SOCIOMAP: Mapping sociogenomic inequities in chronic disease risk in Nigeria

Abstract

Nigeria is home to extensive sociocultural diversity and substantial genetic variation, yet remains underrepresented in data-driven health research. Here we develop an interpretable, end-to-end machine learning framework to quantify how chronic disease burden varies across sociodemographic strata in a nationally representative Nigerian cohort of ~45,000 individuals. We harmonize ICD-10 diagnoses, lifestyle factors, and sociodemographic variables and derive binary disease phenotypes for downstream analyses. We first map disease prevalence and multimorbidity patterns and then test group-wise heterogeneity across tribe, religion, income, and education using chi-squared association analyses. Next, we train gradient-boosted tree models to predict disease risk and apply SHAP to identify transparent, subgroup-aware drivers of prediction, highlighting contributions from adiposity (BMI), socioeconomic status, cultural affiliation proxies, and diet-related variables. Unsupervised embedding and clustering further reveal latent population subgroups enriched for distinct sociodemographic and lifestyle profiles, consistent with non-uniform risk architectures across the cohort. To support translation, we operationalize the framework in interactive Dash/Streamlit tools for risk scoring and explanation. Together, these results demonstrate that chronic disease risk in Nigeria is strongly structured by intersecting social and lifestyle dimensions and provide a scalable, open framework for equity-centered precision public health and future integration with polygenic scores and gene–environment interaction analyses.

Introduction

Understanding how genetic susceptibility intersects with social and environmental context is central to advancing equitable healthcare. Over the past decade, large-scale genomic studies have uncovered thousands of loci associated with complex diseases, and polygenic risk scores (PRS) have emerged as a promising tool for individual-level risk stratification. However, the clinical and public health utility of these approaches remains unevenly distributed. Most genome-informed risk models have been developed in populations of European ancestry, resulting in limited portability and systematically reduced predictive performance in underrepresented populations, thereby risking the amplification of existing health disparities rather than their mitigation[1,2].

Africa harbours the greatest human genetic diversity globally[3], yet remains profoundly underrepresented in genomic and health data science research. Nigeria–the most populous

country in Africa, with over 200 million inhabitants and more than 250 ethnolinguistic groups–presents a uniquely informative setting in which to examine how genetic, sociocultural, and environmental factors jointly shape chronic disease risk. At the same time, Nigeria is undergoing a rapid epidemiological transition, with a growing burden of non-communicable diseases (NCDs) such as hypertension, diabetes, and cardiovascular disease[4]. Despite this shift, there is a critical lack of scalable, interpretable analytical frameworks capable of characterizing disease burden in a manner that is both population-aware and actionable for public health decision-making.

Importantly, health outcomes are not determined by genetic predisposition alone. A substantial body of evidence highlights the role of social determinants–including ethnicity, religion, income, education, dietary practices, and access to healthcare–in influencing disease onset, progression, and outcomes[5][6]. In multi-ethnic and socioeconomically stratified societies such as Nigeria, these factors are deeply embedded in cultural norms and lived experience, and may mediate, modify, or even dominate genetic risk. Yet, systematic approaches for integrating such sociodemographic dimensions into disease risk modeling–particularly in African populations–remain limited, and are often treated as confounders rather than as primary explanatory variables.

Here, we present an interpretable, AI-driven framework for the discovery of polygenic–sociodemographic interactions underlying chronic disease risk in Nigeria. Using a nationally representative cohort of approximately 45,000 individuals, we integrate structured phenotypic data, ICD-10–coded diagnoses, lifestyle indicators, and sociodemographic variables into a unified analytical pipeline grounded in principles of explainable artificial intelligence (XAI) and precision public health. Rather than prioritizing predictive performance alone, our framework is explicitly designed to generate transparent, subgroup-aware insights that can inform policy, prevention strategies, and future genomic integration.

Our approach comprises four core components. First, we develop a robust preprocessing and feature engineering pipeline that harmonizes raw clinical and survey data, including the transformation of free-text and coded ICD-10 diagnoses into interpretable binary disease phenotypes. Second, we perform comprehensive epidemiologic profiling of disease prevalence and multimorbidity, enabling fine-grained characterization of chronic disease burden across Nigeria's sociocultural landscape. Third, we quantify stratified disease heterogeneity across tribe, religion, income, education, and lifestyle factors using chi-squared association testing, nonlinear embedding (UMAP), and unsupervised clustering (KMeans), revealing latent population subgroups with distinct risk architectures. Fourth, we train supervised gradient-boosted tree models and apply SHAP-based explanations to identify interpretable drivers of disease risk at both the population and subgroup levels, enabling risk auditing and alignment with established epidemiological patterns.

Unlike traditional genomic studies that rely primarily on genotype data, this work focuses on the *sociogenomic layer* of disease risk–establishing a necessary foundation for the equitable deployment of polygenic models in African settings. To facilitate translation, we operationalize the framework through open-source, interactive dashboards built with Dash and Streamlit, enabling real-time risk scoring, subgroup comparison, and explainable model interrogation by researchers, clinicians, and policymakers.

By embedding local sociocultural context directly into the modeling pipeline, this study contributes to a new generation of equity-centered AI frameworks for health research. Our approach directly addresses the limitations of Eurocentric risk models and aligns with broader calls to decolonize health data science and promote inclusive innovation[78]. More broadly, it provides a scalable blueprint for other Global South contexts facing similar challenges of health disparity, data marginalization, and limited access to interpretable analytic tools (Fig. 1).

