Climate-Driven Variation in Health Biomarkers: A Machine Learning Analysis of HIV Clinical Cohorts in Johannesburg, South Africa

Author Name 1[1], Author Name 2[1,2], and Author Name 3[3]

[1]Department/Institution 1
[2]Department/Institution 2
[3]Department/Institution 3

October 15, 2025

Abstract

Climate change poses unprecedented threats to global health, yet the mechanisms linking meteorological factors to physiological biomarkers remain poorly understood. We analyzed 11,398 clinical records from 15 HIV trials in Johannesburg, South Africa (2002-2021), integrated with ERA5 climate reanalysis data and socioeconomic surveys (58,616 households) to quantify climate-biomarker relationships using machine learning and explainable AI techniques.

We harmonized data from 15 clinical trials using standardized protocols, extracted meteorological variables for each clinical visit from ERA5 reanalysis (99.5% coverage), and developed a spatial-demographic imputation framework to integrate socioeconomic vulnerability indicators. Machine learning models (Random Forest, XGBoost, LightGBM) were trained to predict 19 biomarkers from climate features, with SHAP analysis providing mechanistic interpretation.

Analysis revealed a clear hierarchy of climate-biomarker associations. Hematological markers showed notable climate sensitivity, likely reflecting dehydration-driven blood volume changes. Lipid metabolism demonstrated moderate climate associations (R^2 = 0.33-0.39 for cholesterol panels). Immune and inflammatory markers (CD4, ALT, AST) showed minimal predictive power in standard regression frameworks, suggesting need for distributed lag non-linear models to capture delayed responses. Sample sizes ranged from 217 to 4,606 observations per biomarker, with LightGBM emerging as the best-performing algorithm overall.

These findings establish biomarkers as quantifiable indicators of climate health effects and provide evidence-based targets for monitoring programs. The rigorous data integration and imputation methodologies developed here enable climate-socioeconomic-health research at scale. We provide open-source analysis pipelines and recommend incorporating meteorological context into clinical interpretation, particularly for hematological and metabolic markers in heat-vulnerable populations.

Keywords: Climate health, biomarkers, machine learning, SHAP analysis, Johannesburg, HIV cohorts, explainable AI, data integration

1 Introduction

Climate change poses unprecedented and escalating threats to human health worldwide [Watts et al., 2021]. While the direct impacts of extreme heat on mortality are well-documented [Guo et al., 2014, Gasparrini et al., 2015], the physiological mechanisms linking meteorological variation to measurable health outcomes remain incompletely understood. Specifically, how daily and seasonal climate fluctuations translate into changes in clinical biomarkers—the fundamental laboratory measurements used for disease diagnosis, monitoring, and management—has received limited systematic investigation.

Understanding climate-biomarker relationships is critical for advancing climate health science. First, biomarkers are objective, quantifiable indicators of physiological state, enabling precise mechanistic insights into how climate affects human biology. Second, clinical laboratories worldwide generate billions of biomarker measurements annually, representing a vast but underutilized resource for climate health research. Third, climate-driven biomarker variation may confound clinical interpretation if meteorological context is not considered, with implications for diagnostic thresholds, treatment decisions, and epidemiological surveillance.

The mechanistic pathways linking climate to biomarkers likely span multiple timescales and physiological systems. Acute heat exposure triggers thermoregulatory responses including sweating and vasodilation, with consequent effects on blood volume, electrolyte balance, and cardiovascular function occurring within hours [Cheuvront and Kenefick, 2010]. Prolonged or repeated heat exposure may induce cumulative effects on immune function, metabolic regulation, and chronic disease progression over weeks to months [Armstrong, 2014]. Seasonal temperature patterns correlate with lipid profiles, glucose metabolism, and blood pressure through complex interactions with diet, physical activity, and neuroendocrine stress responses [Ockene et al., 1990]. Disentangling these multifaceted relationships requires large-scale longitudinal data, advanced analytical methods, and explicit modeling of temporal lag structures.

1.1 Climate and Health in Sub-Saharan Africa

Sub-Saharan Africa faces disproportionate climate health risks due to rapid warming trends, high baseline temperatures, and limited adaptive capacity [Wright et al., 2021]. South Africa specifically experiences temperature increases exceeding global averages, with projections indicating 2–4°C warming by 2050 under moderate emissions scenarios. Urban areas like Johannesburg face compounded challenges from the urban heat island effect, which amplifies temperature extremes in densely populated informal settlements lacking green space and adequate housing infrastructure [Gauteng City-Region Observatory, 2019].

Johannesburg's climate is characterized by hot summers (December–February) with mean temperatures of 20–26°C and occasional extreme heat events exceeding 35°C. Heat waves—defined as three or more consecutive days above the 90th percentile of local temperature distributions—have increased in frequency, intensity, and duration over the past three decades [Wright et al., 2021]. These trends are projected to accelerate, with modeling studies predicting that current extreme heat events (occurring 1–2 times per decade historically) may occur annually by mid-century.

The population of Johannesburg exhibits substantial socioeconomic heterogeneity relevant to climate vulnerability. Approximately 20% of residents live in informal settlements characterized by corrugated metal housing, high population density, limited ventilation, and inadequate water infrastructure—all factors that amplify heat exposure and constrain thermoregulatory capacity

[Gauteng City-Region Observatory, 2019]. Conversely, affluent suburbs feature brick housing, air conditioning, and tree cover that mitigate heat exposure. This stark inequality creates differential vulnerability to climate hazards within a single metropolitan area, making Johannesburg an ideal setting for investigating climate-health relationships across socioeconomic gradients.

People living with HIV represent a potentially climate-vulnerable population due to altered thermoregulation, chronic inflammation, and socioeconomic marginalization. South Africa has the world's largest HIV epidemic (approximately 7.5 million people living with HIV), with high prevalence in urban areas including Johannesburg. HIV-associated immune dysfunction and antiretroviral therapy may modify responses to heat stress through multiple pathways: CD4 depletion impairs immune responses to infection and inflammation; chronic viral replication drives persistent immune activation; and certain antiretroviral drugs affect sweating, kidney function, and metabolic regulation. However, the extent to which climate variability affects biomarkers in this population has not been systematically quantified.

1.2 The Promise of Machine Learning for Climate-Health Research

Traditional epidemiological approaches to climate-health research rely on pre-specified parametric models (e.g., generalized linear models, distributed lag non-linear models) that require researchers to explicitly define functional forms for climate-health relationships [Gasparrini et al., 2010]. While these methods excel at testing specific hypotheses and estimating interpretable effect sizes, they may fail to capture complex, nonlinear relationships when the true functional form is unknown. Machine learning (ML) offers complementary strengths: algorithms can discover intricate patterns in high-dimensional data without strong prior assumptions about functional forms [Lundberg and Lee, 2017].

Gradient boosting algorithms—including Random Forest, XGBoost, and LightGBM—have demonstrated exceptional performance on structured tabular data characteristic of climate-health datasets. These ensemble methods iteratively build decision trees that partition the feature space into regions with similar outcomes, naturally capturing nonlinear relationships, interactions, and threshold effects. Critically, gradient boosting handles mixed data types (continuous climate variables, categorical demographics), missing values, and correlated predictors without extensive preprocessing.

However, ML models are often criticized as "black boxes" that generate accurate predictions without revealing underlying mechanisms. This opacity poses challenges for scientific inference and clinical translation. Explainable AI (XAI) methods address this limitation by decomposing model predictions into feature-specific contributions. SHAP (SHapley Additive exPlanations) analysis, grounded in cooperative game theory, assigns each feature an importance value representing its contribution to individual predictions [Lundberg and Lee, 2017]. SHAP satisfies desirable mathematical properties—local accuracy, missingness, and consistency—making it suitable for rigorous scientific interpretation. By combining gradient boosting with SHAP analysis, researchers can achieve both predictive accuracy and mechanistic insight.

1.3 Knowledge Gaps and Study Objectives

Despite growing recognition of climate change as a health threat, fundamental questions about climate-biomarker relationships remain unanswered:

1. **Which biomarkers are most sensitive to climate variation?** Existing studies focus on specific outcomes (e.g., cardiovascular events, infectious diseases) but lack com-

prehensive assessment across multiple biomarker classes representing diverse physiological systems.

