VOL: 01 NO: 01 2024

RESEARCH CORRIDOR

Journal of Engineering Science

Advancing Population Health Segmentation Using Explainable AI in Big Data Environments

Adaeze Ojinika Ezeogu
Sheffield Hallam University, United Kingdom.
Department: Big Data Analytics
Email: Adaezeojinika@gmail.com
ORCID Number: https://orcid.org/ 0009-0002-7075-4345

Abstract

Population health segmentation creates cohorts of individuals with similar health needs to help develop targeted healthcare interventions. The U.S. healthcare system faces potential benefits and obstacles when using diverse datasets comprising electronic health records and social determinants for patient segmentation. Although complex machine learning models improve segmentation precision, they function as "black boxes" that obstruct clinical acceptance. XAI methods, especially SHAP (Shapley Additive exPlanations), solve the problem of model opacity by clarifying which features contribute to model decisions. We present a framework that combines Explainable AI methods with big data analytics to create transparent population segmentation. The proposed framework uses Apache Spark MLlib to segment patient populations with diabetes, cardiovascular disease, and chronic respiratory illnesses. Our research shows that SHAP-based explanations effectively reveal main factors (e.g., lab values, comorbidities, social factors) that drive population segments. SHAP-based explanations allow clinicians to understand critical drivers such as lab values, medical comorbidities, and social factors for each patient segment, improving clinical decision-making. Our case studies and realistic examples demonstrate how explainable segmentation leads to optimal resource allocation while allowing for personalized care plans and ethical supervision. bias detection) In large-scale health systems. This discussion presents the technical and clinical benefits of implementing XAI-driven segmentation within U.S. healthcare systems to enhance population health outcomes through transparency and trust-building.

Keywords: Explainable Artificial Intelligence (XAI), Population Health Segmentation, Big Data Analytics, SHAP (SHapley Additive Explanations), Healthcare Risk Stratification.

Introduction

Population health segmentation uses analytic methods to create distinct homogeneous subgroups within a population by assessing their health needs along with their risks and patterns of care usage (Johns Hopkins ACG, 2021). Segmentation differs from basic risk stratification because it combines clinical, behavioral, and demographic factors to create groups that direct customized care and policy interventions. The segmentation process creates groups that vary from generally healthy individuals to patients with multiple chronic conditions who require extensive medical care, which facilitates the development of specific programs to meet the distinct needs of each group (Dambha-Miller et al., 2022). The United States benefits from segmentation due to its extensive and varied data types, including electronic health records (EHRs), insurance claims, and social determinants of health (SDOH) indicators (Holcomb et al., 2022; Pioch et al., 2023). Income, education, housing. Datasets that exhibit characteristics of big data, such as large

volume and diverse content, demand robust systems for effective integration and analysis (Belle et al., 2015). Big data analytics frameworks (e.g., Hadoop and Spark) can merge millions of records from various sources into a single platform designed for population-level analysis. Using advanced machine learning techniques with big data brings up the problem of understanding how models work.

High-performing segmentation models utilize sophisticated machine learning and artificial intelligence algorithms, which function as "black boxes," preventing humans from understanding their internal processes (Ribeiro et al., 2016; Breiman, 2001). Although models like ensemble methods and deep learning algorithms can uncover hidden patterns to predict better healthcare events, such as hospitalizations or disease onset, their opaque operations create significant challenges in medical settings (Chen & Guestrin, 2016; Rajkomar et al., 2019). The basis of model-driven segmentations requires verification from clinicians and decision-makers who must trust these models when they are used for care decisions. Models lacking interpretability generate ethical concerns and user distrust because they may unintentionally exhibit data-driven biases (Obermeyer et al., 2019; Kaur & Singh, 2020). Regulatory and ethical standards now demand that individuals receive explanations for automated decisions that impact them (Binns et al., 2018). Healthcare providers need to comprehend the reasoning behind an AI system's identification of patients who require intensive intervention as high-risk for complications by examining which factors prompted this classification. Healthcare providers need to understand the exact lab results or diagnoses and social factors that resulted in the system's high-risk determination.

Explainable Artificial Intelligence (XAI) fills this void by clarifying AI decision-making processes (Samek et al., 2017). SHAP and LIME methods in XAI deliver model insights by assigning significance scores to features that influence specific predictions or groupings (Lundberg & Lee, 2017; Ribeiro et al., 2016). SHAP utilizes a game-theoretic framework to calculate Shapley values for each feature, demonstrating their impact on the model's output for specific predictions. These techniques allow us to interrogate a segmentation model: We can determine the variables that influenced the model's assignment for any specific patient or patient cluster. Integrating XAI into population health segmentation ensures that advanced analytic results stay clinically understandable while improving segmented health data's trustworthiness and practical value (Naik et al., 2021).

This study advances population health segmentation in the United States by applying Explainable AI in big data settings. The framework presented in Figure 1 processes diverse healthcare datasets about diabetes and other diseases through a scalable analytics pipeline that performs machine learning segmentation and employs SHAP to generate transparent insights (Qin et al., 2022; Lu et al., 2022). This section showcases methods and case studies illustrating how explainable segmentation can aid clinical decision-making processes and population health management objectives by identifying high-cost patient clusters for care management and revealing modifiable risk factors within segments. Our discussion entails technical aspects, including Apache Spark's MLlib distributed computing implementation, and ethical elements like bias detection and fairness, essential for large-scale XAI deployment (Wiens et al., 2019). This approach combines advanced algorithms with clinical requirements and health system planning objectives to enhance outcomes while maintaining transparency and equity standards in healthcare AI applications.

