The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 5, No. 1. Online ISSN: 2831-4794

Assessing HIV Mortality Trends in Relation to Prevalence and Diagnosis Rates: A Unified Approach with Machine Learning and Optimal Design

M. Chowdhury[1], S. Balasubramaniam[1], S. Mandal[2*]
[1]Department of Statistics, University of Georgia, Athens, GA 30602, USA
[2]Department of Statistics, University of Manitoba, Winnipeg, MB R3T 5R4, Canada
*Corresponding author. E-mail: saumen.mandal@umanitoba.ca

ABSTRACT

HIV/AIDS remains a significant public health challenge in the world, despite advancements in treatment and prevention over the past few decades. Owing to this fact, we study the relationship between HIV mortality trends, prevalence, and diagnosis rates by employing machine learning methods and optimal design theory. Using advanced machine learning models such as Multiple Linear Regression (MLR), Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR), we predicted HIV-related death rates. The analysis incorporates I-optimal design, a powerful methodology that minimizes prediction variance to enhance model optimization and reliability. Data preprocessing ensured high-quality inputs by addressing missing values, standardizing variables, and handling outliers. The findings reveal that SVR outperformed other models with the lowest mean squared error and the highest R^2. Moreover, integrating I-optimal design improved linear model performance. These results highlight the importance of aligning data design methodologies with model complexity to inform public health interventions. The study underscores the value of optimal design and machine learning in guiding evidence-based resource allocation and improving health outcomes for HIV-affected populations.

Keywords: Human immunodeficiency virus; machine learning; optimal design; regression; random forest; gradient boosting

1. INTRODUCTION

The human immunodeficiency virus is the cause of HIV infection. If HIV is left untreated, it may cause acquired immunodeficiency syndrome (AIDS), the most advanced stage of HIV infection. It continues to spread and affect the lives of millions of people across the globe. First recognized in the early 1980s, the HIV/AIDS epidemic has profoundly impacted communities across the nation, leading to considerable medical, social, and economic implications. As of recent data, an estimated 1.2 million people in the United States are living with HIV, with thousands of new infections reported annually. While progress has been made in reducing transmission rates and improving access to life-saving treatments, disparities persist, particularly among marginalized populations. Understanding the current landscape of HIV/AIDS in the United States involves examining evolving prevention strategies, healthcare interventions, and ongoing efforts to achieve broader health equity for all affected individuals.

Over the time, the Centers for Disease Control and Prevention (CDC) took several initiatives to prevent and control the HIV prevalence rate in the United States using several measures such as increasing HIV testing [1], expanding the use of proven behavioral interventions in prevention programs for high risk populations [2], HIV counseling [3] and etc. According to [3], while numerous interventions have been integrated into prevention programs nationwide, their effectiveness hinges on significantly expanding their reach to expedite progress.

Alarmingly, an estimated quarter of individuals living with HIV are unaware of their status. Moreover, within a recent one-year span, only approximately 15% of men who have sex with men (MSM) engaged in individual-level interventions, and merely 8% participated in group-level interventions—both recognized as highly effective behavioral strategies.

There are several studies concentrated on estimation of HIV Incidence in the United States [4], HIV/AIDS disease burden levels over the globe [5], few studies have reported the global burden and risk factors due to HIV/AIDS. In this study, we use CDC HIV death, prevalence and diagnosis rate data from 2010-2021 in all 50 states to analyze and understand the impact of HIV Prevalence and Diagnosis rate on mortality rates [1].

Combining machine learning methods and optimal design theory, we investigate the relationship between HIV mortality trends, prevalence, and diagnosis rates. We predict HIV-related death rates using advanced machine learning models, including SVR, MLR, RF, GBM, XGBoost, and GBM. I-optimal design, a powerful method that reduces prediction variance to improve model optimization and reliability, is incorporated into the analysis. Analyzing these impacts provides a comprehensive view of the HIV epidemic, facilitating evidence-based decision-making such as improving treatment guidelines to improve health outcomes for affected populations.

The paper is organized as follows: In Section 2, we describe the methodology and data preparation procedures employed in our study. Section 3 details the machine learning models used to predict HIV-related death rates, along with the integration of I-optimal design to enhance model accuracy. Section 4 presents the results of the analysis, highlighting the performance of the model and the improvements achieved by incorporating optimal design principles. Finally, Section 5 concludes with a discussion of the implications of these findings for public health interventions and resource allocation, and offers recommendations for future research and policy initiatives based on the insights derived from this study.

2. METHODOLOGY

The analysis aims to assess the impact of HIV prevalence and diagnosis rates on mortality rates across various U.S. states. We used multiple machine learning models, including Multiple Linear Regression (MLR), Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR), to predict the rate of HIV-related deaths per 100,000 people. The models were evaluated based on their mean squared error (MSE) and R-squared (R^2) values on a holdout test set.

We incorporated I-optimality in our analysis which enhances the study by focusing on optimizing prediction accuracy across the design space. Specifically, I-optimality emphasizes selecting a subset of data points that minimize the average prediction variance for the response variable, in our case, HIV-related death rates. The I-optimality criterion ensures a more efficient and informative design, across a wide range of the predictor variables. By integrating I-optimality, the study can provide deeper insights into the relationship between HIV prevalence, diagnosis rates, and mortality, ultimately leading to more robust and actionable conclusions. We used this criterion to optimize the selection of data points, ensuring minimal average prediction variance across the study region.