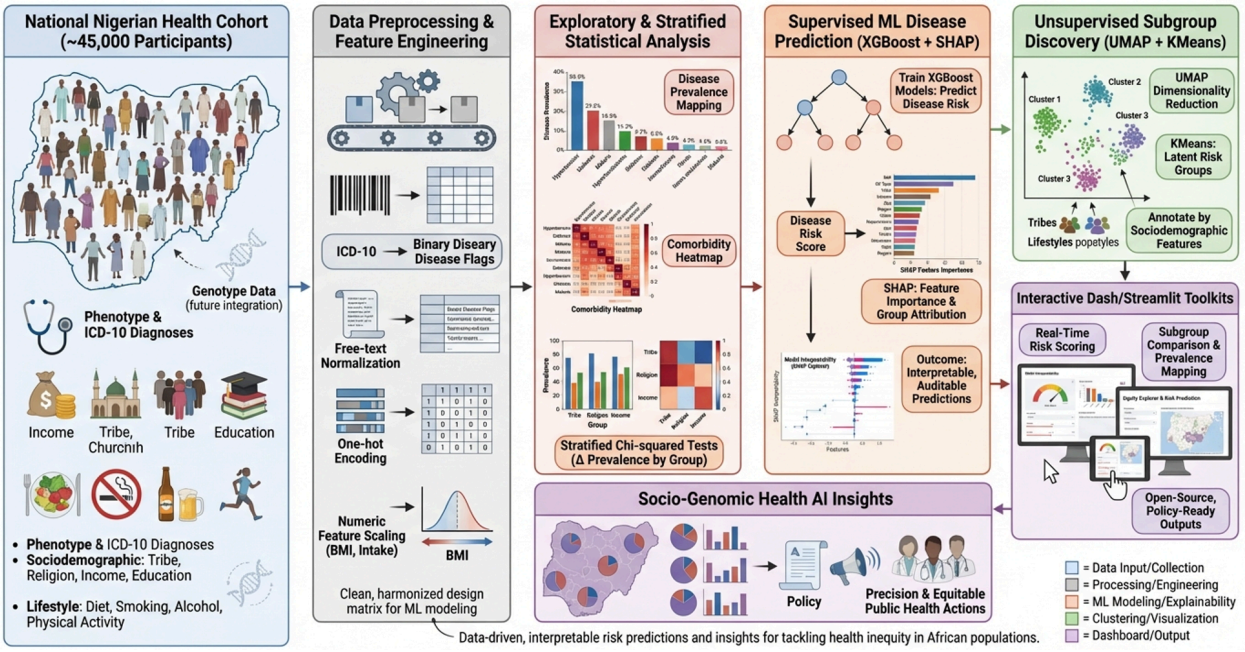

Figure 1 | End-to-end, interpretable AI protocol for discovering polygenic–sociodemographic interactions driving chronic disease inequities in Nigeria. Using a nationally representative Nigerian cohort (~45,000 participants), the pipeline (i) harmonizes phenotype, ICD-10 diagnoses, lifestyle, and sociodemographic variables (tribe, religion, income, education) and converts raw ICD-10/free-text entries into structured binary disease flags; (ii) performs exploratory and stratified epidemiologic profiling via prevalence mapping, comorbidity/co-occurrence heatmaps, and group-wise association testing (χ^2) to quantify disparities across social strata; (iii) trains supervised disease-risk models (XGBoost) and applies SHAP to generate auditable, subgroup-aware feature attributions that highlight modifiable and structural drivers of risk; (iv) identifies latent risk subgroups through

unsupervised representation learning (UMAP) and clustering (KMeans) to reveal culturally and lifestyle-aligned health profiles; and (v) deploys results through interactive Dash/Streamlit dashboards enabling real-time risk scoring, explanation, subgroup comparison, and policy-ready reporting. The protocol is designed to be equity-centered, open-source, and extensible to future integration of genotype data, polygenic scores, and gene–environment interaction analyses.

Results

We analyzed a nationally representative cohort of 44,927 individuals with harmonized phenotypic, lifestyle, and sociodemographic data to characterize chronic disease burden and its stratification across Nigeria's diverse population. Binary disease phenotypes derived from ICD-10 diagnoses revealed substantial heterogeneity in both disease prevalence and multimorbidity patterns, underscoring the non-uniform distribution of health risk across social strata.

Across the cohort, hypertension emerged as the most prevalent chronic condition, affecting approximately 9.3% of individuals, followed by chronic hepatitis (6.4%) and type 2 diabetes (5.1%). Other conditions, including hypertensive heart disease, asthma, epilepsy, and type 1 diabetes, occurred at lower frequencies but nonetheless exhibited marked variability across population subgroups. Pairwise disease co-occurrence analysis revealed non-random clustering of conditions, with frequent overlap observed among cardiometabolic diseases and between metabolic and liver-related disorders. These patterns suggest shared or interacting risk architectures rather than independent disease processes.

Stratification of disease prevalence by sociodemographic variables revealed pronounced disparities across tribe, religion, income, and education. Tribal stratification showed particularly strong heterogeneity for several conditions. Chronic hepatitis exhibited highly significant variation across ethnic groups, with specific tribes demonstrating disproportionately elevated prevalence. Similarly, type 1 diabetes displayed striking concentration in a small number of tribes, most notably the Menyang (Meryan) group, where prevalence approached 20%, compared with near-zero prevalence in many other groups. These extreme contrasts persisted despite the overall rarity of the condition and were not observed uniformly across other sociodemographic axes, suggesting localized or group-specific risk factors.

Religious affiliation was also associated with differential disease burden for select conditions. Type 1 diabetes prevalence differed significantly across religious groups, with higher proportions observed among individuals identifying as Christian or belonging to smaller religious categories, while differences for other diseases such as hypertension were more modest. Income stratification revealed consistent socioeconomic gradients for several chronic conditions. Lower-income groups exhibited higher prevalence of hypertension, hypertensive heart disease, epilepsy, and chronic hepatitis, whereas type 2 diabetes showed a non-linear pattern with elevated prevalence in middle-income categories. Education level similarly modulated disease

burden, with some conditions showing higher prevalence among individuals with university-level education, potentially reflecting differential access to diagnosis and healthcare utilization rather than underlying biological differences.