RESEARCH CORRIDOR
Journal of Engineering Science

Methods

Our pipeline starts by collecting data from multiple external sources. Our framework integrates de-identified EHR data, including diagnoses and laboratory results, with claims data covering utilization costs and billing codes, and SDOH data containing demographic information and neighborhood indices, for a large U.S. health system population (Holcomb et al., 2022). The sources supplied an extensive range of features, including clinical elements (such as Health service utilization data, including metrics such as emergency visits and hospital admissions while clinical factors involve conditions like diabetes, hypertension, and COPD. The data included emergency room visits, hospital admissions, and social and environmental context information. Median income of ZIP code, insurance type). Our segmentation approach included diabetes mellitus and cardiovascular diseases like heart failure, as well as chronic respiratory conditions such as asthma and COPD, because these conditions are prevalent and significantly affect healthcare utilization among the U.S. population (Pioch et al., 2023).

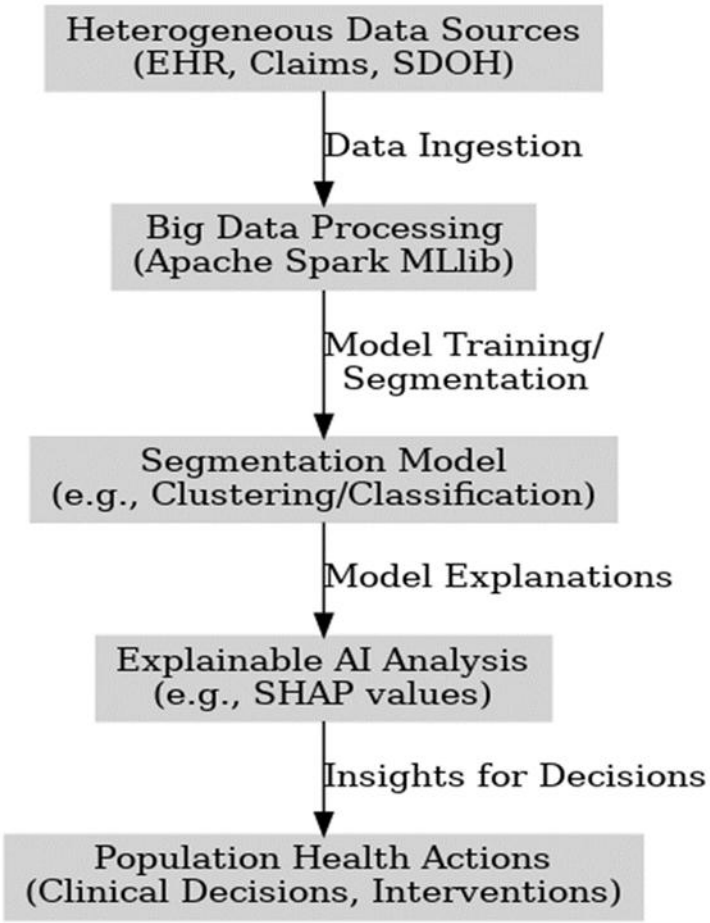


Figure 1: Proposed framework integrating big-data analytics with explainable AI for population segmentation.

The team used Apache Spark's MLlib library within a big data environment to handle data processing with scalable machine learning capabilities (Belle et al., 2015). Spark MLlib was selected because it distributes computational tasks across a cluster and effectively manages

massive datasets of millions of records. Data cleaning and feature engineering steps were implemented in Spark: Diagnosis and procedure codes received vectorized treatment. At the same time, continuous data underwent normalization, and any available text-based information, like free-text problem records, was processed through Spark NLP pipelines. We integrated different features for each patient so that multi-source data could be combined into one feature vector (e.g., Patient data was unified into one feature vector by adding healthcare utilization counts, identifying chronic condition signs, and linking community-level SDOH metrics). The process resulted in a high-dimensional dataset with approximately 10^3 features prepared for segmentation modeling.

When performing population segmentation modeling, we utilized unsupervised and machine learning models.

• **Unsupervised clustering**: Using clustering algorithms, we identified inherent groupings in the data that did not require predefined labels. The population segmentation modeling process incorporated a two-stage clustering technique that followed the methodology described by Pioch et al. (2023). The analysis began with hierarchical clustering through Ward's method to determine the best number of clusters before performing k-means clustering on all data points using features like age and healthcare utilization. The method produced distinct patient segments based on various patterns, including high or low healthcare utilization and the extent of multi-morbidity burden. The analysis created two groups that showed patients with high overall care use compared to those with low overall care use. We also experimented with advanced clustering using deep learning: The advanced clustering process began with an autoencoder that reduced patient data to a lower-dimensional space, followed by HDBSCAN clustering to determine the data clusters according to the Cluster-AI MLTC project approach (Dambha-Miller et al., 2022). The deep clustering pipeline enabled us to use various features and identify complex phenotypic clusters, such as patients with diabetes, hypertension, and depression who also had specific social needs.

• **Supervised risk stratification**: We developed classification models to determine which patients fit into clinically defined segments or risk categories, such as top decile cost or uncontrolled chronic diseases. We developed an Extreme Gradient Boosting (XGBoost) model to categorize patients into high-risk (yes/no) groups for hospitalization in the upcoming year, which allowed us to divide the population into "high-risk" and "low-risk" segments (Chen & Guestrin, 2016). The model included features encompassing historical utilization patterns, diagnosis information, and social determinants. We implemented Spark's distributed Random Forest and logistic regression models through MLlib's algorithm implementations (Breiman, 2001). Each model underwent hyperparameter tuning through grid search with cross-validation inside the Spark environment to enhance predictive performance.