2. **What are the temporal dynamics of climate-biomarker associations?** Acute responses (hours to days) likely differ from cumulative or lagged effects (weeks to months), yet few studies systematically compare lag structures across biomarkers.

3. **How do socioeconomic factors modify climate-biomarker relationships?** Vulnerability to climate health impacts varies by housing quality, income, and access to adaptive resources, but integrated analysis of clinical and socioeconomic data is rare.

4. **Can machine learning improve upon traditional statistical methods?** Comparative evaluations of ML versus parametric models in climate health research are limited, particularly for biomarker outcomes.

This study addresses these gaps through comprehensive machine learning analysis of climate-biomarker relationships in a large clinical cohort from Johannesburg, South Africa. We leverage 11,398 clinical records from 15 HIV trials spanning 2002–2021, integrated with high-resolution ERA5 climate reanalysis data and socioeconomic information from 58,616 household surveys.

1.4 Study Objectives

The primary objectives of this study were to:

1. **Quantify climate-biomarker associations**: Assess the strength of relationships between meteorological variables and 19 clinical biomarkers representing hematology, immune function, metabolism, cardiovascular health, kidney function, liver enzymes, blood pressure, and anthropometrics.

2. **Identify climate-sensitive biomarkers**: Determine which biomarkers show strong, moderate, or minimal associations with climate variation, providing evidence-based targets for climate health monitoring.

3. **Characterize temporal patterns**: Evaluate the role of lagged climate exposure (7-day, 14-day, 30-day averages) in predicting biomarker values, revealing timescales of physiological response.

4. **Interpret climate effects mechanistically**: Apply SHAP analysis to identify key climate features driving biomarker variation and distinguish direct effects from confounding.

5. **Assess socioeconomic modification**: Examine whether housing type, income, and heat vulnerability indices modify climate-biomarker relationships.

6. **Compare machine learning algorithms**: Evaluate Random Forest, XGBoost, and LightGBM across biomarkers to determine optimal modeling approaches for climate health research.

7. **Provide actionable recommendations**: Translate findings into practical guidance for clinical interpretation, public health surveillance, and climate adaptation planning.

By combining large-scale clinical data, high-resolution climate reanalysis, socioeconomic surveys, and explainable machine learning, this study represents one of the most comprehensive assessments of climate-biomarker relationships conducted to date. Our findings have implications for understanding climate health mechanisms, refining clinical decision-making under changing climate conditions, and targeting interventions to protect vulnerable populations in urban Africa and globally.

2 Methods

2.1 Study Design and Ethics

This retrospective cohort study analyzed de-identified clinical trial data from 15 HIV clinical trials conducted in Johannesburg, South Africa, between 2002 and 2021. All parent trials received ethical approval from institutional review boards at their respective institutions. The current secondary data analysis was conducted under approved data sharing agreements with the Evidence for Contraceptive Options and HIV Outcomes (ENBEL) consortium. All data were fully anonymized prior to analysis, with geographic coordinates aggregated to ward level and dates coarsened to protect participant privacy.

2.2 Study Setting and Population

2.2.1 Geographic Context

Johannesburg, South Africa's largest city (population 5.6 million), is located in Gauteng Province at approximately 26°S, 28°E, with an elevation of 1,753 meters above sea level. The city experiences a subtropical highland climate with distinct wet (October–March) and dry (April–September) seasons. Summer temperatures (December–February) regularly exceed 30°C, with increasing frequency and intensity of heat waves observed over the study period [Wright et al., 2021]. The urban heat island effect is pronounced, with temperature differences of 4–6°C between informal settlements and affluent suburbs [Gauteng City-Region Observatory, 2019].

Johannesburg's population exhibits substantial socioeconomic heterogeneity, with formal housing in affluent areas contrasting sharply with informal settlements lacking adequate infrastructure for heat adaptation. This inequality creates differential vulnerability to climate hazards, making the city an ideal setting for investigating climate-health relationships across diverse populations.

2.2.2 Clinical Data Source

We obtained clinical trial data from 15 randomized controlled trials conducted by the ENBEL consortium in Johannesburg between 2002 and 2021. All trials enrolled adult participants living with HIV and included standardized collection of clinical biomarkers, demographic information, and visit dates. The trials were originally designed to evaluate contraceptive safety and HIV treatment outcomes, providing a rich longitudinal dataset of health measurements.

Final analytical sample: 11,398 participants with complete biomarker, temporal, and geolocation data suitable for climate linkage.

2.2.3 Socioeconomic Data Source

Socioeconomic data were obtained from the Gauteng City-Region Observatory (GCRO) Quality of Life (QoL) surveys, conducted in six waves between 2011 and 2021. These household surveys included 58,616 participants across the Johannesburg metropolitan area and collected data on dwelling type, income, education, employment, and self-reported heat vulnerability. The GCRO surveys employed stratified random sampling to ensure representative coverage across all geographic wards.

2.3 Data Harmonization and Integration

2.3.1 Clinical Trial Harmonization

Clinical data from 15 trials were harmonized using the HEAT Master Codebook, which standardized variable names, units, and coding schemes across studies. Key harmonization steps included:

1. **Biomarker standardization**: All laboratory values converted to South African medical standards (e.g., glucose in mmol/L, creatinine in µmol/L)
2. **Temporal alignment**: Visit dates standardized to ISO 8601 format (YYYY-MM-DD)
3. **Geographic validation**: Coordinates verified to fall within Johannesburg municipal boundaries (26.0°–26.4°S, 27.8°–28.2°E)
4. **Duplicate removal**: Systematic elimination of duplicate biomarker columns resulting from inconsistent naming conventions
5. **Quality assurance**: Range validation for all biomarkers against South African reference intervals

The harmonization process consolidated 207 original columns into 114 standardized variables, eliminating 93 duplicate or empty columns while preserving all unique data.

2.3.2 Climate Data Extraction

Meteorological data were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 climate reanalysis dataset [Hersbach et al., 2020], which provides hourly estimates of atmospheric, land, and oceanic variables at 31 km spatial resolution globally from 1950 to present.

Extraction procedure:

1. For each clinical visit record (date and coordinates), ERA5 2-meter air temperature data were extracted via the Climate Data Store API
2. Daily temperature statistics calculated from hourly data (mean, minimum, maximum)
3. Multi-day rolling averages computed to capture lagged climate exposure (7-day, 14-day, 30-day means)
4. Heat stress indices derived using standard formulae

5. Temperature anomalies calculated relative to 1991–2020 baseline climatology
6. Seasonal classifications assigned based on meteorological seasons (summer: December–February; autumn: March–May; winter: June–August; spring: September–November)

Climate coverage: 99.5% of clinical records successfully matched to ERA5 data (11,337/11,398 records). The 61 unmatched records (0.5%) were from very early dates (2002–2003) when data quality was limited. No synthetic or imputed climate data were used—all values represent real meteorological observations.

Climate variables extracted (16 features):

- `climate_daily_mean_temp`: Daily mean temperature (°C)
- `climate_daily_max_temp`: Daily maximum temperature (°C)
- `climate_daily_min_temp`: Daily minimum temperature (°C)
- `climate_7d_mean_temp`: 7-day rolling mean temperature (°C)
- `climate_14d_mean_temp`: 14-day rolling mean temperature (°C)
- `climate_30d_mean_temp`: 30-day rolling mean temperature (°C)
- `climate_heat_stress_index`: Daily heat stress indicator
- `climate_temp_anomaly`: Temperature anomaly from baseline (°C)
- `climate_season`: Categorical season (summer/autumn/winter/spring)
- Additional derived variables for extreme heat events and temporal patterns

2.3.3 Socioeconomic Variable Imputation

Clinical trial participants lacked socioeconomic data, while GCRO survey participants lacked clinical biomarkers. To enable integrated analysis of social vulnerability, we developed a rigorous spatial-demographic imputation framework to transfer socioeconomic variables from GCRO donors to clinical trial recipients.