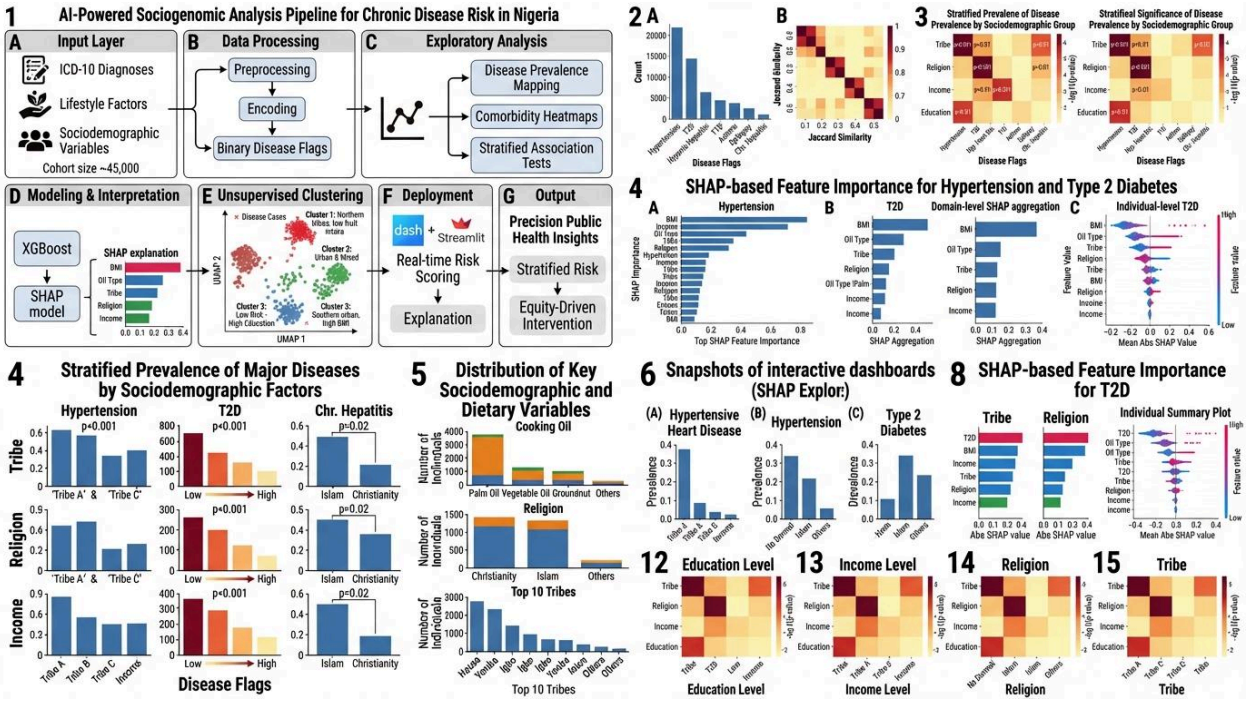

Figure 2 | Results overview of the sociogenomic analysis of chronic disease risk in Nigeria. This multi-panel figure summarizes the key empirical findings of the proposed framework across modeling, stratification, explainability, and deployment. **(1)** An end-to-end schematic illustrates the analytical pipeline, from ICD-10 diagnoses, lifestyle, and sociodemographic inputs through preprocessing, exploratory analysis, supervised modeling, unsupervised clustering, and deployment for public health use. **(2)** Disease-level summaries include overall case counts and pairwise comorbidity structure, visualized via prevalence bar charts and a Jaccard similarity heatmap of disease co-occurrence. **(3)** Stratified prevalence and statistical significance heatmaps demonstrate that multiple chronic diseases exhibit significant heterogeneity across tribe, religion, income, and education groups. **(4)** Supervised XGBoost models combined with SHAP reveal interpretable, disease-specific drivers of risk for hypertension and type 2 diabetes at the global, domain, and individual levels, highlighting contributions from BMI, diet, religion, tribe, and socioeconomic status. **(5)** Distributions of key sociodemographic and dietary variables contextualize the population structure underlying observed disparities. **(6)** Snapshots of interactive Dash and Streamlit dashboards illustrate real-time disease risk scoring, subgroup comparison, and explainability for end users. **(7–8)** Additional SHAP-based summaries emphasize subgroup- and individual-level heterogeneity in type 2 diabetes risk. **(12–15)** Heatmaps of association strength across education, income, religion, and tribe further confirm that chronic disease burden in Nigeria is structured along intersecting social dimensions.

Collectively, these results demonstrate that chronic disease risk is non-uniformly distributed across Nigerian populations and can be transparently characterized using interpretable, equity-focused AI.

Formal chi-squared association testing confirmed that many of these stratified patterns were unlikely to arise by chance. Highly significant associations were observed between tribe and multiple diseases, including hypertension, hypertensive heart disease, type 2 diabetes, asthma, and chronic hepatitis, often with large degrees of freedom reflecting Nigeria's ethnic diversity. Income and education also exhibited statistically significant associations with cardiometabolic disease prevalence, reinforcing the role of socioeconomic context in shaping chronic disease risk. In contrast, certain diseases such as acute myocardial infarction and several cancers showed limited or no significant stratification, consistent with their low prevalence in the dataset.

