This document provides an overview of the data sources and data processing steps for calculating the sample size of the CHIGURU Adivasi birth cohort from the Towards Health Equity and Transformative Action on Tribal Health project dataset.

**1. Data Source**

The primary dataset for this project is the **THETA project dataset**. It is publicly available and can be downloaded from the following link:

* **Dataset Link:** [https://figshare.com/articles/dataset/*b\_Towards\_Health\_Equity\_and\_Transformative\_Action\_on\_tribal\_health\_THETA\_project\_dataset\_b*/23701863](https://figshare.com/articles/dataset/_b_Towards_Health_Equity_and_Transformative_Action_on_tribal_health_THETA_project_dataset_b_/23701863)

**2. Data Processing and Variable Calculation**

The raw data was processed using two separate scripts to generate the final variables for analysis.

**A. Substance Use Variable Calculation (CHIGURU sample calculation.ipynb)**

The recoding and calculation for the substance use variable were performed using a Python script in a Google Collab environment.

**Key Steps:**

1. **Data Loading:** The script begins by loading the dataset into a pandas DataFrame.
2. **Creating the substance use variable:** A new binary variable, substance\_use, was created to summarize substance use among adults. This was done by combining four existing variables: smoking\_tobacco, past\_smoking, smokeless\_tobacco, and alcohol\_use.
   * If an individual responded "yes" to **any** of these four categories, their substance\_use was coded as using.
   * If an individual responded with answers other than "yes" and had no missing data for these columns, they were coded as not using.
   * Rows with blank or missing information for these variables were coded as na.
3. **Dataset Merging:**
   * The script subsets the data into adult (age > 17.9 years) and child (age < 5.01 years) datasets.
   * These two datasets are then merged using a left join on the fulcrum\_id\_parent unique identifier. This links children's data to the corresponding adult (presumably the mother) from the same household.
4. **Output Generation:** The final merged dataset, containing both adult and child information, is exported as merged\_dataset.csv. This file serves as the input for the next stage of analysis in R.

**B. Anthropometric Z-Score Calculation (ctrith\_sample.R)**

The calculation of anthropometric Z-scores and the subsequent classification of malnutrition were performed using an R script.

**Key Steps:**

1. **Data Loading:** The script imports the merged\_dataset.csv (generated from the Python script) and a separate household.xlsx file containing household-level information like caste. These are then merged.
2. **Z-Score Calculation:**
   * The anthro R package is used to calculate standardized Z-scores for child growth indicators.
   * The anthro\_zscores() function calculates Weight-for-Age Z-score (zwei) and Weight-for-Length/Height Z-score (zwfl) using the child's age in days, weight, height, and sex.
3. **Data Filtering:** The dataset is filtered to include only:
   * Children under 36 months of age.
   * Adults (mothers) under 30 years of age.
4. **Malnutrition Classification:**
   * **Wasting:** A new variable, wfhz\_class, is created based on the zwfl score. A child with a zwfl score less than or equal to -2 is classified as "wasting".
   * **Underweight:** A new variable, waz\_class, is created based on the zwei score. A child with a zwei score less than or equal to -2 is classified as "underweight".
5. **Statistical Analysis:**
   * The script performs cross-tabulations (tbl\_cross) to examine the relationship between adult substance use and child malnutrition outcomes (wasting and underweight).
   * A logistic regression model (glm) is run to assess the association between the adult's substance use and the child's likelihood of being underweight (waz\_class).