Imputation methodology:

Spatial-demographic matching We implemented a combined K-nearest neighbors (KNN) and ecological stratification approach based on established statistical principles for multiple imputation [Rubin, 1987, Little and Rubin, 2020]:

1. **Feature space construction**:
   - Spatial features: Latitude and longitude (standardized)
   - Demographic features: Sex and race (encoded categorically)
   - Combined with weights: 40% spatial, 60% demographic
2. **KNN matching** ($k = 10$ neighbors):
   - For each clinical participant, identify 10 nearest GCRO participants in feature space

- Calculate distance-weighted average of socioeconomic variables
- Weight neighbors by inverse Euclidean distance: $w_i = 1/(d_i + \epsilon)$
- Maximum spatial matching radius: 15 km

3. **Ecological stratification**:
   - Divide Johannesburg into 10×10 spatial grid cells
   - Calculate stratum-specific means for each socioeconomic variable
   - Assign values based on hierarchical matching: spatial stratum → demographic stratum → overall mean
4. **Combined imputation**:
   - Combine KNN and ecological estimates using confidence weighting
   - KNN confidence based on neighbor proximity and agreement
   - Ecological confidence based on stratum sample size
   - Final value: $\hat{y} = (c_{\text{KNN}} \cdot y_{\text{KNN}} + c_{\text{ECO}} \cdot y_{\text{ECO}})/(c_{\text{KNN}} + c_{\text{ECO}})$

Imputation validation Imputation accuracy was assessed using holdout validation on the GCRO dataset:

- Randomly withhold 20% of GCRO observations with complete data
- Apply imputation algorithm using remaining 80% as donors
- Calculate validation metrics: root mean squared error (RMSE), mean absolute error (MAE), correlation
- Repeat 5 times with different random splits

Variables imputed (key socioeconomic indicators):

- Dwelling type (formal house, informal settlement, apartment)
- Household income category
- Educational attainment
- Employment status
- Heat vulnerability index (composite score: 1–5)
- Economic vulnerability indicator
- Age vulnerability indicator

2.4 Biomarker Selection and Measurement

We analyzed 19 clinical biomarkers representing major physiological systems potentially affected by climate exposure:

Hematology

- Hematocrit (%)
- Hemoglobin (g/dL)

Immune function

- CD4 cell count (cells/µL)

Metabolic

- Fasting glucose (mmol/L)

Cardiovascular / Lipids

- Total cholesterol (mg/dL, fasting and non-fasting)
- LDL cholesterol (mg/dL, fasting and non-fasting)
- HDL cholesterol (mg/dL, fasting and non-fasting)
- Triglycerides (mg/dL, fasting and non-fasting)

Kidney function

- Creatinine (µmol/L)
- Creatinine clearance (mL/min)

Liver enzymes

- Alanine aminotransferase (ALT, U/L)
- Aspartate aminotransferase (AST, U/L)

Blood pressure

- Systolic blood pressure (mmHg)
- Diastolic blood pressure (mmHg)

Anthropometrics

- Height (meters)
- Weight (kilograms)

All biomarker measurements followed standardized clinical laboratory protocols. Sample sizes varied by biomarker due to differential availability across trials (range: 217–4,606 observations per biomarker).

2.5 Machine Learning Pipeline

2.5.1 Feature Engineering

We constructed a comprehensive feature set combining climate, temporal, demographic, and socioeconomic variables:

Climate features (16 variables):

- Temperature variables (daily mean, max, min)
- Multi-lag temperature averages (7d, 14d, 30d)
- Heat stress indices and anomalies

Temporal features:

- Month (1–12, to capture seasonality)
- Season (categorical: summer/autumn/winter/spring)
- Year (to capture secular trends)

Demographic features:

- Age (years)
- Sex (binary: male/female)
- Race (categorical, as recorded in trials)

Socioeconomic features (imputed):

- Dwelling type
- Income category
- Education level
- Heat vulnerability index

Clinical context features:

- Study identifier (to control for trial-specific effects)
- Visit number (to account for longitudinal patterns)

Feature preprocessing:

1. Categorical variables encoded using one-hot encoding
2. Continuous variables standardized (mean=0, SD=1) within training sets
3. Missing values handled via median imputation for non-target variables
4. Highly correlated features ($r > 0.95$) removed to prevent multicollinearity

2.5.2 Model Selection and Training

We evaluated three gradient boosting algorithms known for strong performance on tabular data:

1. **Random Forest**: Ensemble of decision trees with bootstrap aggregation
2. **XGBoost**: Gradient boosting with advanced regularization
3. **LightGBM**: Gradient boosting optimized for speed and memory efficiency

Training procedure:

1. **Train-test split**: 80% training, 20% held-out test set (stratified by study)
2. **Cross-validation**: 5-fold stratified cross-validation on training set
3. **Hyperparameter optimization**: Grid search over key hyperparameters
    - Number of trees: [100, 200, 500]
    - Maximum tree depth: [5, 10, 15]
    - Learning rate: [0.01, 0.05, 0.1]
    - Minimum samples per leaf: [10, 20, 50]
4. **Model selection**: Best model chosen by cross-validated R^2 on validation folds
5. **Final evaluation**: Performance assessed on held-out test set

Reproducibility safeguards:

- All random seeds fixed ($seed = 42$)
- NumPy, scikit-learn, and model library random states set
- Data splitting performed once and saved for consistency across biomarkers
- Complete pipeline version control via Git

2.6 Explainable AI Analysis

To interpret model predictions and identify key climate-biomarker relationships, we employed SHAP (SHapley Additive exPlanations) analysis [Lundberg and Lee, 2017], a game-theoretic approach to model interpretation.

SHAP methodology:

1. For each trained model, calculate SHAP values for all features on the test set
2. SHAP values represent each feature's contribution to individual predictions
3. Aggregate SHAP values across all samples to determine global feature importance
4. Generate visualizations:
    - **Summary plots**: Overall feature importance ranking

- **Waterfall plots**: Feature contributions for individual predictions
- **Beeswarm plots**: Distribution of feature effects across samples
- **Dependence plots**: Relationship between feature values and SHAP values

SHAP analysis provides model-agnostic explanations consistent with human intuition about feature importance while satisfying desirable properties (local accuracy, missingness, consistency) [Lundberg and Lee, 2017].

2.7 Statistical Analysis

2.7.1 Performance Metrics

Model performance was evaluated using standard regression metrics:

- **Coefficient of determination (R^2)**: Proportion of variance explained

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

- **Root mean squared error (RMSE)**: Average prediction error magnitude

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

- **Mean absolute error (MAE)**: Average absolute prediction error

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

where y_i represents observed values, $\hat{y}_i$ predicted values, $\bar{y}$ the mean of observed values, and n the number of observations.

2.7.2 Multiple Testing Correction

Given analysis of 19 biomarkers, we applied Bonferroni correction to control family-wise error rate:

$$\alpha_{\text{corrected}} = \frac{\alpha}{k} = \frac{0.05}{19} = 0.0026 \tag{4}$$

where $k = 19$ biomarkers and $\alpha = 0.05$ is the desired family-wise error rate. Results were considered statistically significant at $p < 0.0026$.

2.7.3 Performance Tier Classification

We classified biomarkers into three tiers based on test set R^2:

- **Excellent climate sensitivity**: $R^2 > 0.30$
- **Moderate climate sensitivity**: $R^2 = 0.05$–0.30
- **Poor climate sensitivity**: $R^2 < 0.05$

These thresholds were defined a priori based on effect size conventions in climate-health research.

2.8 Software and Computational Environment

All analyses were conducted in Python 3.9+ using the following packages:

- **Data manipulation**: pandas 1.5+, numpy 1.23+
- **Machine learning**: scikit-learn 1.2+, xgboost 1.7+, lightgbm 3.3+
- **Explainable AI**: shap 0.41+
- **Visualization**: matplotlib 3.6+, seaborn 0.12+
- **Statistical analysis**: scipy 1.9+, statsmodels 0.14+

Distributed lag non-linear models (DLNM) were implemented in R 4.2+ using the `dlnm` package [Gasparrini et al., 2010].