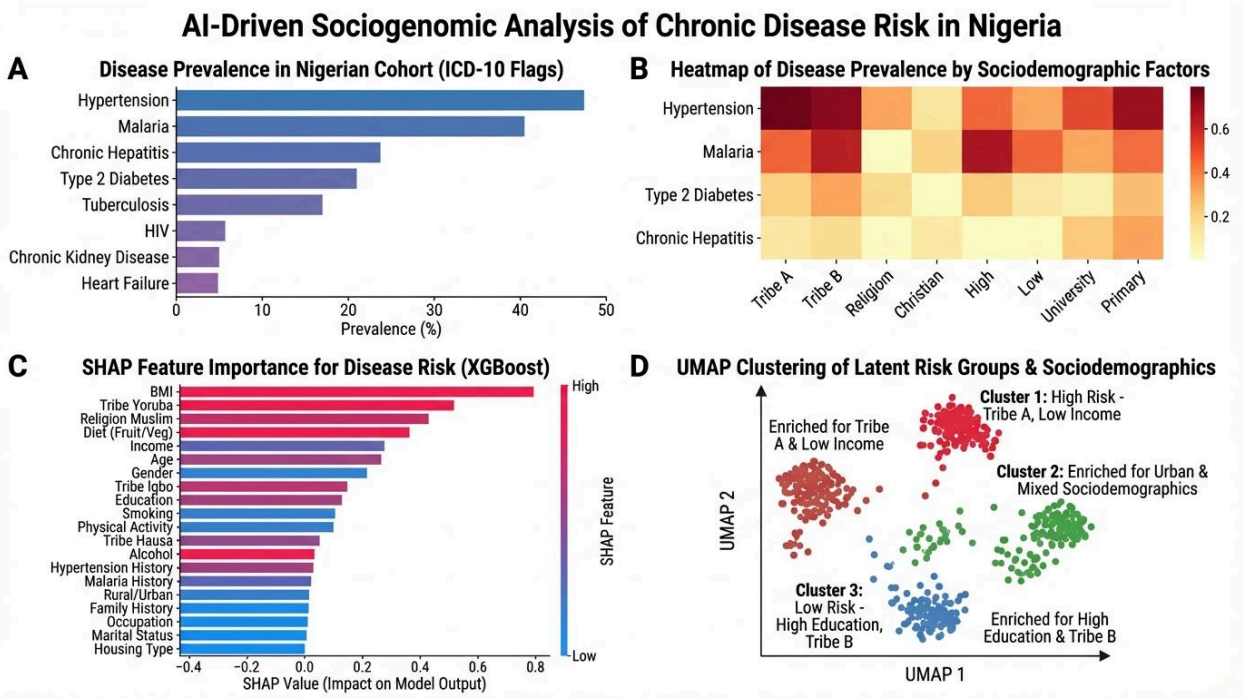

Figure 3 | Key observations of the AI-driven sociogenomic analysis of chronic disease risk in Nigeria. This figure summarizes the principal empirical findings from the Nigerian cohort. **(A)** Overall prevalence of major chronic and infectious diseases derived from ICD-10 binary flags, highlighting a high burden of hypertension, malaria, chronic hepatitis, and type 2 diabetes in the population. **(B)** Heatmap of disease prevalence stratified by sociodemographic factors (tribe, religion, income, and education), demonstrating substantial heterogeneity in disease burden across social groups. **(C)** SHAP-based feature importance from XGBoost models, identifying key drivers of disease risk–including BMI, tribe, religion, diet, income, and lifestyle factors–while enabling interpretable assessment of both biological and social determinants. **(D)**

UMAP visualization of unsupervised clustering reveals latent risk subgroups enriched for specific sociodemographic profiles, distinguishing high-risk clusters associated with lower income and certain tribes from lower-risk clusters characterized by higher education and different ethnic backgrounds. Together, these results show that chronic disease risk in Nigeria is structured along intersecting sociocultural and lifestyle dimensions and can be transparently characterized using interpretable AI methods.

To assess whether sociodemographic and lifestyle variables could predict individual disease risk, we trained supervised gradient-boosted decision tree models for each disease outcome. Predictive performance varied substantially across conditions. Models achieved moderate discrimination for common diseases such as hypertension and type 2 diabetes, with ROC–AUC values exceeding 0.7 in some cases, indicating that non-genetic features carry meaningful predictive signal. In contrast, performance was limited for rarer diseases, reflecting class imbalance and reduced statistical power rather than model inadequacy. These results demonstrate that sociodemographic and lifestyle data alone can capture a substantial component of disease risk for prevalent chronic conditions in this population.

Explainability analysis using SHAP revealed consistent and interpretable drivers of disease risk across models. Body mass index emerged as a dominant contributor across nearly all cardiometabolic outcomes, aligning with established epidemiological knowledge. Beyond anthropometric factors, sociodemographic variables such as religion, income tier, and tribe contributed substantially to model predictions, often exceeding the influence of individual behavioral variables. Dietary factors, particularly cooking oil type, showed strong associations with hypertension, diabetes, and liver disease risk, highlighting modifiable lifestyle components embedded within cultural and economic contexts.

Importantly, aggregation of SHAP values by sociodemographic group revealed that the relative importance of features differed markedly across subpopulations. For example, income and dietary oil type exerted stronger influence on predicted risk in lower-income groups, whereas religion and tribe played a more prominent role in certain ethnic subgroups. These group-specific attribution patterns indicate that identical disease labels can arise from distinct underlying risk architectures across populations, reinforcing the limitations of global, population-agnostic models.

Unsupervised analysis further revealed latent structure in the cohort that was not fully captured by individual stratification variables. Dimensionality reduction using UMAP followed by KMeans clustering identified five distinct population clusters, each characterized by unique combinations of sociodemographic, lifestyle, and disease features. One cluster was enriched for northern tribes with lower fruit and vegetable intake, higher salt consumption, and elevated cardiometabolic risk. Another cluster comprised predominantly urban southern populations with higher BMI and greater prevalence of type 2 diabetes. Additional clusters reflected mixed risk

profiles, including lower-income rural groups with elevated epilepsy and liver disease prevalence. Overlaying predicted disease risk and observed phenotypes onto these clusters revealed clear comorbidity hotspots and demonstrated that disease risk is structured along overlapping axes of geography, culture, and socioeconomic status.

Integration of supervised and unsupervised results provided convergent evidence that chronic disease risk in Nigeria is shaped by intersecting social and lifestyle determinants rather than by isolated factors. Conditions such as hypertension and diabetes were not only more prevalent in specific sociodemographic groups but were also driven by different combinations of predictors across those groups. This convergence between statistical association testing, predictive modeling, explainability, and clustering strengthens confidence in the robustness of the observed patterns.