2.1 I-optimality and its implementation

As one of our main objectives is to have good prediction for the model, the I-optimality will play an important role in this present work. I-optimality minimizes the average variance of the predicted responses over the design space (experimental region). Other criteria, such as D-optimality, focus on maximizing the determinant of the information matrix (which influences parameter estimation), whereas I-optimality focuses on minimizing prediction uncertainty across the design space. When the objective is to provide predictions that are reliable in a variety of contexts rather than just fitting a model to actual data, I-optimality is particularly pertinent. Every

model has some uncertainty. An I-optimal design mathematically calculates the total amount of uncertainty across the entire design space. It then chooses data points that minimize the average amount of uncertainty. Its goal is to make the average prediction as reliable as possible, rather than focusing on just the best or worst case. It prioritizes the quality of the model's output (the prediction). Thus, the main purpose is to ensure that the predictions are reliable and accurate across a wide range of the predictors.

Consider the problem of selecting an optimal design to furnish information on models of the type

$$y \sim \pi(y|\boldsymbol{x}, \boldsymbol{\theta}, \sigma) \tag{1}$$

where y is the response variable; $\boldsymbol{x} = (x_1, x_2, ..., x_m)^\top$ are design variables, $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^m$, $\mathcal{X}$ is the design space; $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)^\top$ are unknown parameters; σ is a nuisance parameter, fixed but unknown; and $\pi(\cdot)$ is a probability model. The true value of $\boldsymbol{\theta}$ is known to belong to a set $\Theta \in \mathbb{R}^k$. For each $\boldsymbol{x} \in \mathcal{X}$, an experiment can be performed whose outcome is a random variable (response) $y(\boldsymbol{x})$, where $\text{var}(y(\boldsymbol{x})) = \sigma^2$. Further it is assumed that $y(\boldsymbol{x})$ has an expected value of the explicit form $E(y|\boldsymbol{v}) = \boldsymbol{v}^\top \boldsymbol{\theta}$, where $\boldsymbol{v} \in V$, $V = \{\boldsymbol{v} \in \mathbb{R}^k : \boldsymbol{v} = \eta(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\}$ with $\eta(\boldsymbol{x}) = (\eta_1(\boldsymbol{x}), \eta_2(\boldsymbol{x}), ..., \eta_k(\boldsymbol{x}))^\top$. Thus, V is the image under a set of regression functions η of $\mathcal{X}$, called the induced design space or design locus. Clearly choosing a vector $\boldsymbol{x}$ from the design space $\mathcal{X}$ is equivalent to choosing a k-vector $\boldsymbol{v}$ from the closed bounded k-dimensional induced design space V. Note that typically this design space is continuous however we can assume that V is discrete. Suppose that the space V consists of J distinct vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_J$. In order to acquire an observation on the response y, a value for $\boldsymbol{v}$ must first be selected from the J elements of V to be the point at which to take this observation. An approximate design is characterized by a probability measure $\boldsymbol{p}$ defined on the design space $\mathcal{X}$ and hence on the induced design space V. In practice we discretize these spaces. For example, in the case of one design variable, we approximate the design space by a grid of points equally spaced at intervals between the end points of the design space. Let $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_J\}$. Then $\boldsymbol{p}$ is characterized by a set of weights $p_1, p_2, ..., p_J$ satisfying $p_j \geq 0, j = 1, 2, ..., J$ and $\sum p_j = 1$. The weight p_j is assigned to $\boldsymbol{v}_j$. We wish to optimally choose $\boldsymbol{p} = (p_1, p_2, ..., p_J)$.

If $\hat{\boldsymbol{\theta}}$ is the least squares estimator of $\boldsymbol{\theta}$, then $\text{cov}(\hat{\boldsymbol{\theta}}) \propto M^{-1}(\boldsymbol{p})$, where $M(\boldsymbol{p})$ is the per observation information matrix

$$M(\boldsymbol{p}) = \sum_{j=1}^{J} p_j \boldsymbol{v}_j \boldsymbol{v}_j^\top = VPV^\top = \sum_{j=1}^{J} p_j \eta(\boldsymbol{x}_j) \eta^\top(\boldsymbol{x}_j) \tag{2}$$

where V is the $k \times J$ matrix $[\boldsymbol{v}_1 \boldsymbol{v}_2 ... \boldsymbol{v}_J]$, $P = \text{diag}(p_1, p_2, ..., p_J)$. We wish to choose the proportion p_j of observations taken at $\boldsymbol{x}_j$ to ensure good estimation of $\boldsymbol{\theta}$ by optimizing the I-optimality criterion, $\phi_I(p)$. Our problem is to maximize $\phi_I(\boldsymbol{p})$ subject to $p_j \geq 0, j = 1, 2, ..., J, \sum p_j = 1$. Mathematically, we can write the I-optimality criterion in terms of integrated variance. I-optimality minimizes