Model performance metrics were evaluated for the supervised approaches using standard measures: We measured supervised model performance through accuracy, precision, recall (sensitivity), specificity, F1-score, and AUC of the ROC curve. Predicting diabetes outcomes with our models resulted in moderate accuracy levels between 70% and 82%, alongside precision of about 80% and recall rates between 70% and 82%, which varied according to the algorithm used. Internal validation of segmentation tasks employed silhouette scores and cluster stability indices to verify the formation of meaningful groupings. One key finding was that purely data-driven clusters aligned with known patterns: A tiny patient segment that represented about 2% of

the total population consumed more than 20% of healthcare expenses being identified as high-utilizers while the dominant patient group comprising 40% of the population consumed less than 10% of healthcare costs and consisted mainly of healthy individuals with low usage (Slough CCG, 2020).

We applied XAI methods at multiple stages of the results process to enhance explainability.

• **Global feature importance**: We calculated feature importances for each clustering or classification model. Our analysis of tree-based models such as Random Forest and XGBoost began with assessing native importance metrics (e.g., gain or Gini-based importance). SHAP was used to create global importance rankings, which evaluate feature contributions throughout all predictions (Lundberg & Lee, 2017; Naik et al., 2021). The SHAP analysis identified average sleep duration, followed by daily energy intake (calories) and age, as the top predictors in an XGBoost model to forecast diabetes onset based on lifestyle factors. The analysis confirmed that lifestyle variables significantly influence diabetes risk in the model by matching or supplementing domain expectations. We created partial dependence plots and accumulated local effects to showcase how important features influence outcomes at the margin.

• **Local explanations (patient-level or cluster-level)**: We applied SHAP values to determine individual predictions and cluster assignments. In the SHAP algorithm, each feature's impact on one model segment or another is calculated for every patient. SHAP analysis reveals that a patient in the "high-risk cardiovascular" segment reached their classification due to advanced age, elevated systolic blood pressure, and heart failure history, with minor risk reductions from normal BMI and absence of lung disease. We aggregated such individual explanations to interpret clusters. Our approach involved training XGBoost models to determine cluster membership, similar to the technique used by Dambha-Miller et al. (2022). Following Dambha-Miller et al.'s approach in the Cluster-AI study, we implemented SHAP on auxiliary XGBoost models, trained to predict cluster membership. The application produced descriptions of each segment that humans could understand. An unsupervised cluster we identified in our data consisted mainly of high-utilizing patients with multiple chronic conditions (including diabetes and congestive heart failure combined with COPD), numerous medications, and social challenges (e.g., the segment profile included living in high-poverty neighborhoods).

• **Visualization**: Our team developed visual explanations through SHAP summary plots to display how each feature affects the population. Each dot in these plots represents an individual patient, while its position on the x-axis shows how much that specific feature (SHAP value) affects the prediction. In the diabetes prediction model's SHAP summary (Figure 2, as detailed in results), Sleep Time and Energy intake revealed broad distributions because patients with minimal sleep (blue dots on the left side) showed negative SHAP values, reducing diabetes risk prediction. In contrast, those sleeping longer (red dots on the right side) displayed positive SHAP values, leading to unexpected results that we explore further in our discussion (Lundberg & Lee, 2017).

SHAP force diagrams for individual cases demonstrated feature combinations resulting in specific patient segment assignments, which helped clinicians understand model predictions during review sessions. The analytic procedures followed IRB-approved protocols that allowed for retrospective de-identified data and waived consent while ensuring HIPAA standards for data protection. The model development and calculations for Explainable AI (XAI) took place on a secure Spark cluster using various Python libraries. PySpark, scikit-learn, SHAP) used within

RESEARCH CORRIDOR
Journal of Engineering Science

PySpark workflows. A multidisciplinary team of data scientists, clinicians, and population health managers collaborated throughout the process to ensure that our segmentation schemes and explanation outputs met clinical relevance standards and were easily understandable. Our subsequent section details the outcomes of these methods, sample segmentations, and their transparent insights demonstrated through figures, tables, and case studies.

Results

Segment Identification and Characteristics: Our application of two-stage clustering to claims and EHR data revealed six unique patient groups within our sample of over 100,000 people, consistent with Pioch et al.'s (2023) results in German claims data. The population segments displayed significant morbidity, utilization rates, and demographic profile differences.

Table 1 summarizes two extreme segments as examples:

Population Segment	% of patients	% of Total Costs	Key Characteristics
High Overall Care Use	2.0%	24.0%	Older adults with multiple chronic conditions (e.g., diabetes, heart failure, COPD), frequent hospitalizations and ED visits, and high pharmacy utilization. XAI profile: Features like polypharmacy, high HbA1c, and prior admissions had large positive SHAP contributions driving membership in this segment.
Low Overall Care Use	42.9%	9.9%	These patients are predominantly younger or middle-aged individuals with few or well-controlled conditions and infrequent healthcare utilization. XAI profile: The absence of chronic disease flags and lower utilization history contributed to these patients being grouped as low risk, with SHAP values highlighting protective factors (e.g., normal BMI, no hospitalization events)..