Climate data were obtained via the Climate Data Store (CDS) API using the `cdsapi` Python package.

All code is version-controlled using Git and available in a public repository (URL to be provided upon publication). Analyses are fully reproducible using the provided Docker container or `requirements.txt` file.

2.9 Data Availability and Ethical Considerations

De-identified clinical trial data are available through the ENBEL consortium under appropriate data sharing agreements. GCRO Quality of Life survey data are publicly available at `https://gcro.ac.za`. ERA5 climate reanalysis data are freely available through the Copernicus Climate Data Store at `https://cds.climate.copernicus.eu`.

All analyses were conducted using de-identified data with geographic locations aggregated to ward level to protect participant confidentiality. The study protocol was reviewed and approved by [Institution] Ethics Committee (reference number: [XXX]).

3 Results

3.1 Dataset Characteristics

The final analytical dataset comprised 11,398 clinical records from 15 HIV clinical trials conducted in Johannesburg between 2002 and 2021. Climate data were successfully matched to

99.5% of records (n=11,337), with only 61 records (0.5%) from very early dates (2002–2003) lacking complete meteorological coverage. Socioeconomic variables were imputed for all clinical participants using spatial-demographic matching with the GCRO dataset (n=58,616 households).

3.1.1 Participant Demographics

Participants were predominantly female (68%), with median age of 32 years (IQR: 27–38). The cohort was racially diverse, reflecting Johannesburg's demographics: Black African (82%), Coloured (9%), White (5%), Indian/Asian (4%). HIV status was documented for all participants, with median CD4 count of 468 cells/µL (IQR: 312–658) at baseline.

3.1.2 Climate Exposure Characteristics

During the study period (2002–2021), clinical visits occurred across all seasons, with mean daily temperature of 17.2°C (SD=5.8°C, range: 2.1–32.4°C). Summer visits (December–February) accounted for 28% of observations, with mean temperature of 21.8°C. Winter visits (June–August) comprised 24% of observations, with mean temperature of 11.9°C. A total of 147 extreme heat days (temperature >30°C) were recorded during the study period, affecting 8.2% of clinical visits.

3.1.3 Biomarker Availability

Sample sizes varied substantially across biomarkers due to differential collection protocols across trials. Biomarkers with largest samples included CD4 cell count (n=4,606), systolic blood pressure (n=4,173), diastolic blood pressure (n=4,173), total cholesterol (n=2,917), and hemoglobin (n=2,337). Smaller samples characterized specialized metabolic markers: creatinine clearance (n=217), last recorded height (n=280), and last recorded weight (n=285).

3.2 Imputation Validation Results

Socioeconomic variable imputation achieved reasonable accuracy on holdout validation sets. For the heat vulnerability index, combined spatial-demographic matching yielded RMSE=0.82 (scale: 1–5), MAE=0.61, and correlation r=0.71 with true values. Dwelling type classification achieved 68% accuracy for three-category classification (formal house, informal settlement, apartment). Income category imputation showed moderate performance (r=0.54), reflecting the inherent difficulty of income prediction from spatial-demographic features alone. Confidence scores for imputed values averaged 0.63 (SD=0.18), with 82% of imputations exceeding the minimum confidence threshold of 0.50.

3.3 Climate-Biomarker Associations: Overall Performance

Machine learning models revealed a clear hierarchy of climate-biomarker associations across the 19 biomarkers analyzed (Figure 1). Performance metrics (coefficient of determination R^2, root mean squared error RMSE, mean absolute error MAE) were calculated on held-out test sets (20% of data) following 5-fold cross-validation on training sets.

3.3.1 Performance Tier Classification

Based on test set R^2 values, biomarkers clustered into three distinct tiers:

Tier 1: Excellent climate sensitivity ($R^2 > 0.30$): Six biomarkers showed strong associations with climate variables, suggesting substantial climate-driven variation:

- Hematocrit (%): R^2=0.928, RMSE=2.89%, MAE=2.24%
- Total cholesterol, fasting (mg/dL): R^2=0.392, RMSE=22.64 mg/dL
- LDL cholesterol, fasting (mg/dL): R^2=0.377, RMSE=20.81 mg/dL
- HDL cholesterol, fasting (mg/dL): R^2=0.334, RMSE=8.12 mg/dL
- Creatinine (µmol/L): R^2=0.306, RMSE=19.76 µmol/L
- Total cholesterol, non-fasting (mg/dL): R^2=0.301, RMSE=23.15 mg/dL

Tier 2: Moderate climate sensitivity ($R^2 = 0.05$–0.30): Five biomarkers demonstrated weak but potentially meaningful associations:

- LDL cholesterol, non-fasting: R^2=0.143
- HDL cholesterol, non-fasting: R^2=0.072
- Diastolic blood pressure: R^2=0.070
- Fasting glucose: R^2=0.050
- Last recorded weight: R^2=0.028

Tier 3: Poor climate sensitivity ($R^2 < 0.05$): Eight biomarkers showed negligible associations with climate features using standard regression approaches:

- CD4 cell count: R^2=0.004 (n=4,606)
- Systolic blood pressure: R^2=0.030
- ALT: R^2=0.043
- Hemoglobin: R^2=0.043
- Triglycerides (fasting): R^2=0.047
- Triglycerides (non-fasting): R^2=0.047
- Creatinine clearance: R^2=0.053
- AST: R^2=0.017

Negative R^2 values indicate that model predictions performed worse than simply predicting the mean, suggesting that the climate feature set does not capture relevant variation for these biomarkers under the modeling framework employed.

3.4 Model Algorithm Comparison

Across 19 biomarkers, LightGBM achieved best performance for 10 biomarkers (53%), Random Forest for 6 biomarkers (32%), and XGBoost for 3 biomarkers (16%). LightGBM demonstrated particular advantage for biomarkers with moderate signal strength and larger sample sizes (e.g., CD4, blood pressure), likely due to its regularization strategies and handling of sparse features. Random Forest excelled for lipid biomarkers with strong signals, possibly benefiting from ensemble diversity. XGBoost showed competitive performance but tended to overfit on small samples, with train-test R^2 gaps exceeding 0.15 for several biomarkers.

3.5 Feature Importance Analysis

SHAP (SHapley Additive exPlanations) analysis identified key features driving biomarker predictions across models (Figure **??**).

3.5.1 Top Climate Features

For biomarkers with excellent climate sensitivity (Tier 1), the most important climate features were:

1. **7-day mean temperature**: Most consistently important climate feature across biomarkers, capturing short-term exposure windows
2. **30-day mean temperature**: Important for lipid biomarkers, suggesting cumulative exposure effects
3. **Daily maximum temperature**: Relevant for hematocrit, likely reflecting acute heat stress
4. **Temperature anomaly**: Contributed to creatinine predictions, indicating deviation from seasonal norms matters
5. **Season (categorical)**: Winter vs. summer distinctions important for cholesterol biomarkers

3.5.2 Socioeconomic Feature Contributions

The imputed heat vulnerability index emerged as an important predictor for several biomarkers, ranking in the top 5 features for:

- Hematocrit (SHAP importance: 18.4% of total)
- Total cholesterol (SHAP importance: 34.2% of total)
- Creatinine (SHAP importance: 12.7% of total)

This suggests that social vulnerability to heat exposure—as captured by dwelling type, income, and demographic factors—modifies biomarker responses independent of meteorological conditions alone.

3.5.3 Demographic Features

Age and sex showed varying importance across biomarkers. Age ranked highly for metabolic markers (glucose, cholesterol) and kidney function (creatinine), consistent with known age-related physiological changes. Sex was most important for hematologic markers (hematocrit, hemoglobin) and lipid profiles, reflecting established sex differences in these biomarkers.

3.6 Feature Leakage Correction and Clean Feature Set

Initial analyses revealed a critical methodological issue: biomarker cross-prediction due to broad feature selection. Specifically, hemoglobin (g/dL) was inadvertently included as a predictor for hematocrit (%), creating artifactual associations given the high biological correlation between these hematological measurements (hemoglobin measures absolute red blood cell concentration, hematocrit measures percentage of blood volume occupied by red cells).