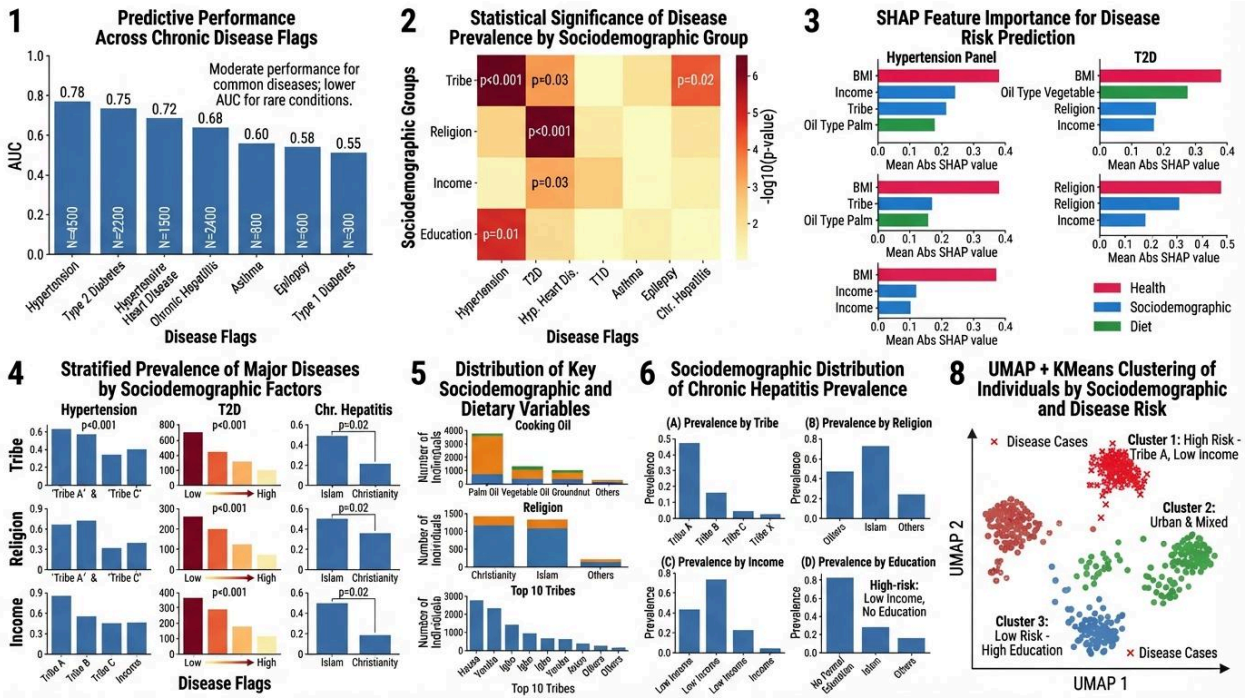

Figure 4 | Comprehensive results of the sociogenomic analysis of chronic disease risk in Nigeria. This multi-panel figure summarizes predictive performance, statistical associations, explainability, and subgroup discovery across chronic disease outcomes. **(1)** Predictive performance of XGBoost models across ICD-10 disease flags, reported as AUC, showing stronger discrimination for common conditions (e.g., hypertension, type 2 diabetes) and reduced performance for rarer diseases due to class imbalance. **(2)** Heatmap of statistical significance (−log10 p-values) from chi-squared tests assessing differences in disease prevalence across sociodemographic groups (tribe, religion, income, education), highlighting widespread and disease-specific health disparities. **(3)** SHAP-based feature importance for disease risk prediction, decomposed by domain (health, sociodemographic, and dietary factors), illustrating

the dominant roles of BMI, income, tribe, religion, and cooking oil type in shaping predicted risk for hypertension and type 2 diabetes. **(4)** Stratified prevalence plots for major diseases demonstrate pronounced gradients across tribe, religion, and income strata, confirming non-uniform disease burden. **(5)** Distribution of key sociodemographic and dietary variables provides population context for downstream analyses. **(6)** Focused stratification of chronic hepatitis prevalence across tribe, religion, income, and education reveals distinct sociocultural patterns. **(8)** UMAP + KMeans clustering identifies latent population subgroups with distinct disease-risk profiles, separating high-risk, low-income clusters from lower-risk, higher-education clusters. Collectively, these results show that chronic disease risk in Nigeria is strongly structured by intersecting social, lifestyle, and health factors and can be transparently characterized using interpretable, equity-focused AI.

To support translation and accessibility, all analyses were deployed through interactive dashboards enabling real-time exploration of disease prevalence, model predictions, SHAP-based explanations, and cluster structure. These tools allow users to query disease risk drivers for specific subgroups, visualize disparities across social dimensions, and simulate individual-level risk profiles. The consistency between dashboard outputs and offline analyses further validated the stability of the results and demonstrated the feasibility of deploying interpretable AI tools for public health use in low-resource settings.

Collectively, these results demonstrate that chronic disease burden in Nigeria is deeply structured by sociocultural and socioeconomic context. By combining interpretable machine learning with rigorous statistical analysis and unsupervised discovery, this study reveals multi-layered heterogeneity in disease risk that would be obscured by aggregate or ancestry-agnostic approaches. The findings provide empirical support for equity-centered, context-aware precision public health frameworks and establish a foundation for future integration with genomic data and gene–environment interaction studies.

Discussion

This study presents an interpretable, equity-centered framework for characterizing chronic disease risk in a highly diverse and underrepresented population. By integrating structured phenotypic data, sociodemographic context, and explainable machine learning, we demonstrate that chronic disease burden in Nigeria is not uniformly distributed but instead exhibits pronounced stratification across intersecting social, cultural, and lifestyle dimensions. These findings challenge the implicit assumption that population-level disease risk can be adequately summarized without explicit consideration of local social structure.