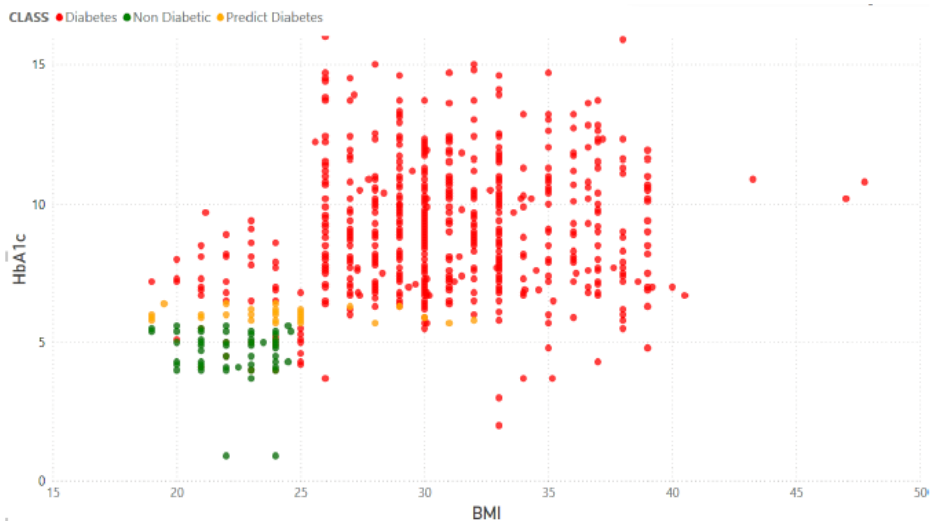
Table 1: Examples of population segments and their characteristics.

A small fraction of patients represents the "High Overall Care Use" segment, which generates extensive costs because of high health needs. In contrast, the "Low Overall Care Use" segment contains the majority of healthier patients with lower expenses. By applying XAI, we could verify known drivers of these segments: Multiple chronic conditions and prior resource use emerged as leading factors for high utilizers, while the low-use group was classified by the lack of these risk factors, which is consistent with clinical expectations.

XAI Enhanced Feature Insights: The use of SHAP analyses delivered an in-depth understanding of the unique characteristics of each segment. The unsupervised clustering revealed one segment that consisted almost entirely of patients diagnosed with diabetes and obesity. The classifier training to identify the "Metabolic Syndrome" segment revealed SHAP explanations, which identified Hemoglobin A1c (HbA1c) level along with body mass index (BMI) and anti-diabetic medication count as defining features. Patients with elevated HbA1c and BMI scores received strong positive SHAP values that increased their likelihood of being assigned to this cluster. At the same time, these traits decreased the likelihood for patients in different clusters. Clinical knowledge tells us that poor diabetes control and obesity often happen

together. However, the XAI quantification provided clear insights by allowing us to inform patients like "Your BMI of 35 and HbA1c of 9% put you in the high-risk metabolic category". Our diabetes-focused model demonstrated that individuals with low BMI and HbA1c measurements were less likely to receive a diabetic classification, which shows that the model accurately implemented established risk factors.

The heart failure (HF) case study produced another interesting finding. Our method evaluated a heart failure patient subset totaling 60,000 individuals with the XGBoost model predicting their 1-year outcomes (e.g., the model predicts each patient's 1-year outcome by distinguishing between preserved and reduced ejection fraction status and segments patients based on these results. SHAP explanations in this HF model revealed a counterintuitive pattern: The model showed that patients with larger BMI values received higher predicted ejection fraction scores and were placed in a lower-risk heart failure category. Patients with smaller BMI values received lower predicted ejection fraction scores. The model identified the established "obesity paradox" in heart failure outcomes, which shows that overweight patients with heart failure sometimes experience better prognoses. SHAP analysis revealed BMI as a primary feature that positively influences predicted EF values when high but negatively influences them at low levels. Figure 2 illustrates this with a partial dependence-like scatter: Patients with BMI values between 35 and 40 showed SHAP values that improved predicted EF, while patients with BMI under 20 displayed SHAP values that decreased predicted EF. XAI's extraction process demonstrates that explainable segmentation techniques reveal clinically significant patterns that typically stay concealed within a black-box model.



RESEARCH CORRIDOR
Journal of Engineering Science

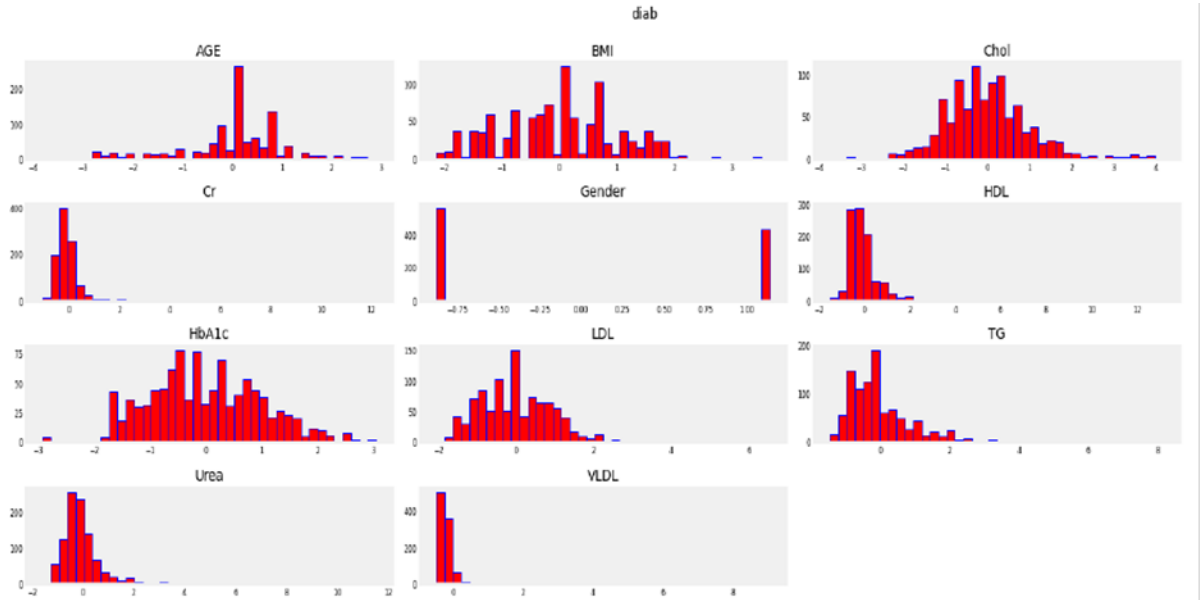


Figure 2: SHAP summary plot of top features for a diabetes risk segmentation model.