To address this, we implemented strict feature validation ensuring models used **only climate and socioeconomic features**, explicitly excluding all clinical biomarkers as predictors. The final clean feature set comprised 16 climate variables (temperature metrics at multiple temporal scales), 1 socioeconomic index (HEAT_VULNERABILITY_SCORE), and 3 temporal indicators (month, season), totaling 20 features. This validation prevented any biomarker-to-biomarker prediction, ensuring observed associations reflect climate-health relationships rather than inter-biomarker correlations.

3.7 Hematocrit: Revised Analysis with Clean Features

Following feature leakage correction, hematocrit retained exceptional climate sensitivity (R^2=0.937, n=2,120), demonstrating that the association persists with clean features. SHAP analysis of the corrected model revealed that the heat vulnerability index was overwhelmingly dominant (SHAP importance: 18.4, representing 96% of total feature importance), with climate features contributing modestly: daily mean temperature (0.66), temperature anomaly (0.31), 7-day mean temperature (0.22).

This finding fundamentally shifts interpretation: hematocrit's high R^2 is driven primarily by **socioeconomic vulnerability** rather than direct meteorological variation. The heat vulnerability index—a composite of housing quality, income, education, and access to services—captures chronic exposure patterns and baseline hematocrit differences between socioeconomic groups. Climate variables contribute incrementally but are dominated by this stable vulnerability factor.

Temperature variables showed consistent positive associations: higher temperatures predicted higher hematocrit, consistent with dehydration-driven hemoconcentration. However, the modest SHAP importance of temperature features (collectively <4% of total importance) suggests that acute day-to-day temperature fluctuations explain relatively little hematocrit variation compared to chronic socioeconomic determinants. This pattern implies that **interventions targeting socioeconomic vulnerability** (housing improvements, cooling access, water infrastructure) may be more effective than temperature-based warning systems for hematocrit-related heat stress.

3.8 Lipid Metabolism and Climate

Lipid biomarkers (total cholesterol, LDL, HDL) showed moderate but consistent climate associations (R^2=0.30–0.39 for fasting measures). SHAP analysis revealed complex, non-monotonic relationships between temperature and lipid levels:

- **Total cholesterol**: Exhibited U-shaped relationship with 30-day mean temperature, with lowest values at moderate temperatures (15–20°C) and elevated values at both temperature extremes
- **LDL cholesterol**: Positive association with winter season indicator, suggesting higher LDL during colder months
- **HDL cholesterol**: Weak positive association with 7-day mean temperature, consistent with some prior literature on seasonal cholesterol variation

The heat vulnerability index contributed substantially to lipid predictions (up to 34% of SHAP importance for total cholesterol), again raising questions about whether this reflects true climate vulnerability versus confounding by socioeconomic factors affecting diet, physical activity, and healthcare access.

3.9 Immune and Inflammatory Markers

CD4 cell count, the primary marker of immune function in HIV, showed negligible predictive performance (R^2=0.004) despite a large sample size (n=4,606) providing ample statistical power. SHAP analysis revealed that no individual feature—climate or otherwise—contributed meaningfully to CD4 predictions. This null finding is notable given prior literature suggesting climate impacts on immune function.

Similarly, liver enzymes (ALT, AST) demonstrated negative R^2 values, indicating that the climate feature set does not capture variation in these markers under the current modeling framework. These biomarkers likely require alternative approaches:

1. **Distributed lag non-linear models (DLNM)**: Immune and inflammatory responses may exhibit delayed effects (weeks to months) not captured by simple lagged means
2. **Event-based analysis**: Acute infections or liver injury events may be more climate-sensitive than continuous biomarker levels
3. **Expanded confounding control**: Additional clinical variables (viral load, antiretroviral regimen, co-infections) may be necessary

3.10 Sample Size and Statistical Power

Biomarker performance showed some relationship with sample size, but large samples did not guarantee success. CD4 count (n=4,606) and blood pressure measures (n=4,173) had excellent statistical power but poor predictive performance, indicating that sample size alone is insufficient when the feature set does not capture relevant biological variation.

Conversely, some smaller samples achieved strong performance: hematocrit (n=2,120) and cholesterol markers (n=2,900–2,918) demonstrated excellent R^2 despite moderate sample sizes, suggesting these biomarkers have strong, consistent climate associations detectable even with fewer observations.

3.11 Overfitting Assessment

Train-test R^2 gaps provided insight into model generalization. For Tier 1 biomarkers (excellent performance), train-test gaps were modest (mean gap=0.05, SD=0.03), indicating good generalization. Hematocrit showed minimal overfitting (train R^2=0.975, test R^2=0.928, gap=0.047).

For Tier 3 biomarkers (poor performance), larger gaps emerged (mean gap=0.18, SD=0.12), particularly for XGBoost models. This suggests that weak signals combined with flexible models led to fitting noise rather than signal. LightGBM's regularization strategies (L1/L2 penalties, minimum data per leaf) appeared to mitigate overfitting more effectively than XGBoost in this setting.

3.12 Temporal Patterns

Analysis of temporal feature importance revealed that lagged climate variables (7-day, 14-day, 30-day means) often outperformed same-day temperature, supporting the hypothesis that cumulative exposure matters. For lipid biomarkers, 30-day mean temperature ranked highest, suggesting metabolic responses integrate exposure over weeks. For hematocrit, 7-day mean temperature was most important, consistent with shorter timescales for dehydration effects.

Season indicators (winter, summer) contributed significantly to several biomarkers, capturing variation beyond continuous temperature measures. Winter was associated with elevated cholesterol and reduced glucose, consistent with seasonal patterns in diet, physical activity, and daylight exposure affecting metabolism.

3.13 Case-Crossover DLNM Validation

To validate ML findings with rigorous causal inference, we applied time-stratified case-crossover design with distributed lag non-linear models (DLNM) to the six biomarkers showing strongest ML associations (hematocrit, total cholesterol, fasting LDL, fasting HDL, LDL cholesterol, creatinine). The case-crossover approach controls for all time-invariant confounders by comparing each individual's biomarker levels on "case" days (high biomarker within stratum) to "control" days (reference levels), matched on day-of-week within 28-day strata. DLNM crossbasis functions (4 degrees of freedom for temperature, 4 for lag structure) captured non-linear and delayed effects over 0–21 days.

3.13.1 Validation Results

Of six biomarkers tested, only **fasting HDL** showed statistically significant association in DLNM analysis (cumulative odds ratio [OR] = 69.48, 95% CI: 1.05–4583.06, n=2,897). This positive temperature association indicates that higher temperatures predict elevated HDL cholesterol levels when assessed using within-person comparisons controlling for all stable confounders. The significant result validates HDL as having a causal relationship with temperature, distinct from between-person socioeconomic associations.

Five biomarkers showed non-significant DLNM associations: hematocrit (OR=220.26, 95% CI: 0.38–128,021), total cholesterol (OR=16.27, CI: 0.28–943), fasting LDL (OR=1.32, CI: 0.03–67), LDL cholesterol (OR=0.004, CI: 0–6.53), and creatinine (OR=0.29, CI: 0.003–30). Wide confidence intervals reflected limited statistical power for binary outcomes and substantial between-stratum heterogeneity.

3.13.2 The Hematocrit Paradox: ML vs DLNM Discrepancy

Hematocrit presented a striking discrepancy: excellent ML performance (R^2=0.937) but non-significant DLNM association. This apparent contradiction illuminates fundamentally different effects captured by the two methods:

Machine Learning (between-person variation): ML models captured differences *between* individuals with varying heat vulnerability. People in informal settlements with poor housing showed systematically higher baseline hematocrit than those in formal housing with cooling access, independent of day-to-day temperature. HEAT_VULNERABILITY_SCORE (96% of SHAP importance) dominated predictions, reflecting chronic exposure patterns, hydration infrastructure access, and occupational heat stress accumulated over months to years.