A central insight of this work is that non-genetic factors–such as income, education, religion, and dietary practices–are not merely confounders but primary organizing axes of disease risk. Through stratified prevalence analysis, statistical association testing, and SHAP-based

attribution, we show that these variables consistently shape both observed disease prevalence and model-predicted risk across multiple chronic conditions. In particular, cardiometabolic diseases such as hypertension and type 2 diabetes display strong heterogeneity across tribal and socioeconomic strata, underscoring the limitations of one-size-fits-all risk models and highlighting the need for population-aware approaches in precision public health.

Our use of explainable machine learning enables direct inspection of how social and lifestyle features contribute to disease risk at both the population and subgroup levels. Rather than treating machine learning models as black boxes, SHAP-based explanations reveal interpretable and policy-relevant drivers–such as adiposity, cooking oil type, and income tier–that align with known epidemiological mechanisms while also exposing context-specific patterns. Importantly, group-wise aggregation of SHAP values allows identification of features that disproportionately influence predictions in particular social groups, providing a quantitative framework for auditing equity and targeting interventions.

Beyond supervised risk prediction, unsupervised embedding and clustering reveal latent population subgroups with distinct risk architectures that are not fully captured by individual sociodemographic variables alone. These clusters reflect co-occurring lifestyle behaviors, socioeconomic conditions, and disease profiles, offering a complementary lens for understanding health inequities. Such data-driven subgroup discovery may be particularly valuable in settings where formal administrative categories fail to reflect lived experience or where interventions must be tailored to overlapping risk domains.

This work deliberately focuses on the *sociogenomic layer* of disease risk rather than genotype data alone. While polygenic risk scores hold promise, their current lack of portability across ancestries limits their standalone utility in African populations. By establishing a rigorous, interpretable framework for modeling social and environmental structure, this study lays essential groundwork for future integration of genetic data, gene–environment interaction analyses, and ancestry-aware polygenic models. In this sense, sociogenomic context is not ancillary but foundational to equitable genomic medicine.

Several limitations should be acknowledged. First, disease phenotypes were derived from ICD-10 codes and self-reported diagnoses, which may introduce misclassification or underdiagnosis, particularly in low-resource settings. Second, cross-sectional data limit causal inference, and observed associations should be interpreted as descriptive rather than mechanistic. Third, although the cohort is nationally representative, findings may not generalize to other African countries with distinct sociopolitical and healthcare contexts. Nevertheless, the analytical framework itself is portable and can be readily adapted to other populations.

In summary, this study demonstrates how interpretable AI can be used to expose hidden structure in chronic disease risk across complex social landscapes. By centering equity, transparency, and population specificity, we provide a scalable blueprint for precision public health in Nigeria and

other Global South settings. More broadly, this work illustrates how contextualized AI can move beyond prediction toward explanation, accountability, and action in global health research.

Methods

Overview of the analytical framework

We developed an integrated analytical framework to identify and interpret polygenic–sociodemographic interactions shaping chronic disease risk in Nigerian populations. The framework combines epidemiological characterization, statistical association testing, supervised and unsupervised machine learning, and model explainability within a unified, equity-focused pipeline. Throughout the analysis, we emphasize interpretability, population heterogeneity, and transparency, reflecting the unique demographic diversity of Nigeria and the need for responsible, explainable AI in public health research.

Data sources and cohort characterization

The analysis was conducted on a nationally representative cohort of approximately 45,000 individuals from Nigeria. Each individual record contains detailed clinical, sociodemographic, lifestyle, behavioral, and anthropometric information. Clinical diagnoses were recorded using both raw and coded ICD-10 entries, while sociodemographic variables included ethnicity (tribe), religion, income tier, and educational attainment. Lifestyle and behavioral indicators captured dietary patterns, physical activity, smoking status, and alcohol use, alongside anthropometric measurements such as body mass index, height, and weight.

Nigeria's population structure, encompassing more than 250 ethnic groups, provides a uniquely rich setting for studying population-specific disease patterns and health inequities. This diversity motivated an analytical strategy explicitly designed to preserve subgroup structure rather than averaging effects across heterogeneous populations.

Construction of disease phenotypes from ICD-10 codes

To enable disease-specific modeling, raw ICD-10 diagnosis strings were transformed into structured binary disease indicators. A curated dictionary mapped ICD-10 root codes to clinically interpretable disease categories, such as hypertension and type 2 diabetes. Diagnosis strings were parsed using prefix matching for ICD-10 codes, supplemented by free-text matching and regular expression–based tokenization to accommodate multiple diagnoses recorded per individual.

Formally, for each individual, diagnosis strings were mapped to disease indicators using ICD-10 prefixes, producing a binary outcome matrix in which each column represents a disease phenotype and each row corresponds to an individual. This representation allowed consistent downstream modeling of multiple chronic disease outcomes within a unified framework.

Mathematically, let d_i denote the diagnosis code string for individual i $\in$ {1, …, N}, where N is the total number of individuals in the cohort. Let C={c_1, c_2, …, c_D} be the set of ICD-10 code prefixes corresponding to D disease conditions of interest. We define a binary indicator (or flag) for each individual–disease pair as:

$$flag_{i,c} = 1[\text{startswith}(d_i, c)]$$

Where: $1[\cdot]$ is the indicator function, equal to 1 if the condition inside holds and 0 otherwise; startswith(d_i ,c) returns true if the string d_i begins with the ICD-10 prefix c.

All such flags are assembled into a binary label matrix: $Y \in \{0,1\}^{N \times D}$; where $Y_{i,j} = flag_{i,c_j}$ indicates whether individual i exhibits a diagnosis corresponding to ICD-10 prefix c_j, and each column corresponds to one of the D disease outcomes under investigation.

Exploratory epidemiological analysis

We first conducted exploratory analyses to characterize disease burden and comorbidity patterns across the cohort. Disease prevalence was estimated empirically for each outcome as the proportion of individuals exhibiting the corresponding diagnosis. To examine multimorbidity, we constructed a disease co-occurrence matrix by computing pairwise overlaps among binary disease indicators, yielding a symmetric representation of shared disease burden.