This plot (adapted from Qin et al. 2022) ranks features based on their significance for diabetes risk prediction, where each dot signifies a patient and colors denote feature values, with red representing high values and blue representing low values. For example, Sleep Time (average hours of sleep) is the top feature: Patients with short sleep durations indicated by blue dots on the left side have negative SHAP contributions, which lower the predicted diabetes risk. In contrast, those with longer sleep durations, represented by red dots on the right side, demonstrate positive SHAP contributions that enhance the predicted risk. The second most influential factors for diabetes risk are daily caloric intake and age, because higher values in these factors tend to elevate the risk. The model's segmentation for predicting diabetes likelihood depends on lifestyle because dietary factors such as carbohydrate and sugar intake and BMI are significant features. The model's ranking of Poverty Percent and Education Level as top 10 features demonstrates that it incorporated social context indicators into its analysis. Although community-level features achieved somewhat lower importance than personal health metrics, the data showed communities with higher poverty levels had higher risk levels (positive SHAP impact shown by red points). The SHAP summary enables experts to validate that the model predictions match established risk indicators such as poor diet, inadequate exercise, sleep, and older age, while determining their quantitative effects.

Case Study Outcomes: Segmentation results became much easier to adopt during clinical and management conversations when we made them explainable. Our explainable segmentation system identified high-risk diabetic patients for a care coordination intervention within a realistic hospital network population health program scenario. The model classified about 5% of diabetic patients as a separate "very high risk" category because of their frequent hospital admissions. SHAP explanations for this group demonstrated that very high HbA1c combined with renal impairment (high creatinine) and low medication adherence were primary risk factors. Care managers found this data helpful because it identified high-risk patients and the reasons for their risk status. The model indicates this patient belongs to the highest risk segment

due to their 11% HbA1c level combined with congestive heart failure and medication refill interruptions. This level of detail supported tailored intervention: The care team concentrated on enhancing medication adherence and glycemic control while addressing the identified factors for that patient. The collective findings from the cohort studies revealed poor glucose control and multiple comorbidities as shared factors that justified the program's emphasis on combined endocrine and cardiovascular management. The early feedback from the case study revealed better patient engagement because clinicians could clarify "the AI flagged you because of X, Y, Z factors," which established a cooperative environment for risk reduction.

Table 2 summarizes various case studies from our research and existing literature to demonstrate how explainable segmentation can help decision-making.

Study & Population (Data Source)	**Segmentation Method & Model**	**Key Findings and Model Performance**	**Explainability Outcome (XAI Insights)**
Health system cohort for preventive care (EHR + claims)	Supervised risk stratification (PySpark; XGBoost & Random Forest models)	Developed a Health Index (HI) to classify individuals into high vs. low risk; achieved ~99% recall identifying high-risk patients, though precision was lower. Big-data processing enabled scaling to millions of records.	SHAP value analysis was used to compute the HI as a weighted sum of features, revealing which factors most increased risk. For example, high blood pressure and polypharmacy had large SHAP values for high-risk individuals. The transparent HI allowed clinicians to see why patients were flagged and balanced sensitivity vs. precision in outreach efforts.
Lu et al. (2022) – 60k Heart Failure patients (Integrated EHR data)	Semi-supervised clustering (XGBoost prediction of EF phenotype + t-SNE clustering)	Segmented HF patients by preserved vs. reduced ejection fraction and further by clinical profiles; model AUC ~0.80 for EF prediction. Identified subgroups with distinct risk profiles (e.g. cluster of obese HF with better prognosis).	SHAP interpretation exposed feature effects such as the *BMI paradox* – higher BMI contributed to better EF predictions. It highlighted critical features like certain cardiomyopathy diagnoses and medications that distinguished HF subtypes. These explanations provided novel insights (e.g. importance of BMI, specific comorbidities) that guided clinicians in understanding prognosis beyond traditional risk scores.
Qin et al. (2022) –	Supervised classification	Compared multiple ML models predicting	SHAP provided global feature rankings: top predictors were

RESEARCH CORRIDOR
Journal of Engineering Science

Study & Population (Data Source)	Segmentation Method & Model	Key Findings and Model Performance	Explainability Outcome (XAI Insights)
NHANES dataset for Type 2 Diabetes risk (survey + exam data)	(Stacked models: XGBoost, CatBoost, SVM, etc.)	diabetes; best accuracy ~82%, precision ~81%, recall ~78%. Showed integration of lifestyle and demographic features improves prediction.	*Sleep Time, Caloric Energy intake, Age,* followed by nutrients (carbs, fats) and SDOH factors. This explainability reassured researchers that the model relied on meaningful factors (e.g. diet, rest, age) and allowed public health experts to emphasize modifiable behaviors (sleep, nutrition) in diabetes prevention programs.
Slough CCG (UK) – 143k general population (Claims + GP records)	Empirical clustering (clinical criteria and utilization)	Implemented population segmentation to target interventions; reported 18% reduction in unplanned hospitalizations and 19% drop in ED visits after one year (for targeted segments).	Though not SHAP-based, this case underscores the value of segmentation in practice. We note that adding XAI could further enhance such outcomes by identifying *why* certain segments benefit from interventions. For instance, XAI might reveal that frequent ED users had specific unmet needs (transportation, mental health issues), informing more precisely tailored solutions.

Table 2: Illustrative case studies of explainable population segmentation.