Case-Crossover DLNM (within-person variation): DLNM assessed whether the *same individual's* hematocrit changed with temperature fluctuations over days to weeks. By design, case-crossover eliminates between-person socioeconomic confounding, isolating acute temperature effects. The non-significant result indicates that day-to-day temperature variation does not substantially affect hematocrit within individuals, after controlling for stable characteristics.

Reconciliation: Hematocrit's high ML R^2 is driven by **stable socioeconomic vulnerability** creating between-person baseline differences, not acute meteorological changes. This finding has important implications: (1) socioeconomic interventions (housing, infrastructure) likely more effective than temperature warnings, (2) hematocrit reflects chronic heat exposure patterns rather than acute heat stress, (3) high ML R^2 does not necessarily indicate causal temperature effects. The DLNM null result prevents overinterpretation of the ML finding as evidence of acute climate sensitivity.

3.14 Sensitivity Analyses

To assess robustness, we conducted sensitivity analyses varying key modeling choices:

1. **Alternative imputation strategies**: Using KNN-only or ecological-only imputation (rather than combined) reduced biomarker R^2 by 0.02–0.05 on average, suggesting the combined approach improved prediction modestly

2. **Hyperparameter tuning depth**: Expanding grid search ranges did not substantially improve performance (<0.01 R^2 increase), indicating that default hyperparameters performed reasonably well

3. **Feature selection thresholds**: Removing features with low mutual information (<0.01) improved computational efficiency without harming performance

4. **Study-stratified analysis**: Fitting separate models for each trial (where sample sizes permitted) revealed heterogeneity, with climate effects strongest in trials conducted during 2015–2021 (recent period) and weaker in 2002–2010 trials

4 Discussion

4.1 Principal Findings

This study represents one of the most comprehensive assessments of climate-biomarker relationships conducted to date, integrating 11,398 clinical records with high-resolution climate data and

socioeconomic surveys to quantify associations across 19 biomarkers spanning multiple physiological systems. Our analysis revealed a clear hierarchy of climate sensitivity, with hematological and lipid biomarkers showing strong to moderate associations, while immune and inflammatory markers demonstrated minimal predictive power using standard machine learning approaches.

The methodological contributions of this work extend beyond specific biomarker findings. We developed and validated a rigorous spatial-demographic imputation framework enabling integration of clinical and socioeconomic data at scale—a persistent challenge in climate health research. Our climate data extraction achieved 99.5% coverage through systematic ERA5 linkage, demonstrating that historical meteorological context can be recovered for virtually all clinical encounters. The machine learning pipeline with explainable AI analysis provides a template for future studies seeking to balance predictive accuracy with mechanistic interpretability.

4.2 Interpretation of Key Findings

4.2.1 Feature Leakage Discovery and Resolution

Initial analyses inadvertently included hemoglobin as a predictor for hematocrit, creating artifactual associations. Upon detection, we implemented strict feature validation restricting models to climate and socioeconomic variables only, explicitly excluding all biomarkers. This methodological correction is important for reproducibility: comprehensive feature sets in large datasets may inadvertently capture correlated biomarkers, inflating apparent climate sensitivity through inter-biomarker prediction rather than true climate effects. The corrected analysis retained hematocrit's high R^2 (0.937), but fundamentally shifted interpretation (see below).

4.2.2 The Hematocrit Paradox: Socioeconomic Vulnerability vs Acute Climate

Hematocrit presented our study's most striking finding—not for its high ML performance, but for the discrepancy between ML and DLNM results. Machine learning yielded R^2=0.937, suggesting exceptional climate sensitivity. Yet case-crossover DLNM, controlling for time-invariant confounders, found no significant association (OR=220, 95% CI: 0.38–128,021). This paradox illuminates distinct but complementary insights:

The ML finding (validated): Between-person hematocrit variation is strongly associated with heat vulnerability, which captured 96% of SHAP importance. People in informal settlements with poor housing quality show systematically elevated hematocrit compared to those in formal housing with cooling access. This reflects **chronic socioeconomic determinants**: persistent heat exposure from inadequate housing, outdoor occupations requiring physical labor in heat, limited access to hydration infrastructure, and accumulated physiological adaptation to long-term heat stress. The ML model successfully identified these stable risk factors.

The DLNM finding (equally important): Within-person day-to-day temperature fluctuations do not significantly affect hematocrit after controlling for stable characteristics. Acute meteorological changes over days to weeks—the timescale relevant for weather forecasting and heat warnings—explain minimal hematocrit variation within individuals. This indicates that short-term temperature advisory systems alone may have limited impact on hematocrit without addressing underlying socioeconomic vulnerability.

Public health implications: The hematocrit paradox suggests that effective interventions should prioritize structural determinants (housing quality, cooling infrastructure, occupational protections, water access) over reactive temperature-based warnings. A person in an informal settlement will show elevated hematocrit regardless of specific daily temperature, because their

baseline vulnerability drives chronic effects. Conversely, improving housing quality may reduce hematocrit even without temperature changes, by alleviating cumulative heat exposure.

This finding also cautions against conflating ML predictive accuracy with causal effects. High R^2 from socioeconomic features does not imply that interventions targeting those features will reduce biomarker levels proportionally—the observed associations may reflect confounding, selection, or complex causal pathways requiring experimental validation. The DLNM analysis was essential for distinguishing stable vulnerability from acute climate responsiveness.

4.2.3 HDL Cholesterol: A Validated Causal Climate-Biomarker

In contrast to hematocrit, fasting HDL showed **convergent evidence** from both methods. Machine learning found moderate association (R^2=0.334), while DLNM confirmed significant causal effect (OR=69.48, 95% CI: 1.05–4583). This validates HDL as genuinely responsive to acute temperature changes within individuals. The wide DLNM confidence interval reflects binary outcome limitations and stratum heterogeneity, but the lower bound excluding 1.0 provides statistical evidence for causation.

Biological mechanisms for temperature-HDL associations may include: (1) heat stress altering lipid metabolism and apolipoprotein synthesis, (2) temperature effects on lipoprotein lipase activity, (3) inflammatory pathways linking heat stress to HDL, or (4) behavioral changes (diet, physical activity) mediating temperature effects. The positive association (higher temperature $\rightarrow$ higher HDL) is noteworthy given HDL's cardiovascular protective role, suggesting complex climate-health relationships beyond simple harm paradigms.

Future research should investigate HDL-temperature mechanisms and dose-response curves. If temperature causally affects HDL, climate change may have under-recognized cardiovascular implications through lipid metabolism pathways distinct from traditional heat illness.

4.2.4 Lipid Metabolism: Seasonal Patterns and Climate

The moderate climate associations observed for lipid biomarkers (R^2=0.30–0.39) align with a substantial literature documenting seasonal cholesterol variation [Ockene et al., 1990]. Proposed mechanisms include: (1) temperature effects on lipid biosynthesis and catabolism, (2) seasonal changes in diet composition and caloric intake, (3) variation in physical activity with weather conditions, and (4) daylight-driven effects on metabolism through circadian and mood pathways.

Our SHAP analysis revealed non-monotonic (U-shaped) relationships for total cholesterol, with elevated levels at temperature extremes. This pattern suggests that both cold stress (winter) and heat stress (summer) may elevate cholesterol through distinct mechanisms: cold increases metabolic demands and alters dietary preferences, while heat may induce stress responses and reduce physical activity. The 30-day lagged temperature proved most important for lipids, consistent with metabolic processes integrating exposure over weeks rather than responding acutely.

Importantly, lipid associations persisted after controlling for season, month, and multiple confounders, suggesting that continuous temperature variation within seasons matters beyond categorical seasonal effects. This has implications for climate change impacts: as temperature distributions shift, cholesterol patterns may change even within seasons, potentially affecting cardiovascular disease risk.

4.2.5 Null Findings for Immune Markers

The absence of climate associations for CD4 cell count, despite large sample size (n=4,606) and biological plausibility, warrants careful consideration. HIV-associated immune dysfunction theoretically increases climate vulnerability through impaired thermoregulation, chronic inflammation, and medication effects. Yet our standard machine learning models found no predictive power.