$$\phi_I = \int_{\mathcal{X}} \text{V}(\hat{y}(x))\, dx \tag{3}$$

In practice, we choose a set of design points (or factor levels) that lead to minimize ϕ_I. So we discretize the design space $\mathcal{X}$ over a set of points, say $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_J\}$. Then we minimize $\phi_I = \sum_{i=1}^{J} V(\hat{y}(\boldsymbol{x}_i))$ or equivalently, we minimize the total variance. In order to find the I-optimal design, we first define optimality conditions in terms of point to point directional derivatives. Using differential calculus, we exploit the directional derivative of [6]. The directional derivative $F_{\phi_I}\{\boldsymbol{p}, \boldsymbol{q}\}$ of $\phi_I(\cdot)$ at $\boldsymbol{p}$ in the direction of $\boldsymbol{q}$ is defined by

$$F_{\phi_I}\{\boldsymbol{p}, \boldsymbol{q}\} = \lim_{\varepsilon \downarrow 0} \frac{\phi_I((1-\varepsilon)\boldsymbol{p} + \varepsilon \boldsymbol{q}) - \phi_I(\boldsymbol{p})}{\varepsilon} \tag{4}$$

The derivative $F_{\phi_I}\{\boldsymbol{p}, \boldsymbol{q}\}$ exists even if $\phi(\cdot)$ is not differentiable. If $\phi_I(\cdot)$ is differentiable, it can be shown that $F_{\phi_I}(\boldsymbol{p}, \boldsymbol{q}) = (\boldsymbol{q} - \boldsymbol{p})^\top \partial\phi / \partial \boldsymbol{p}$ [7]. Let $F_j = F_{\phi_I}(\boldsymbol{p}, \boldsymbol{e}_j)$ where $\boldsymbol{e}_j$ is the j^{th} unit vector in $\mathbb{R}^J$. We call F_j the vertex directional derivative of $\phi_I(\cdot)$ at $\boldsymbol{p}$.

We will use the Federov algorithm and make use of the I-optimality. The Fedorov algorithm for I-optimality is a powerful tool in experimental design that helps to optimize the precision of parameter estimates in a regression model. This is an iterative procedure as given in the following:

Initialization: A random subset of design points is selected from the candidate set provided in the data argument. This is a set of points in the design space $\mathcal{X}$. For each design point in the current design, calculate the information matrix $M(\boldsymbol{p})$ as given in (2). This initial design is evaluated using the I-optimality criterion.

Iterative Improvement: At each iteration of the algorithm, a design point is swapped with a candidate point from the pool. The swap is accepted if it improves the I-optimality criterion value (reduces the average or total prediction variance, as described in the above methodology).

Convergence: The process continues until no further improvements can be made, or a specified number of iterations is reached. The optimal design must satisfy the optimality conditions. As the I-optimality criterion $\phi_I(\cdot)$ is differentiable at an optimizing distribution $\boldsymbol{p}$, then the first-order optimality conditions are

$$F_j = F_{\phi_I}\{\boldsymbol{p}, \boldsymbol{e}_j\} \begin{cases} = 0 & \text{for } p_j > 0 \\ \leq 0 & \text{for } p_j = 0. \end{cases}$$

where F_j are the vertex directional derivatives of $\phi_I(\cdot)$ at $\boldsymbol{p}$ (4). If a criterion is not concave or convex, a second order condition is used [8]. The above optimality condition is known as the general equivalence theorem in optimal design [9].

Result: The final design is returned, which minimizes the average prediction variance across the design space.

To identify a subset of training points that maximizes predictive efficiency, we used I-optimal design, which selects observations that minimize the average prediction variance across the design space [10, 11]. The I-optimal criterion is well-suited for regression-based prediction because it prioritizes selecting points that improve overall predictive precision rather than only parameter estimation.

I-optimal Design and the Fedorov Exchange Algorithm

The Fedorov exchange algorithm selects an optimal subset of design points from a larger candidate pool. In our study, it operates in four steps:

- **Candidate set.** Within each training fold, we sort the training data by the two predictors and select 100 evenly spaced points.
- **Initial design.** A random subset of 40 candidate points is chosen as the starting design.
- **Swap-and-improve.** Points in the design are iteratively swapped with points outside it, and a swap is accepted if it reduces the I-optimality criterion (average prediction variance).
- **Convergence.** Swapping stops when no further improvement is possible, yielding the I-optimal subset.

Role in the Modeling Pipeline

I-optimal selection is performed only within each training fold to prevent data leakage. Models are then trained on either the full training data or the I-optimal subset, and both are evaluated on the held-out test fold. The full workflow is shown in Figure 1.