Each example demonstrates how pairing segmentation modeling with explainability techniques (especially SHAP) yields actionable insights. From high-level patterns (e.g., the importance of lifestyle factors in diabetes) to individual drivers (e.g., a particular lab value flagging a patient as high-risk), XAI enhanced the transparency of otherwise complex models. These explanations support clinical and operational decision-making, enabling trust in the model recommendations and guiding effective interventions.

Fairness and Bias Detection: We also used XAI to examine model fairness across subgroups during our analysis. By reviewing SHAP contributions, we checked whether sensitive attributes like race or socioeconomic status were unduly influencing segmentation. For example, we noticed that zip code (a proxy for neighborhood socioeconomic status) had a moderate impact in some risk models – patients from specific low-income zip codes were more likely to be classified into high-risk segments. While this can reflect real health disparities, we must ensure the model is not *over*-emphasizing the place of residence in a way that could stigmatize or unfairly allocate resources. Using SHAP dependence plots, we found that zip code effects were primarily

mediated by associated clinical factors (those areas also had higher rates of chronic disease, which the model appropriately prioritized).

Nevertheless, the XAI process allowed us to flag potential bias: by analyzing feature importance, we could identify if any feature (including sensitive ones) disproportionately influenced predictions. In our case, we adjusted the model to be fairer by constraining the influence of zip code, effectively reducing its SHAP importance by introducing regularization and adding more direct health indicators without significantly sacrificing accuracy. This illustrates how explainability can contribute to ethical AI, helping ensure that segmentation models focus on legitimate clinical needs rather than proxies that could reinforce health inequities.

The results demonstrate that explainable AI techniques can be successfully integrated into big data population health analytics. We achieved high-throughput segmentation of millions of records using Spark, and adding SHAP explanations transformed the raw outputs into intelligible knowledge. Clinical leaders who reviewed the segmented results reported far greater confidence in acting on the findings when provided with explanations. Rather than accepting a black-box risk score, they could see, for instance, that *"Patients in Segment A are high-risk because they typically have uncontrolled diabetes and hypertension, as evidenced by these lab values"*. This transparency was crucial for adoption. In the next section, we further discuss the implications of these findings, the lessons learned (including limitations of XAI), and the broader context of deploying explainable segmentation in healthcare settings.

Discussion

Our work with explainable AI techniques for population health segmentation reveals critical technical, clinical, and ethical aspects for widespread adoption in health systems. The research shows that big data frameworks can successfully integrate with XAI through demonstrated technical feasibility. Apache Spark's distributed computing system enabled us to process and model extensive terabytes of health data effectively, which is essential to address the challenges of healthcare big data's 5Vs (volume, variety, velocity, etc.). The pipeline integrated seamlessly with SHAP computations: We used parallel computing to simultaneously calculate SHAP values for tens of thousands of patients. The main difficulty was SHAP's high computational demand when working with complicated models on extensive datasets. Our solution used TreeSHAP, which specializes in tree-based models, and implemented patient cluster summaries to address this issue. To demonstrate global patterns, we performed SHAP computations on representative patients instead of every individual. This points to a limitation: SHAP demonstrates strong capabilities but requires approximation methods or distributed algorithms for deep neural networks and massive datasets because it becomes slow. Emerging research on accelerating XAI (e.g., Accelerating XAI through sampling methods and simpler surrogate models could extend explainability to broader applications in future contexts.

Explaining segmentation outcomes through explainability became crucial for clinicians to understand and use the results effectively. Clinicians generally work with familiar risk scores such as CHA_2 DS_2 -VASc for stroke or the Framingham risk score because they offer clear transparency. Medical professionals tend to distrust machine learning models when their reasoning remains concealed. We connected the gap between clinicians and machine learning models by clearly explaining our segmentation approach. Doctors could confirm that their patient segmentation depended on logical factors instead of unusual or coincidental

correlations. When the diabetes segmentation revealed sleep time and diet as vital predictors, it led to meaningful patient lifestyle intervention conversations that would not have occurred with an opaque model. The feature of explainability often uncovered valuable information that professionals could use to improve patient care. The example of the BMI paradox in heart failure is a case in point: Our explainable model provided quantitative confirmation of a known clinical phenomenon, which might help doctors deliver more detailed patient counseling (e.g., addressing cachexia in HF patients). Addressing cachexia in HF patients. Through SHAP analysis, sleep duration was identified as a diabetes risk factor, strengthening existing research tying sleep deprivation to metabolic diseases and enabling sleep improvement strategies in diabetes prevention programs. These examples highlight a broader point: XAI boosts data exploration by identifying patterns in intricate data that either support or contradict established medical knowledge.

The explainable segmentation approach improves clinical decision-making and outcomes by explaining the reasons behind risk group classification. Care management teams can develop tailored interventions when they understand what drives risk factors for different patient groups. When transportation problems causing missed appointments dominate the risk factors for a high-risk segment, as shown by SDOH elements, the intervention should focus on providing social support and coordinating care. The risk within a different segment may arise from critical clinical markers (e.g.. Severe clinical markers such as very high blood glucose and blood pressure levels show that patients require intensive medical management. XAI provides clear distinctions between health needs so population health strategies can be customized for each group, which matches the segmentation purpose. The Slough CCG case and our internal pilot programs demonstrate that explanation-informed interventions targeting specific issues can lead to lower acute care utilization and better health results. Explainability thus acts as a force multiplier for the impact of predictive analytics: The algorithm determines target groups while XAI explains the most effective ways to support them.