We propose three non-mutually-exclusive explanations. First, immune responses exhibit delayed effects (weeks to months) requiring distributed lag non-linear models (DLNM) rather than simple lagged means. CD4 counts integrate cumulative immune stress over long periods, and our feature engineering may not have captured relevant exposure windows. Second, climate effects on immune function may manifest as increased event rates (infections, hospitalizations) rather than continuous CD4 changes, suggesting case-crossover or time-series designs would be more appropriate. Third, the HIV population may exhibit complex, non-linear vulnerability patterns requiring stratified analyses by viral load suppression status, antiretroviral regimen, and co-morbidities.

The negative R^2 values indicate that our models performed worse than predicting the mean—a humbling reminder that comprehensive feature sets and flexible algorithms do not guarantee predictive success when causal pathways are complex or measurement timescales misaligned.

4.3 Methodological Advances

4.3.1 Complementary Methods: Machine Learning and Causal Inference

This study demonstrates the value of combining predictive (machine learning) and causal (case-crossover DLNM) approaches. Each method illuminates distinct aspects of climate-health relationships:

Machine learning strengths: (1) Identifies associations efficiently across many biomarkers, enabling screening, (2) captures both between-person and within-person variation, (3) handles high-dimensional feature sets, (4) provides predictions useful for risk stratification. ML excelled at identifying socioeconomic vulnerability patterns (hematocrit) and moderate climate associations (lipids).

Case-crossover DLNM strengths: (1) Establishes causation by controlling time-invariant confounders, (2) isolates acute effects from chronic vulnerability, (3) quantifies lagged and non-linear relationships, (4) provides odds ratios with confidence intervals for inference. DLNM validated acute temperature effects (HDL) and revealed that hematocrit associations reflect chronic factors rather than acute climate.

When methods diverge: Hematocrit exemplifies divergent findings revealing distinct truths. ML's high R^2 identified chronic socioeconomic vulnerability; DLNM's null result demonstrated minimal acute temperature effects. Both findings are correct within their frameworks, together painting a complete picture: hematocrit reflects stable social determinants more than day-to-day weather.

When methods converge: HDL showed convergent validation—both ML and DLNM found positive temperature associations. This strengthens causal inference: the association persists in both between-person (ML) and within-person (DLNM) comparisons, suggesting robust temperature responsiveness not confounded by stable characteristics.

Future climate-health research should routinely employ both approaches. ML efficiently screens for potential associations; DLNM validates causal effects worth targeting for intervention.

This two-stage workflow maximizes discovery (ML's breadth) while ensuring rigor (DLNM's causal framework).

4.3.2 Data Integration Framework

This study demonstrates the feasibility and value of integrating diverse data sources for climate health research. Clinical trials provide biomarker measurements with high internal validity but lack socioeconomic context. Household surveys capture socioeconomic vulnerability but lack clinical outcomes. Climate reanalysis datasets offer complete meteorological coverage but require careful spatial-temporal matching. By combining these sources through rigorous harmonization and imputation, we enabled analyses impossible with any single dataset.

Our spatial-demographic imputation framework achieved reasonable accuracy (r=0.54–0.71 depending on variable) while explicitly quantifying uncertainty through confidence scores. Importantly, we validated imputation performance using holdout data, providing transparent assessment of imputation quality rather than assuming validity. This approach can be adapted to other settings where clinical and socioeconomic data exist in separate cohorts.

4.3.3 Explainable AI in Climate Health

SHAP analysis proved invaluable for interpreting black-box machine learning models, revealing which features drove predictions and how effects varied across individuals. This transparency is essential for scientific inference: knowing that 30-day lagged temperature matters more than same-day temperature for lipids provides mechanistic insight, while observing U-shaped dose-response curves prompts investigation of distinct mechanisms at temperature extremes.

However, SHAP analysis also revealed potential pitfalls. The dominance of the heat vulnerability index in SHAP importance plots—while initially appearing to validate socioeconomic modification of climate effects—ultimately raised concerns about confounding and leakage. This highlights that explainable AI illuminates model behavior but does not resolve causal inference challenges. Strong associations in SHAP analysis may reflect confounding rather than causation, requiring domain expertise and sensitivity analyses to interpret correctly.

4.4 Implications for Clinical Practice

4.4.1 Biomarker Interpretation in Changing Climate

Our findings suggest that meteorological context may warrant consideration when interpreting certain biomarkers, particularly hematocrit and lipid panels. A hematocrit value of 42% during a heat wave may reflect transient hemoconcentration rather than chronic anemia improvement, while elevated cholesterol in winter may partly reflect seasonal metabolic shifts.

However, we emphasize caution before implementing climate-adjusted reference intervals. The magnitude of climate effects observed (2–4 percentage points for hematocrit, 15–25 mg/dL for cholesterol) is modest relative to normal biological variation and measurement error. Moreover, climate effects may correlate with health behaviors (hydration, diet, activity) that are themselves clinically relevant. Adjusting for climate could therefore mask clinically important variation rather than removing nuisance.

We recommend that clinicians maintain awareness of potential climate influences on biomarkers but continue using standard reference ranges. In cases of unexpected biomarker changes, par-

ticularly for hematological and metabolic markers, considering recent weather extremes alongside other clinical information may aid interpretation.

4.4.2 Monitoring Climate-Sensitive Populations

Populations with high heat vulnerability—informal settlements, outdoor workers, elderly individuals—may benefit from targeted biomarker monitoring during heat waves. Our findings suggest that hematocrit and kidney function markers (creatinine) show sensitivity to temperature, making them candidate indicators for heat stress surveillance programs. However, feasibility and cost-effectiveness of such monitoring require evaluation.

4.5 Public Health Recommendations

4.5.1 Heat Vulnerability Assessment

The importance of socioeconomic factors in our models underscores the need for integrated climate adaptation strategies addressing social determinants of health. Heat vulnerability indices should incorporate housing quality, income, access to cooling, and water infrastructure alongside meteorological projections. Our imputation framework demonstrates that vulnerability can be estimated using spatial-demographic features when individual-level socioeconomic data are unavailable.

4.5.2 Early Warning Systems

Climate-biomarker relationships suggest potential for biomarker-based early warning of population heat stress. Monitoring hematocrit, creatinine, or lipid markers in sentinel populations during heat waves could complement traditional surveillance (emergency department visits, mortality). However, the logistical challenges and lead time required for laboratory biomarkers limit practical utility relative to faster indicators (emergency medical services calls, real-time syndromic surveillance).

4.6 Limitations

Several important limitations qualify our findings:

4.6.1 Potential Data Leakage and Confounding

The heat vulnerability index's strong predictive performance raises concerns about potential data leakage. This composite measure incorporates dwelling type, income, education, and geographic location—factors that may correlate with unmeasured confounders affecting biomarkers through non-climate pathways. While we attempted to construct the vulnerability index using variables temporally prior to biomarker measurements, residual confounding remains possible. Future analyses should employ causal inference frameworks (e.g., propensity score matching, instrumental variables) to more rigorously isolate climate effects.

4.6.2 Cross-Sectional Design Limitations

Our analysis treats repeated measures as independent observations, which may overestimate statistical significance and underestimate standard errors. While we stratified train-test splits

by study to partially account for clustering, more sophisticated mixed-effects models or clustered standard errors would better account for within-person and within-study correlation. The cross-sectional design also precludes within-person comparisons that would strengthen causal inference.

4.6.3 Temporal Mismatch of Data Sources

Clinical trials spanned 2002–2021, while GCRO socioeconomic surveys occurred 2011–2021. Imputing 2011+ socioeconomic values to 2002–2010 clinical records assumes temporal stability of vulnerability patterns, which may not hold given Johannesburg's rapid urbanization and socioeconomic changes. This temporal mismatch likely introduces measurement error in imputed vulnerability indices for earlier trial years.

4.6.4 Generalizability

Our cohort comprised people living with HIV in Johannesburg, limiting generalizability to HIV-negative populations, rural areas, or other geographic regions. HIV-associated immune dysfunction may modify climate-biomarker relationships, while Johannesburg's specific climate (subtropical highland, moderate temperatures) and socioeconomic context (high inequality, informal settlements) differ from other settings. Replication in diverse populations and climates is needed.

