Modeling Climate-Attributable Sickle Cell Mortality Risk in Africa

This report describes the data preparation, modeling procedures, and forecasting methods used to estimate and project the climate-attributable risk of sickle cell disease (SCD) mortality across African countries. The objective of the modeling pipeline is to generate data-driven forecasts of climate-linked mortality risk to inform regional health preparedness and climate adaptation strategies.

Data Sources: Two primary data sources are used: annual country-level mortality counts for sickle cell disease from the Institute for Health Metrics and Evaluation (IHME) and monthly climate indicators (temperature, precipitation, and aerosol optical depth) sourced from Africa Data Hub. These datasets are merged by country and year to construct a panel dataset spanning multiple African countries over time and used as input for downstream processing. The merged dataset contains monthly observations of climate variables and associated annual mortality totals at the country level.

Temporal Disaggregation of Annual Mortality Data: The mortality data from IHME is only available at the yearly resolution. To enable monthly-resolution modeling, annual sickle cell mortality values are disaggregated to monthly estimates using a Denton-Cholette approach disaggregation method to estimate plausible monthly mortality values that aggregate to the known annual total. For each country, a linear regression is fitted using yearly average temperature and precipitation as predictors of annual mortality. The fitted coefficients serve as country-specific weights to combine temperature and precipitation into a monthly indicator series. A disaggregation function is used to allocate annual mortality totals across the 12 months by minimizing the sum of squared second-order differences in the monthly series. This enforces temporal smoothness, in line with the Denton-Cholette method. The optimization ensures that the monthly series sums exactly to the annual total and remains non-negative. This is applied to all country-year pairs with complete monthly data. In cases where monthly mortality values are partially missing, a follow-up imputation step ensures consistency with annual totals by distributing residual counts evenly across missing months, subject to bounding constraints.

Feature Engineering: Following disaggregation, feature engineering is applied to construct predictors for the regression model. Monthly raw values of average temperature, precipitation, and aerosol optical depth (AOD) are retained as base features. To account for temporal accumulation and lag effects, rolling means over 1-, 2-, 3-, 6-, and 12-month windows are computed for each variable, grouped by country. Lagged features at 1-, 2-, and 3-month intervals are also generated for the same variables. Additionally, a composite interaction term (De martonne's Aridity Index) representing aridity and humidity scales is constructed using monthly temperature and precipitation values. To encode seasonal variation, the month variable is transformed using sinusoidal functions, resulting in two features representing cyclic month progression.

Feature Selection: Prior to model training, the feature space is subjected to multicollinearity diagnostics. A correlation matrix is computed for the full set of engineered features, and features with pairwise absolute correlation exceeding 0.8 are flagged. In addition, the variance inflation factor (VIF) is calculated for each feature to detect linear dependencies in the design matrix. Low importance features identified as highly collinear with other high importance features, through either correlation thresholding or VIF analysis are excluded from the final predictor set.

Model Training: Gradient Boosting Regression: The climate-attributable mortality signal is modeled using Gradient Boosting Regression, implemented via the XGBRegressor class from the xgboost library. The model is trained to predict monthly disaggregated mortality values as a function of climate features alone, allowing isolation of the climate-attributable signal. The training and evaluation procedure includes a standard split into training (80%) and validation (20%) sets. Predictions are generated and performance is evaluated using Mean Absolute Error (MAE) and Mean Squared Error (MSE).

SHAP-Based Attribution: To interpret the model output, SHAP (SHapley Additive exPlanations) values are computed. These decompose the prediction into additive contributions of each input feature for each observation, enabling identification of the relative importance of different climate factors. SHAP values corresponding to each monthly prediction are summed to produce a monthly climate-attributable mortality estimate.

Forecasting Climate-Attributed Mortality Signal: To project the climate-attributable component of mortality into the future, the time series of SHAP-derived climate-attributable mortality values is then modeled using a Seasonal Autoregressive Integrated Moving Average

(SARIMA) process. The SARIMA model is fitted separately to the climate-attributable component for each country to capture seasonal and autoregressive patterns for estimating projections conditioned on historical statistical relationships between climate variables and SCD mortality. Forecasts are generated for monthly values through the year 2030 and can be used to inform regional health adaptation planning.

Output Integration: The entire pipeline, from disaggregation through forecasting, is implemented in Python and the output of the model feeds into a Power BI interactive dashboard which is hosted on Streamlit, allowing users to explore forecasted climate-linked mortality risk across African countries. All components are grounded in open-source data and code, ensuring transparency and reproducibility of the modeling results.