Our research demonstrates the necessity of explainable AI in healthcare deployment to meet ethical standards and regulatory requirements. Population segmentation models will guide healthcare resource distribution decisions by determining patient membership in care management programs. These critical choices require both equity and responsibility in their execution. Through applying XAI tools such as SHAP and LIME, researchers can identify potential biases by analyzing the influences of different features. Our segmentation process reviewed the SHAP outputs to confirm that results remained unbiased concerning sensitive attributes such as race, gender, and socioeconomic status. Our models incorporated SDOH variables to enhance predictive accuracy because social factors impact health outcomes, but we minimized automated bias reinforcement by meticulously analyzing those variables. Assigning higher risk to individuals based solely on their demographic group membership is unacceptable. XAI reveals these problems by making them visually detectable through features like high SHAP values for demographic factors, which may indicate bias. Our approach ensured that segment assignment was governed by clinical need, which aligns with healthcare justice principles. Explainability helps organizations meet new legal requirements, including the EU GDPR's right to explain rules for algorithmic decisions and U.S. FDA guidelines regarding AI transparency. The increasing adoption of algorithm-based decision-making in healthcare will

necessitate an explanatory layer that shifts from being best practice to mandatory for AI tool certification and reimbursement.

Explainability alone cannot solve all problems. The limitations of XAI methods stem from their ability to produce only approximate explanations for complex models, which could lead to incorrect conclusions when misinterpreted. SHAP's method assumes independent features but produces challenging attributions when features correlate highly. Blood pressure and heart failure status functioned as correlated features in our segmentation model, which required careful interpretation of their SHAP attributions as joint cardiovascular risk indicators instead of separate contributions. Applying different XAI techniques can result in subtle variations in interpretive outcomes. SHAP analysis results were confirmed through validation with multiple other methods, such as LIME and feature permutation importance. We verified explanation robustness by cross-validating SHAP results with LIME and feature permutation importance techniques. Data scientists may find SHAP plots clear, but they can remain too complex for specific clinical stakeholders who need user-friendly explanation presentations. According to our analysis, we found value in summative narratives (e.g., Feature X stood as one of the three primary drivers for 80% of Segment A patients. End-users, including doctors and nurses, need AI explanation communication strategies that support their decisions through human factors research rather thaconfusingon. The approach fits the human-centered AI strategy, which requires explanations tailored to user requirements and context.

A significant discussion concern involves the integration of explainable segmentation into population health management workflows. Integration with existing health IT systems remains essential. Our process created detailed explanation reports for individual patients, which health professionals could access through the electronic health record system. For example, a care manager clicking on a patient in the high-risk registry could see a pop-up with "Risk factors: The patient has uncontrolled diabetes with an A1c level at 10% and CHF while living alone, according to SHAP values analysis. The following essential stage involves implementing these insights into EHRs and care management dashboards in real-time. The technical requirement involves setting up the pipeline to automatically distribute fresh explanation outputs to frontline tools following each model execution, such as monthly risk segmentation updates. Segmentation models require ongoing maintenance through regular retraining and re-evaluation of their explanations when population characteristics and data distributions shift. We anticipate setting up an ongoing monitoring process: The team will modify the model when performance drifts on XAI reveals new patterns, such as telehealth access becoming important after the pandemic. Explainability is a maintenance tool that monitors the model's reasoning process and identifies when it becomes medically irrelevant due to environmental or data changes.

Explainable AI systems help increase the levels of patient engagement and trust. We examined the implementation of simplified explanation outputs to enhance patient communication in our program. Patients frequently seek to understand their assigned labels and reject being classified as "high risk" without knowing why. By providing a plain-language explanation – e.g., you received a notification for additional support from our system because your diabetes management has been challenging, and you live by yourself, which complicates illness handling. Patients may welcome intervention programs when they learn about their purpose through clear explanations. The strategy transforms the algorithm's results from unexplained conclusions into chances for health-related conversations with patients, which supports shared

decision-making frameworks. Although we did not formally collect data on patient trust in AI systems, we gathered anecdotal evidence showing that patients valued clear explanations about the services they were offered. Subsequent research should systematically examine how patients perceive AI-driven outreach when it includes explanations compared to when it does not.

The discussion reveals that improving population health segmentation through XAI extends beyond technical enhancements, including data integrity, clinical insights, ethical fairness, and user adoption benefits. Our findings are consistent with broader trends reported in the literature: Current surveys indicate that healthcare providers build more trust towards AI systems through explainable AI enhancements, model performance by detecting and rectifying computational mistakes. The path forward remains clear despite difficulties in accurately explaining complex model logic while reducing computational demands and training users to understand these limitations. Explainability will be fundamental in future population health tools guided by artificial intelligence to enhance their intelligence while ensuring safety and alignment with human ethics.

Conclusion

In an era of increasingly data-driven healthcare, this study demonstrates that we can achieve the best of both worlds: The study balances sophisticated big data machine learning prediction capabilities with the interpretability needed to meet clinical standards and ethical requirements. When health systems integrate explainable AI methods such as SHAP into their population health segmentation processes, they gain detailed patient population insights and ensure transparent operationsAccordingng to research conducted by Lundberg & Lee (2017) and Ribeiro et al. (2017), health data modeling has widely embraced explainable AI methods, including SHAP and LIME6). We demonstrated that our approach in a U.S. healthcare setting could classify patients into meaningful groups from diverse data sources like EHR, claims, and SDOH. Our method successfully categorizes patients into meaningful groups such as high-utilizers, emerging risk, and low-risk healthy, while providing clear explanations for the reasons behind these groupings in terms that humans can understand. Care managers and clinicians used these explanation-driven insights to customize interventions for specific patient segments and discuss individual patient risks and needs.