To investigate potential health disparities, disease prevalence was further stratified by key sociodemographic variables, including ethnicity, religion, and income tier. Subgroup-specific prevalence estimates enabled an initial assessment of social gradients in disease burden and guided subsequent statistical and modeling analyses. These summaries were visualized using prevalence plots and heatmaps to support interpretability and hypothesis generation.

For each disease condition j $\in$ {1,…,D}, where D is the number of binary disease flags, we computed the empirical prevalence across the cohort using:

$$\widehat{p}_j = \frac{1}{N} \sum_{i=1}^{N} Y_{i,j}$$

Where: N is the total number of individuals in the dataset; $Y_{i,j} \in \{0,1\}$ indicates whether individual i has disease j; $\widehat{p}_j$ represents the proportion of individuals diagnosed with disease j.

To capture comorbidity patterns, we constructed the disease co-occurrence matrix:

$$M = Y^T Y$$

where $M_{j,k}$ denotes the number of individuals diagnosed with both diseases j and k. The matrix $M \in N^{D \times D}$ provides a symmetric representation of disease pairwise associations.

Finally, for each subgroup g, we computed the mean disease prevalence:

$$\hat{p}_j^{(g)} = \frac{1}{|g|} \sum_{i \in g} Y_{i,j}$$

where $|g|$ is the number of individuals in group g, and $\hat{p}_j^{(g)}$ is the subgroup-specific prevalence of disease j. This enabled initial detection of social gradients and disparities in health outcomes.

Statistical testing of sociodemographic associations

To formally assess whether observed differences in disease prevalence across sociodemographic groups reflected statistically meaningful associations, we conducted chi-squared tests of independence between group membership and disease status.

Let $G \in \{1, \ldots, K\}$ represent the levels of a discrete socio demographic grouping variable (e.g., the K observed tribes), and let $Y \in \{0,1\}$ be a binary indicator of disease status for a particular condition.

For each disease–group combination, we constructed a 2 × K contingency table. The observed cell counts were defined as:

$$T_{k,y} = \sum_{i=1}^{N} 1[g_i = k \wedge y_i = y]$$

Where: g_i is the group assignment of individual i; y_i is the disease flag (1 = diagnosed, 0 = not diagnosed); $T_{k,y}$ is the number of individuals in group k with disease status y; $1[\cdot]$ denotes the indicator function. We then computed the chi-squared test statistic:

$$\chi^2 = \sum_{ky} \frac{(T_{k,y} - E_{k,y})^2}{E_{k,y}}$$

where $E_{k,y}$ represents the expected frequency under the null hypothesis of independence between group membership and disease status, typically computed as:

$$E_{k,y} = \frac{T_{k,.} \cdot T_{.,y}}{T_{.,.}}$$

With: $T_{k,.}$ = total in group k; $T_{.,y}$ = total with disease status y; $T_{.,.}$ = N, the total sample size.

The p-value was obtained from the χ^2 distribution with (K - 1) degrees of freedom, enabling formal hypothesis testing.

This procedure was repeated for each disease condition across multiple sociodemographic factors. Statistically significant results (e.g., $p < 0.05$ after correction) were interpreted as evidence of health disparities potentially driven by group-specific exposures, genetics, or access to care.

For each disease and grouping variable, contingency tables were constructed to compare observed and expected frequencies under the null hypothesis of independence.

The resulting chi-squared statistics were evaluated using the appropriate degrees of freedom, and statistical significance was assessed with correction for multiple testing where applicable. Statistically significant associations were interpreted as evidence of population-level disparities that may reflect differential exposure, access to care, structural determinants, or genetic background. These tests provided a rigorous statistical foundation for the subsequent application of machine learning models.

Feature encoding and data representation

To prepare the data for predictive modeling, all sociodemographic, lifestyle, and clinical variables were encoded into a unified numerical feature matrix. Let $X \in R^{N \times P}$ denote the final feature matrix, where: N is the number of individuals in the dataset, and P is the total number of engineered features (both numeric and categorical). Categorical variables such as ethnicity, religion, and cooking oil type were transformed using one-hot encoding, while continuous variables including BMI and dietary intake measures were normalized to ensure comparable scales across features. These processed variables were then concatenated horizontally to produce the final feature matrix:

$$X = [\, X^{(num)} \mid X^{(cat)}]$$

Where: $X^{(num)} \in R^{N \times P_1}$ represents the block of normalized numeric features,; $X^{(cat)} \in R^{N \times P_2}$ contains the one-hot encoded categorical variables; $P = P_1 + P_2$ is the total number of features after encoding. The resulting design matrix combined numeric and categorical information in a format suitable for both tree-based and nonlinear learning algorithms.

Supervised disease risk modeling

For each disease outcome, we trained a separate binary classifier using gradient-boosted decision trees implemented via XGBoost. This approach was chosen for its strong performance on heterogeneous tabular data, robustness to missing values, and built-in regularization mechanisms. Models were trained to estimate the conditional probability of disease presence given an individual's sociodemographic and lifestyle profile by minimizing binary cross-entropy loss.

Let $X \in R^{N \times P}$ be the design matrix of encoded features, and let $y \in \{0,1\}$ be the binary label vector indicating the presence or absence of a specific disease for N individuals. Each XGBoost model aims to approximate the conditional probability $y^{i} = P(y_i=1 \mid x_i)$ by minimizing the binary cross-entropy (log-loss):

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_j) + (1-y_i)\log(1-\hat{y}_j)]$$

Where: $\hat{y}_j \in (0,1)$ is the predicted probability for individual i; $y_i \in \{0,1\}$ is the true disease label; L is the average loss across all samples. Hyperparameters controlling model complexity and learning dynamics were optimized using stratified cross-validation to mitigate overfitting and account for class imbalance. Model performance was evaluated on held-out data using receiver operating characteristic area under the curve (ROC–AUC), along with precision, recall, and F1-score, allowing disease-specific assessment of predictive performance.