2.2 Data Preparation

The dataset for this study was obtained from the CDC and spans the years 2010 to 2021. The data is available in public domain (Centers for Disease Control and Prevention) [12]. It encompasses various HIV-related variables across different states in the U.S., including:

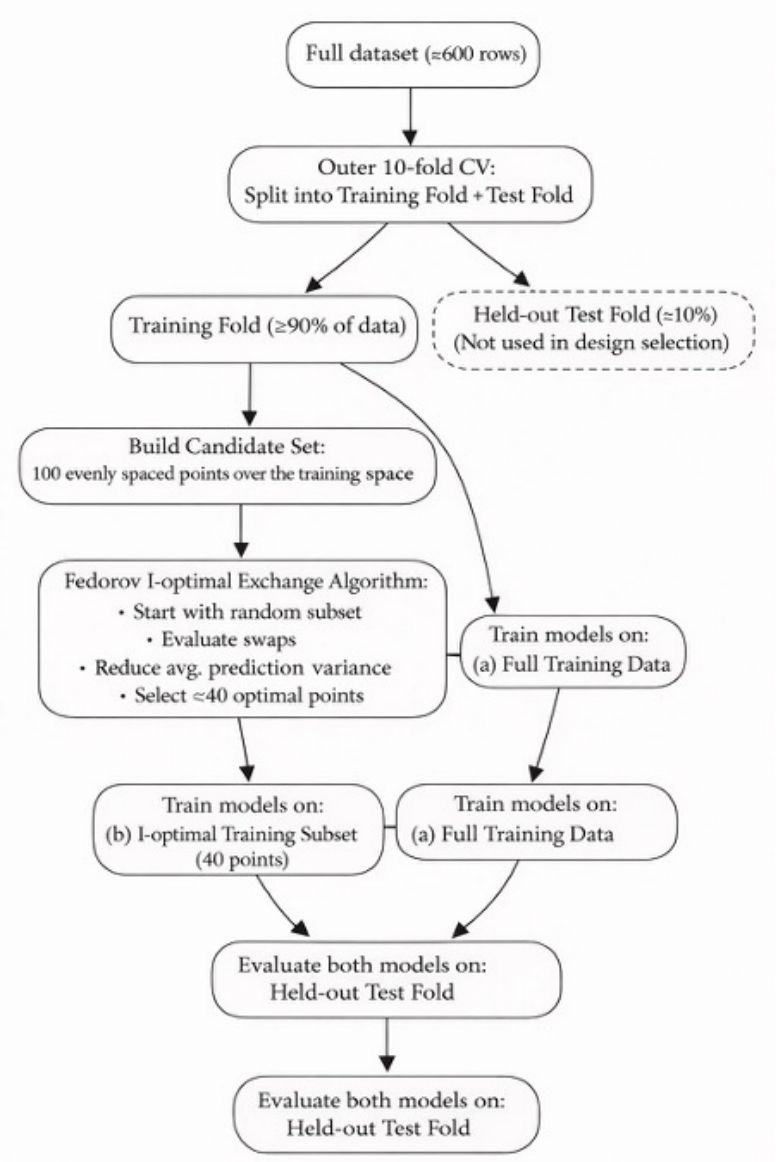


Figure 1: Workflow for integrating the I-optimal design into the modeling pipeline.

- **Geography:** The state or region where the data was collected.
- **Year:** The year in which the data was recorded.
- **HIV Deaths:** The number of deaths due to HIV in each state.
- **HIV Diagnoses:** The number of new HIV diagnoses reported in each state.
- **HIV Prevalence:** The rate of HIV prevalence in each state.

The data preparation process involved several critical steps to ensure the dataset's quality and suitability for analysis:

1. **Handling Missing Values:** Missing values in the dataset were identified as strings such as "Data not available" and "NA." These entries were addressed to prevent them from affecting the analysis. These missing values were excluded due to their impact and prevalence.

2. **Data Dimensions Check:** The dimensions of the dataset were verified to confirm that it loaded correctly and included all necessary variables. This step ensured that the data structure aligned with expectations.

3. **Categorical Variables Conversion:** The variables 'Year' and 'Geography' were converted into factors. This conversion was crucial for facilitating categorical data analysis and modeling, as it allows the software to treat these variables appropriately during statistical analyses.

4. **Filtering Relevant Features:** The dataset was filtered to retain only the key features: HIV deaths, HIV diagnoses, and HIV prevalence. This subset of data focused on the variables directly relevant to the study's objectives. The CDC HIV surveillance dataset did not include regional or socio-economic variables such as income, education, or poverty rates. Therefore, the analysis focused on the epidemiological indicators available across all states and years—HIV deaths, HIV diagnoses, and HIV prevalence. To account for unobserved differences over time and between states, both Year and Geography were modeled as categorical factors.

5. **Correction of Inconsistencies:** Inconsistencies in the 'Geography' variable were identified, such as state names with extraneous characters. These were corrected to ensure consistency and accuracy in the dataset.

6. **Data Reshaping:** The dataset was reshaped from a long format to a wide format. In the long format, data were stacked in a single column with multiple rows, while the wide format has separate columns for each variable. This transformation made it easier to apply machine learning models and perform analysis.

7. **Exclusion of Outliers:** The 'District of Columbia' was excluded from the dataset as it was considered an outlier. This decision was based on the dataset's context and the need for consistent regional data.

8. **Removal of Unnecessary Columns:** Several columns that were deemed irrelevant to the analysis, such as metadata and identifiers, were removed. This step streamlined the dataset and focused it on the variables essential for the prediction tasks.

2.3 Missing Data and Sensitivity Analysis

Following data cleaning, additional diagnostic checks were performed to evaluate the extent and mechanism of missingness and to assess the influence of potential outliers such as the District of Columbia. These sensitivity analyses ensured that the dataset used for modeling was both statistically sound and representative of national HIV trends.

