Vector}_i = [\text{mean\_age}, \text{sd\_age}, \text{prop\_male}, \dots]$$

2. **Calculate Target $\texttt{Error}_i$:** This is the target variable for the calibration model. Let $\text{ATE}_{\text{rct},i}$ be the true ATE reported in the publication for RCT i.
$$\texttt{Error}_i = \text{ATE}_{\text{transformer},i} - \text{ATE}_{\text{rct},i} \tag{2}$$
You will have one such error value for each outcome (e.g., `Error_HbA1c`, `Error_MACE`). You will train a separate GP for each outcome.

3 Phase 3: Analysis and Calibration Model Training

3.1 Step 3.1: Visualize the Manifold Hypothesis (t-SNE)

To build intuition and visually validate our core hypothesis, we will visualize the error manifold.

1. Take the set of all $N = 25$ `Population_Vectors` you created. This is a dataset where each row is an RCT, and columns are population characteristics.
2. Use a dimensionality reduction algorithm, such as t-SNE or UMAP, to project these high-dimensional vectors into a 2D space.
3. Create a scatter plot of the 2D-projected points.
4. Color each point on the plot by its corresponding $\texttt{Error}_i$ value. Use a continuous color scale (e.g., blue for negative error, red for positive error).
5. **Interpretation:** If the manifold hypothesis holds, this plot will show structure. You should see "blobs" or gradients of color, indicating that RCTs with similar populations (and are therefore close in the t-SNE plot) also have similar prediction errors. A random-looking "salt and pepper" plot would invalidate the hypothesis.

3.2 Step 3.2: Train the GP Calibration Model

1. Create the training dataset for the GP:
   - Input Features: The $N = 25$ `Population_Vectors`.
   - Target Variable: The $N = 25$ `Error` values.
2. Using a library like GPyTorch, scikit-learn, or GPflow, train a Gaussian Process Regressor. The GP will learn the function:
$$\text{GP}(\texttt{Population\_Vector}) \rightarrow (\mu_{\text{error}}, \sigma^2_{\text{error}})$$
3. Choose a suitable kernel, such as the RBF (Radial Basis Function) kernel, and optimize its hyperparameters (e.g., lengthscale, variance) by maximizing the log marginal likelihood on the training data.

4 Phase 4: Deployment for Calibrated Prediction

This is the final workflow for predicting the outcome of a new target population (e.g., a new trial that is underway).

1. **Define Target Population:** Obtain the baseline characteristics (the "Table 1") for the new target population. This forms your new `Population_Vector_new`.

2. **Generate Raw ATE:** Simulate a cohort matching `Population_Vector_new` (using the same SDG model and raking procedure) and run it through the **Base Predictor** (the Causal Transformer) to get the raw prediction, `ATE_raw`.

3. **Predict Error:** Feed `Population_Vector_new` into the trained **GP Calibration Model**. It will output the predicted error mean ($\delta_{\text{predicted}}$) and variance ($\sigma_\delta^2$).

4. **Calculate Final Estimate:** Combine the outputs to get the final calibrated ATE and its 95% confidence interval.

$$\texttt{ATE\_calibrated} = \texttt{ATE\_raw} - \delta_{\text{predicted}} \tag{3}$$

$$95\%\ \text{CI} = \texttt{ATE\_calibrated} \pm 1.96 \cdot \sqrt{\sigma_\delta^2} \tag{4}$$

This provides a final, trustworthy prediction that is consistent with the body of existing clinical trial evidence, complete with a principled uncertainty estimate.