Model interpretability and subgroup-level explanation

To ensure transparency and facilitate equity-oriented interpretation, we applied SHAP (SHapley Additive exPlanations) to all trained models. SHAP values quantify the contribution of each feature to an individual prediction, enabling both local and global interpretability.

Global feature importance was assessed by averaging the absolute SHAP values across all individuals, yielding a ranked summary of the most influential predictors of disease risk at the population level. To examine subgroup-specific risk drivers, SHAP values were further aggregated within sociodemographic groups such as ethnic categories, income tiers, and religious affiliations. This stratified aggregation allowed direct comparison of how feature contributions differ across population subgroups, highlighting potential sources of health inequity.

Essentially, mathematically, to assess overall feature importance across the population, we computed global SHAP values by averaging the absolute contributions of each feature over all individuals:

$$\phi_j = \frac{1}{N}\sum_{i=1}^{N} |\phi_{ij}|$$

Where: ϕ_{ij} denotes the SHAP value of feature j for individual i; N is the total number of individuals; ϕ_j quantifies the average marginal contribution of feature j to the model's predicted log-odds or probability output. These global values provide a ranked summary of which features were most influential in disease risk prediction across the entire dataset.

To identify subgroup-specific risk drivers and potential sources of disparity, we stratified the SHAP values by socio demographic group g (e.g., tribe, religion, income tier). For each feature j, we computed the average SHAP value within group g as:

$$\overline{\phi}_j^{(g)} = \frac{1}{|g|}\sum_{i \in g} \phi_{ij}$$

Where: $|g|$ is the number of individuals in group g; $\overline{\phi}_j^{(g)}$ represents the average contribution of feature j within that subgroup. This group-level aggregation enabled population-level interpretation of risk factors, revealing whether certain social groups were disproportionately influenced by specific features.

Unsupervised population stratification

To uncover latent subpopulations with distinct health profiles, we applied unsupervised learning techniques combining nonlinear dimensionality reduction and clustering. High-dimensional feature representations were embedded into a two-dimensional latent space using Uniform Manifold Approximation and Projection (UMAP), which preserves both local neighborhood relationships and global data structure.

We employed UMAP to project the high-dimensional feature matrix $X \in R^{N \times P}$ into a low-dimensional space that preserves both local neighborhoods and global data structure. Formally, the embedding is given by:

$$Z = \text{UMAP}(X),\ Z \in R^{N \times 2}$$

Where: N is the number of individuals, P is the number of input features, Z is the resulting 2D latent representation for visualization and clustering.

UMAP was selected due to its ability to maintain the manifold geometry of the input space, making it particularly well-suited for identifying meaningful structure in high-dimensional sociodemographic and health-related data.

We then applied KMeans clustering to the UMAP embedding Z to partition individuals into K latent clusters. For each individual i, cluster assignment was defined as:

$$C_i = argmin_k ||Z_i - \mu_k||^2$$

Where: $C_i \in \{1, \ldots, K\}$ is the assigned cluster label for individual i; $Z_i \in R^2$ is the UMAP embedding of individual i; μ_k is the centroid of cluster k. The number of clusters K was determined using silhouette analysis or the elbow method, depending on the specific disease context.

Clustering was performed on the UMAP embeddings using KMeans, with the number of clusters determined via standard cluster validation criteria. The resulting clusters were visualized and annotated with predicted disease risk and sociodemographic attributes, revealing emergent subgroups characterized by distinct combinations of social context and health risk.

Interactive visualization and decision-support tools

To promote accessibility, transparency, and real-world applicability, we developed interactive web-based dashboards for exploring model outputs and population-level patterns. Using the Dash framework, we implemented an interface for interactive inspection of SHAP values, subgroup comparisons, and disease-specific risk factors. A complementary Streamlit application enables real-time disease risk prediction, interactive cluster exploration, and population-level summaries, supporting scenario analysis and targeted intervention planning.

Together, this framework provides a scalable, interpretable, and equity-centered approach to modeling chronic disease risk in diverse populations. By integrating statistical testing, machine learning, explainability, and interactive visualization, the pipeline enables both rigorous scientific inference and actionable public health insights tailored to Nigerian populations.

Code Availability

All analysis code is available at https://github.com/1234-Ariel-code/sociomap/.

Acknowledgement

We thank all members of the MIRA Lab for their dedication, collaborative spirit, and valuable contributions throughout this project. Their collective efforts were instrumental in the development of the analytical framework and the generation of the results presented in this study. We are grateful to the participants and communities whose data made this research possible. We also acknowledge the broader public health, open science, and machine learning communities for developing the tools, frameworks, and resources that informed our methodology. This work represents an important milestone for the MIRA Lab, and we look forward to building on it in future efforts to advance equitable, interpretable, and data-driven health research across African populations.

References

1. Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M. & Daly, M. J. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

2. Knowles, J. W. & Ashley, E. A. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med.* **15**, e1002547 (2018).

3. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

4. World Health Organization. *Noncommunicable Diseases Country Profiles 2021: Nigeria* (WHO, 2021).

Marmot, M. Social determinants of health inequalities. *Lancet* **365**, 1099–1104 (2005).

5. Braveman, P., Cubbin, C., Egerter, S., Williams, D. R. & Pamuk, E. Socioeconomic disparities in health in the United States: What the patterns tell us. *Am. J. Public Health* **100**, S186–S196 (2010).

6. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).

7. Nature Editorial. How to bring more diversity into polygenic risk scores. *Nature* **606**, 9 (2022).

8. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 785–794 (ACM, 2016).

9. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).

10. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *J. Open Source Softw.* **3**, 861 (2018).

11. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G. & Newell, E. W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).

12. Lundberg, S. M., Erion, G., Chen, H. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

13. H3Africa Consortium. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).

14. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).

15. Khoury, M. J., Iademarco, M. F. & Riley, W. T. Precision public health for the era of precision medicine. *Am. J. Prev. Med.* **50**, 398–401 (2016).

16. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**, 866–872 (2018).

17. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).