To evaluate data quality and ensure the robustness of the analytical dataset, additional diagnostics were conducted to examine patterns of missingness and potential outliers. Little's MCAR test indicated that missingness in the original dataset was not completely at random ($\chi^2(9) = 3039$, $p < 0.001$). In response, analyses were restricted to complete indicators—HIV deaths, HIV diagnoses, and HIV prevalence—to minimize potential bias associated with nonrandom missingness. After applying these filters, the final analytic dataset consisted of 600 complete state–year observations with no missing values.

The District of Columbia (DC) was further assessed for its potential influence due to its disproportionately high HIV rates compared to other states. To determine its impact, models were refitted with and without DC observations. The resulting predictions were highly consistent (correlation $r \geq 0.996$), with only modest differences in fit statistics ($\Delta R^2 = -0.3\%$ to $+14.5\%$; ΔRMSE up to -68%) and year-trend estimates (17–46%

change). Despite this, DC exhibited moderate influence (|DFBETAS| up to 4.2), suggesting that its inclusion introduced instability in parameter estimates and reduced cross-validation consistency. Therefore, to enhance model stability and maintain comparability across states, all subsequent analyses were conducted excluding the District of Columbia.

2.4 Data Split

After preparing the dataset, the final version contained 600 rows, each representing a unique geographical region. The dataset was structured to include several HIV-related indicators with the following columns:

- **HIV Deaths:** The count of HIV-related deaths per region.
- **HIV Diagnoses:** The number of new HIV diagnoses per region.
- **HIV Prevalence:** The rate of HIV prevalence per region.

Each indicator provided two types of information:

- **Cases:** The raw count of reported cases.
- **Rate per 100,000:** The standardized rate of occurrence per 100,000 population.

2.5 Data Split and Validation

To facilitate the development and validation of machine learning models, the dataset was partitioned into training and testing sets using the following criteria.

Because the analytic dataset was modest in size (approximately 600 state–year observations), we evaluated all models using nested k-fold cross-validation rather than a single train–test split. Specifically, we used an outer 10-fold cross-validation to estimate out-of-sample performance. In each outer fold, approximately 90% of the data were used for training and 10% were held out for testing. Within each outer training fold, we performed an inner 5-fold cross-validation to tune hyperparameters for the non-linear models (Random Forest, Gradient Boosting, XGBoost, and Support Vector Regression). Model performance was summarized across the 10 outer folds using the mean and standard deviation of the mean squared error (MSE) and coefficient of determination (R^2).

To evaluate the impact of I-optimal design, we implemented the design selection within each training fold only, thereby avoiding any information leakage from the test data. For a given outer fold, the training data were first sorted by HIV diagnosis and HIV prevalence rates, and a candidate set of up to 100 evenly spaced points was constructed. The Fedorov algorithm with an I-optimality criterion was then applied to this candidate set to select a subset of 40 design points. Models were trained either on (i) the full training data or (ii) the I-optimal training subset, and in both cases were evaluated on the same held-out test fold.

2.6 Software and Computational Environment

All analyses were conducted in R using RStudio. The main packages employed included *tidyverse* for data wrangling and visualization, *caret* for model training and cross-validation, *randomForest*, *gbm*, and *xgboost* for ensemble learning, *e1071* for support vector regression, and *AlgDesign* for implementing I-optimal design. Additional packages such as *performance, broom, naniar, BaylorEdPsych*, and *mvnmle* were used for model diagnostics, handling missing data, and statistical testing.

3. MODEL SELECTION AND APPROACHES

There are a diverse array of machine learning algorithms can be employed to model and predict continuous outcomes such as multiple linear regression, k-nearest neighbors, decision trees, random forest, gradient boosting, extreme gradient boosting and support vector machines. In this study, we have implemented several of these approaches. These methods fall under the category of supervised machine learning, which utilizes labeled datasets to train algorithms that either classify data or predict outcomes with high accuracy. Supervised learning enables organizations to address various real-world challenges on a large scale, such as separating spam emails from your inbox. In supervised learning, each training example is paired with an output label, and the algorithm learns to map inputs to these outputs. The goal is to build a model that can make accurate predictions or classifications on new, unseen data based on what it has learned from the training data. In essence, supervised learning algorithms learn from the training data to make informed predictions or decisions based on new inputs. Supervised learning can be separated into two types of problems when data mining—classification and regression. Classification algorithms assign test data to specific categories, recognizing entities and determining their labels. Regression analyzes the relationship between dependent and independent variables.

Five predictive models were evaluated to capture both linear and nonlinear relationships between HIV mortality and the explanatory indicators (HIV diagnoses and HIV prevalence): Multiple Linear Regression (MLR), Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR). This combination balances interpretability with predictive flexibility across algorithmic families. Methods such as k-Nearest Neighbors and Neural Networks were not applied because the dataset (600 observations $\times$ 9 variables) was relatively small, increasing the risk of overfitting for high-capacity models. All models were implemented in R using 10-fold cross-validation. Key tuning parameters were: RF ($ntree = 100$, $mtry = \sqrt{p}$); GBM ($n.trees = 100 - 1000$, $interaction.depth = 1 - 5$, $shrinkage = 0.1$, $bag.fraction = 0.6$); XGBoost ($nrounds = 50 - 500$, $eta = 0.05 - 0.3$, $max_depth = 1 - 5$, $subsample = 1$); and SVR (radial kernel, $cost = 1$, $epsilon = 0.2$). The MLR model was fitted using ordinary least squares without regularization.