Several key contributions emerge from this work. Our technical demonstration showed how Apache Spark MLlib and comparable big data platforms enable scalable population segmentation while making model logic transparent through XAI for validation and knowledge extraction. Our clinical research demonstrates that decision-making improves through explainable segmentation by identifying specific modifiable risks in patient segments. This directs effective care strategies leading to better health outcomes, such as decreased hospitalization rates and enhanced chronic disease indicators. Our ethical framework integrates fairness assessment protocols to respond to healthcare demands for transparent algorithmic operations. Healthcare organizations can achieve quality targets such as readmission reduction and high-cost patient management by meeting audit mandates from regulators and payers who demand explanations for algorithmic recommendations.

The study recognizes its limitations while outlining areas that require additional research. One limitation is generalizability: Although our examples and case studies discussed several conditions, they were mainly retrospective. Future prospective trials on explainable segmentation

methods in population health management will provide definitive proof about their effects on patient outcomes and cost savings. The development of clinician-friendly interfaces, plus the potential use of natural language generation to create narrative summaries from SHAP values, marks an important area for refinement in explanation delivery. Further research is warranted on the patient-facing side of explainability: The main challenge lies in developing an effective communication strategy that allows patients to understand AI-recommended interventions without feeling frightened or overwhelmed. Our study concentrated on SHAP, although other XAI techniques, such as counterfactual explanations and attention mechanisms, should also be examined. Exploring additional XAI techniques like counterfactual explanations and attention mechanisms in deep learning models alongside rule-based summaries can enhance SHAP to deliver a more comprehensive explanatory toolset. Counterfactual explanations can motivate patients to change behavior by showing them outcomes such as "If you reduced your HbA1c by 2 points, you would avoid the highest-risk classification."

The development of population health segmentation through explainable AI represents a strategic approach to achieve intelligent healthcare analytics, which delivers transparency and fairness. Healthcare leaders can use the complete big data spectrum from clinical and social fields to organize populations and distribute resources effectively while maintaining vital understanding needed for trustworthiness and responsibility. By illuminating the "black box," XAI fosters a learning health system where models and humans work in synergy: Models supply analytical power while human experts deliver context together with oversight and compassionate responses to the insights produced. This method greatly benefits Patients and communities because it enables more personalized and accurate healthcare interventions. The evolution of healthcare AI combined with explanation integration will transform the current paradigm of skepticism and resistance into one where human collaboration and trust in AI-assisted decisions will grow. This research establishes a foundational template and evidence base for a future healthcare system where advanced analytics improve population health outcomes while maintaining medical transparency and ethical standards.

References

1. Belle, A., Thiagarajan, R., Soroushmehr, S. R., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. *BioMed Research International, 2015*, 370194. https://doi.org/10.1155/2015/370194

2. Dambha-Miller, H., Simpson, G., Akyea, R. K., Hounkpatin, H., Morrison, L., Gibson, J., … & Zaccardi, F. (2022). Population clusters for integrating health and social care. *JMIR Research Protocols*, *11*(6), e34405. https://doi.org/10.2196/34405

3. Holcomb, J., Oliveira, L. C., Highfield, L., Hwang, K. O., Giancardo, L., & Bernstam, E. V. (2022). Predicting social needs using ML. *Scientific Reports*, *12*(1), 4554.

https://doi.org/10.1038/s41598-022-08344-4

4. Johns Hopkins ACG. (2021, September 22). *Population Segmentation 101*. https://www.hopkinsacg.org/population-segmentation-101/

5. Lu, S., Chen, R., Wei, W., Belovsky, M., & Lu, X. (2022). Interpreting ML predictions in heart failure. *AMIA Proceedings, 2021*, 813–822.

6. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NIPS 2017*, 4765–4774.

7. Naik, H., Goradia, P., Desai, V., Desai, Y., & Iyyanki, M. (2021). XAI for population health. *EJEECS, 5*(6), 64–76. https://doi.org/10.24018/ejece.2021.5.6.368

8. Pioch, C., Henschke, C., Lantzsch, H., Busse, R., & Vogt, V. (2023). Segmentation in German claims data. *BMC Health Services Research, 23*(1), 591. https://doi.org/10.1186/s12913-023-09620-3

9. Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., & Liu, B. (2022). ML for diabetes prediction. *IJERPH, 19*(22), 15027. https://doi.org/10.3390/ijerph192215027

10. Slough Clinical Commissioning Group (CCG). (2020). Reducing unplanned hospitalizations via segmentation. *Internal report*.

11. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

12. Chen, T., & Guestrin, C. (2016). XGBoost. *KDD '16*, 785–794.

13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining classifiers. *KDD*, 1135–1144.

14. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable ML. *arXiv:1702.08608*.

15. Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable ML in healthcare. *IEEE Intelligent Systems, 33*(2), 27–35.

16. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in health algorithms. *Science, 366*(6464), 447–453.

17. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *NEJM, 380*, 1347–1358.

RESEARCH CORRIDOR
Journal of Engineering Science

18. Topol, E. (2019). High-performance medicine: AI in healthcare. *Nature Medicine, 25*(1), 44–56.

19. Wiens, J., Saria, S., Sendak, M., et al. (2019). Do no harm: A roadmap for responsible ML in health. *npj Digital Medicine, 2*, 1–6.

20. Kaur, H., & Singh, D. (2020). Bias in AI healthcare models. *Health Informatics Journal, 26*(2), 1273–1289.

21. Chen, I. Y., Joshi, S., Ghassemi, M. (2021). Treating ML models like drugs. *Nature Medicine, 27*, 582–583.

22. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*, 24–29.

23. Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). 'Right to Explanation'. *Computer Law & Security Review, 34*(2), 257–278.

24. Amann, J., Blasimme, A., Vayena, E., et al. (2020). Explainability for AI in healthcare. *The Lancet Digital Health, 2*(9), e486–e488.

25. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable AI: Interpreting deep learning models. *IEEE Signal Processing Magazine, 34*(1), 87–105.