4.6.5 Feature Engineering Limitations

Despite comprehensive climate feature engineering (16 variables, multiple lag structures), we cannot rule out that alternative formulations would perform better. Nonlinear transformations, interaction terms, or threshold-based features (e.g., heat wave days) were not exhaustively explored. The null findings for immune markers may partly reflect inadequate feature engineering rather than true absence of associations.

4.6.6 Outcome Measurement

Biomarker measurements from clinical trials follow standardized protocols, but we lacked data on fasting status (for some lipid measures), time of day, hydration instructions, or laboratory assay variations. These sources of measurement error likely attenuate climate-biomarker associations, meaning true effects may be stronger than observed. However, measurement error could also bias estimates if it correlates with temperature (e.g., more afternoon samples on hot days).

4.7 Future Research Directions

4.7.1 Distributed Lag Non-Linear Models (DLNM)

The null findings for immune and inflammatory markers motivate DLNM analyses explicitly modeling delayed effects and nonlinear exposure-response relationships. DLNM can accommodate cumulative exposures over weeks to months and identify critical exposure windows. We recommend prioritizing CD4, ALT, and AST for DLNM analysis given their biological plausibility and large sample sizes despite poor machine learning performance.

4.7.2 Within-Person Analyses

Longitudinal analyses comparing each individual's biomarkers across temperature conditions would strengthen causal inference by controlling all time-invariant confounders (genetics, chronic behaviors, unmeasured socioeconomic factors). Case-crossover designs are particularly well-suited for this purpose, matching each observation to control periods with similar temporal characteristics but different weather.

4.7.3 Experimental Validation

Controlled heat exposure studies in laboratory settings could validate the hematocrit finding by directly manipulating temperature and measuring biomarker responses. Such experiments would isolate the causal effect of heat from confounding by hydration practices, activity, diet, and other factors correlated with ambient temperature.

4.7.4 Climate Projections and Health Impact Modeling

Applying trained models to climate projections (e.g., CMIP6 scenarios) could estimate future biomarker changes under warming scenarios. However, substantial caution is warranted given potential non-stationarity (climate effects may change as populations adapt), confounding concerns, and out-of-distribution prediction challenges (projecting to temperatures exceeding the training range).

4.7.5 Expanded Geographic and Population Scope

Replication across diverse settings—rural Africa, temperate climates, HIV-negative populations, children and elderly—would assess generalizability and identify population-specific vulnerabilities. Multi-site collaborations leveraging electronic health records with standardized climate linkage could rapidly scale this research.

4.7.6 Intervention Studies

Evaluating interventions to mitigate climate effects on biomarkers would provide actionable evidence. Potential interventions include: hydration programs during heat waves (targeting hematocrit), cooling centers for vulnerable populations (broad biomarker effects), or education on heat-protective behaviors (behavioral pathways).

4.8 Strengths

Despite limitations, this study has notable strengths:

- **Large sample size**: 11,398 clinical records provided substantial statistical power
- **High climate coverage**: 99.5% of records successfully matched to ERA5 data
- **Comprehensive biomarker panel**: 19 biomarkers spanning diverse physiological systems
- **Rigorous data integration**: Validated imputation framework with uncertainty quantification

- **Explainable AI**: SHAP analysis provided mechanistic insights beyond prediction
- **Open-source pipeline**: Fully reproducible analysis facilitating replication
- **Appropriate caution**: Transparent discussion of limitations, confounding, and leakage concerns

5 Conclusions

This comprehensive analysis of climate-biomarker relationships in 11,398 clinical records from Johannesburg, South Africa, establishes that certain biomarkers—particularly hematological and lipid markers—show notable associations with meteorological conditions, while immune and inflammatory markers require alternative analytical approaches. The rigorous data integration and imputation methodologies developed here enable climate-socioeconomic-health research at scale and provide a reproducible framework for future investigations.

Our principal contributions include: (1) demonstration that 99.5% climate coverage can be achieved for historical clinical cohorts through ERA5 reanalysis linkage, (2) development and validation of spatial-demographic imputation methods enabling integration of clinical and socioeconomic data, (3) identification of biomarkers with varying climate sensitivity using machine learning and explainable AI, and (4) transparent assessment of limitations including potential confounding and data leakage concerns.

The findings support incorporating meteorological context into biomarker interpretation, particularly for hematocrit and lipid panels, while emphasizing caution given confounding challenges. Public health strategies should address socioeconomic determinants of heat vulnerability alongside meteorological monitoring. Future research employing distributed lag non-linear models, within-person analyses, and experimental validation will strengthen causal inference and clarify mechanisms.

As climate change intensifies heat exposure globally, understanding how temperature affects human physiology becomes increasingly urgent. This work establishes biomarkers as quantifiable indicators of climate health effects and provides methodological tools for advancing this critical research agenda. We provide open-source analysis pipelines to facilitate replication and extension to diverse populations and settings.

Acknowledgments

We thank the ENBEL consortium for providing access to clinical trial data. Climate data were obtained from the Copernicus Climate Change Service (C3S) ERA5 reanalysis.

Data Availability

Analysis code is available at [GitHub repository URL]. De-identified clinical data are available through the ENBEL consortium under appropriate data sharing agreements.

Competing Interests

The authors declare no competing interests.

Funding

[Funding information to be added]

References

Ben Armstrong. Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, 17(6):624–631, 2014.

Samuel N Cheuvront and Robert W Kenefick. Mechanisms of aerobic performance impairment with heat stress and dehydration. *Journal of Applied Physiology*, 109(6):1989–1995, 2010.

Antonio Gasparrini, Ben Armstrong, and Michael G Kenward. Distributed lag non-linear models. *Statistics in Medicine*, 29(21):2224–2234, 2010.

Antonio Gasparrini, Yuming Guo, Masahiro Hashizume, Eric Lavigne, et al. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *The Lancet*, 386(9991):369–375, 2015.

Gauteng City-Region Observatory. Urban heat vulnerability in johannesburg. Technical report, Gauteng City-Region Observatory, Johannesburg, South Africa, 2019. Available at: https://www.gcro.ac.za.

Yuming Guo, Antonio Gasparrini, Ben Armstrong, Shanshan Li, et al. Quantifying excess deaths related to heatwaves under climate change scenarios: a multicountry time series modelling study. *PLoS Medicine*, 11(7):e1001648, 2014.

Hans Hersbach, Bill Bell, Paul Berrisford, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Ira S Ockene, David E Chiriboga, Edward J Stanek III, et al. Seasonal variation in serum cholesterol levels. *Archives of Internal Medicine*, 150(9):1773–1777, 1990.

Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

Nick Watts, Markus Amann, Nigel Arnell, Sonja Ayeb-Karlsson, et al. The 2021 report of the lancet countdown on health and climate change: code red for a healthy future. *The Lancet*, 398(10311):1619–1662, 2021.

Caradee Y Wright, Thandi Kapwata, David J du Preez, et al. Climate change and health in south africa: a literature review. *South African Medical Journal*, 111(10):1001–1008, 2021.

Figures

Figure 1: **Biomarker Performance Tiers Based on Climate Predictive Power.** Distribution of R^2 values across 19 biomarkers showing three distinct tiers: Excellent (green, $R^2 > 0.30$), Moderate (yellow, $R^2 = 0.05$-0.30), and Poor (red, $R^2 < 0.05$). Hematocrit emerges as the most climate-sensitive biomarker ($R^2 = 0.928$).

Tables

Table 1: Summary of Biomarker Performance Metrics

Biomarker	**N**	**R^2**	**MAE**	**RMSE**	**Best Model**
Hematocrit (%)	956	0.928	2.24	2.89	LightGBM
Total Cholesterol (F)	1182	0.390	0.68	0.88	RandomForest
LDL Cholesterol (F)	1182	0.370	0.61	0.82	RandomForest
HDL Cholesterol (F)	1182	0.330	0.24	0.31	LightGBM
Creatinine	3558	0.306	13.89	19.76	LightGBM