3.1 Linear Regression Model

Linear Regression is a fundamental statistical method used to examine the relationship between a dependent variable and one or more independent variables. This technique is widely applied to understand how changes in independent variables affect the dependent variable and to make predictions about future outcomes. When the analysis involves only one independent variable and one dependent variable, the method is known as simple linear regression. When two or more independent variables are used to predict a single dependent variable, the analysis is called multiple linear regression [13, 14]. This extension of simple linear regression accommodates more complex scenarios in which multiple factors influence the dependent variable. The relationship is represented by a plane or hyperplane in higher dimensions and is described by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{5}$$

For multiple linear regression, the coefficients are estimated using the least squares method to minimize the difference between the observed and predicted values, providing the best-fitting hyperplane to the data.

To explore the relationship between the response variable, the rate of HIV-related deaths per 100,000 population, and the two predictors—rate of new HIV diagnoses per 100,000 population, and the rate of HIV prevalence per 100,000 population—we employed multiple linear regression to build a predictive model. By developing this model, we can identify which predictors are insignificant and remove those that do not contribute effectively to predicting the rate of HIV-related deaths per 100,000 population. To ensure the precision of our predictive

model for the rate of HIV-related deaths per 100,000 population, we checked for collinearity among the predictors. This process helps us refine the model to make it more appropriate for prediction.

3.2 Random Forest

Random forest regression is a supervised learning algorithm that employs an ensemble method by building multiple decision trees and aggregates their predictions. In this approach, multiple trees are generated using bootstrap samples from the training data, and efforts are made to reduce the correlation between the trees [15]. By combining the results from several trees, the random forest algorithm improves accuracy and reduces overfitting compared to individual decision trees. It is a special case of bagged tree model that improves upon bagged trees by introducing a small tweak that decorrelates the trees. In bagging certain features may be very important, leading to highly correlated decision trees. To mitigate this, random forest randomly selects a subset of features, decorrelating the trees, and reducing the variance when the trees are averaged.

Similarly to bagging, random forests build multiple decision trees on bootstrapped training samples. However, when constructing these trees, a random selection of m predictors is chosen as candidates from the full set of p predictors, and the split is allowed to use only one of those m predictors. A new selection of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$, meaning that the number of predictors considered at each split is approximately equal to the square root of the total number of predictors.

3.3 Gradient Boosting Machine

Gradient Boosting Machine, a powerful machine learning technique that builds models sequentially by optimizing the residual errors of previous models. It belongs to the family of ensemble methods and is commonly used for both regression and classification tasks. GBM [16] is based on the boosting technique, which combines the strengths of multiple weak learners (typically decision trees) to create a strong predictive model. Each new tree is built to correct the errors made by the previous ones. GBM uses gradient descent to minimize the loss function (the error or difference between the predicted and actual values). The model iteratively adds trees that point in the direction of the steepest descent, effectively reducing the error with each step. Each tree in the sequence is trained on the residuals (errors) of the previous tree. The idea is to reduce the residuals at each step, gradually improving the model's predictions. GBM includes a learning rate parameter that controls how much each tree contributes to the final model. A smaller learning rate generally leads to better performance but requires more trees.

3.4 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an advanced implementation of the gradient boosting algorithm designed for high efficiency and performance. It is a scalable and highly optimized machine learning algorithm based on the gradient boosting framework. Unlike traditional gradient boosting, XGBoost [17] introduces a number of algorithmic enhancements that make it particularly effective for large-scale and complex predictive modeling tasks. This method is known for its speed, accuracy, and robustness, making it a popular choice in both academic research and industry applications. It is a "super charged" version of boosted trees and is one of the main state-of-the-art methods. While it might not work better than bagging or boosting in all circumstances, it has demonstrated impressive performance on very hard ML problems. It's particularly well-suited for structured/tabular data and has gained significant popularity due to its success in various machine learning competitions and practical applications.

3.5 Support Vector Regression

A support vector machine is a popular supervised learning model developed by Vladimir Vapnik, used for both data classification and regression. It is a widely used classifier recognized for accuracy, flexibility and robustness. Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) [18, 19] algorithm, which is primarily used for classification tasks, into the domain of regression. SVR is particularly useful in scenarios where the goal is to predict a continuous outcome variable. It is built upon the principles of Support Vector Machines (SVMs), a popular classification algorithm, and adapts these principles to predict continuous outcomes rather than discrete class labels. The core idea of SVR is to find a function that approximates the relationship between input variables and the continuous output, while maintaining a margin of tolerance within which predictions are allowed to deviate from the actual observed values.

4. RESULTS AND DISCUSSION

4.1 Model Performance Using Full Data vs. I-Optimal Subsets

In this study, we applied five different machine learning models to predict HIV death rates per 100,000 people based on the HIV diagnosis and prevalence rates. The models used were multiple linear regression (MLR), random forest (RF), gradient boosting machine (GBM), extreme gradient boosting with hyperparameter tuning (XGB-tuned), and support vector regression (SVR) - using nested 10-fold cross-validation. For each outer fold, models were trained either on the full training set or on a 40-point I-optimal subset constructed exclusively from that training fold. Performance was assessed using mean squared error (MSE) and the coefficient of determination (R^2) on the held-out test fold.

Visual Comparison of Model Performance (Full vs I-optimal)

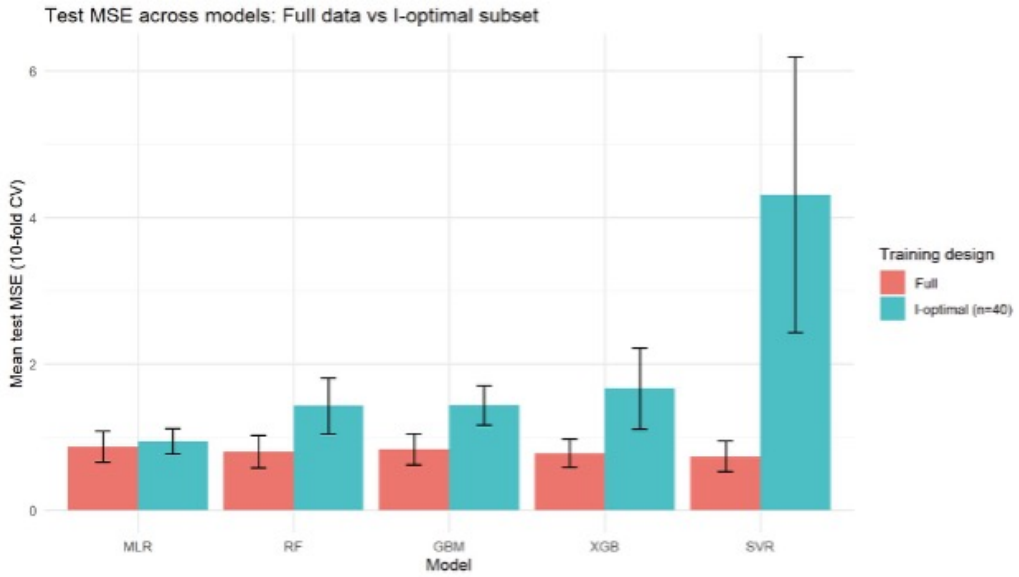


Figure 2: Comparison of mean test MSE across models trained on the full dataset versus the I-optimal subset.

Figure 2 shows the mean MSE for each model under the full-data and I-optimal training regimes. When trained on the full dataset, the nonlinear models exhibited the strongest predictive accuracy, with SVR achieving the lowest MSE (0.74), followed by XGBoost (0.78), RF (0.80), GBM (0.83), and MLR (0.87).

Training on the I-optimal subset increased prediction error for all models, but especially for the nonlinear methods. MSE rose to 1.43 for RF, 1.44 for GBM, 1.67 for XGBoost, and 4.31 for SVR, indicating reduced ability to learn complex patterns when trained on a much smaller design-optimized subset. In contrast, MLR showed only a modest increase in error (MSE 0.94), consistent with the fact that I-optimality is specifically designed to improve prediction efficiency for linear models.

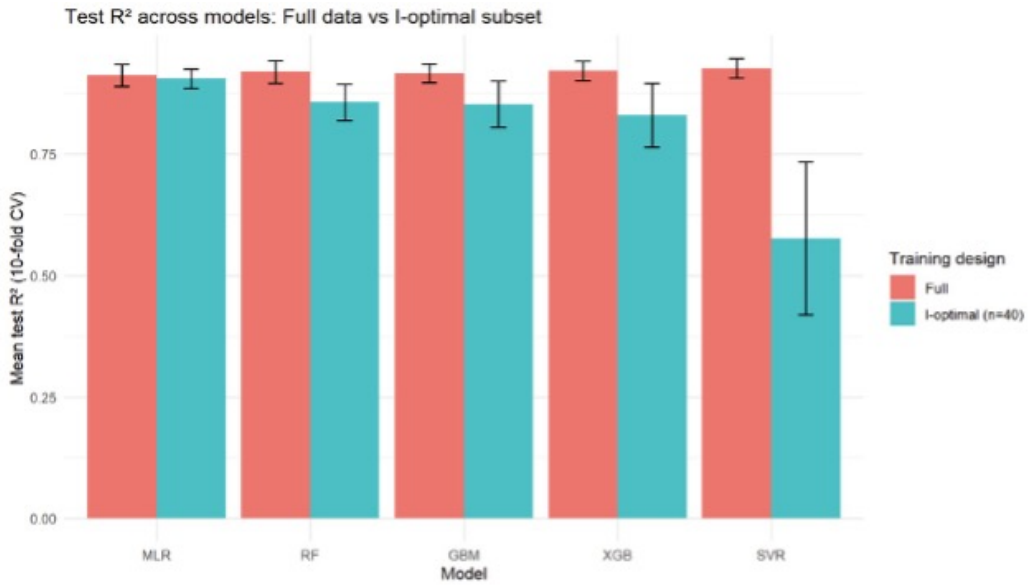


Figure 3: Comparison of mean test R^2 values across five predictive models trained on the full dataset versus the I-optimal subset, based on 10-fold cross-validation.

Figure 3 displays the corresponding R^2 values. Full-data models achieved high predictive power (R^2 ranging 0.91–0.93), with SVR and XGBoost performing best. Under I-optimal subsampling, R^2 values declined for all models, again most substantially for nonlinear methods (e.g., SVR R^2 dropped from 0.93 to 0.58). MLR maintained a relatively high R^2 (0.90) even with only 40 design points, highlighting the suitability of I-optimality for linear prediction tasks.

4.2 I-optimal Design and Data Geometry

To illustrate how I-optimality selects training points, Figure 4 overlays the 40-point I-optimal subset on the full predictor space defined by HIV diagnosis and prevalence rates. The selected points are spread across the domain and cover regions that contribute most to reducing average prediction variance under a linear model.

Figure 5 plots the prediction standard error (SE) surface from the MLR model, with I-optimal points overlaid. The I-optimal subset tends to cluster in areas with higher SE, demonstrating that the algorithm preferentially samples locations where additional information is most valuable. This visual evidence helps explain why linear modeling performance remains stable even with substantial data reduction, whereas nonlinear models—which rely on dense sampling to learn complex interactions—show a pronounced decline in accuracy when trained on the reduced subset. Together, Figures 4 and 5 show how I-optimal design shapes the geometry of the training data and why its benefits are model-dependent.

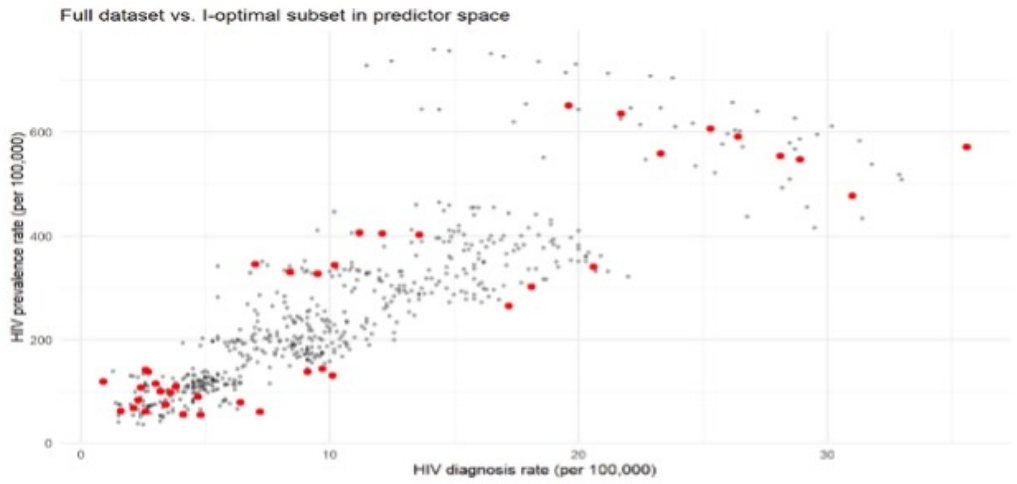


Figure 4: I-optimal design points overlaid on the full predictor space. The scatterplot shows how the I-optimal subset (red points) spans the range of HIV diagnosis and prevalence rates in the training data, illustrating how the design selects points that efficiently cover the predictor domain.

4.3 Summary of Performance Patterns

The cross-validated results highlight clear differences in how models respond to I-optimal subsampling. Linear models such as MLR retain most of their predictive performance because I-optimality directly minimizes linear prediction variance. In contrast, nonlinear models—including RF, GBM, XGBoost, and SVR—experience systematic degradation in accuracy, as the 40-point I-optimal subset does not preserve the richer interaction structure these algorithms rely on to capture complex patterns. SVR, which performs best on the full dataset, is the most adversely affected, reflecting its sensitivity to the diversity and density of the training data. Overall, these findings underscore the importance of aligning data design strategy with the modeling approach: while optimal design can enhance efficiency for linear models, nonlinear machine learning methods generally require larger and more comprehensive datasets to achieve optimal predictive performance.

4.4 Discussion

The study highlights the trade-offs between dataset design and model performance, particularly when comparing a full dataset to an I-optimal design subset of 40 design points. Although I-optimality is theoretically designed to reduce prediction variance for linear models, our cross-validated results showed that Multiple Linear Regression (MLR) maintained comparable—but slightly lower—predictive accuracy when trained on the I-optimal subset relative to the full dataset. This indicates that while the I-optimal subset preserves essential linear structure, it does not empirically improve MLR performance under repeated subsampling. The modeling results further show that HIV diagnosis rates are inversely associated with mortality, suggesting that expanding early testing and diagnosis reduces HIV-related deaths through timely treatment initiation and improved retention in care. In contrast, higher HIV prevalence rates are positively associated with mortality, reflecting the cumulative disease burden and ongoing care demands within more heavily affected states. Over the period 2010–2021, the models captured a gradual decline in HIV mortality, consistent with national improvements in antiretroviral therapy (ART) access and testing coverage, although declines were slower in Southern states where prevalence remains disproportionately high.

From a modeling standpoint, I-optimal design selection maintained reasonable stability for linear models

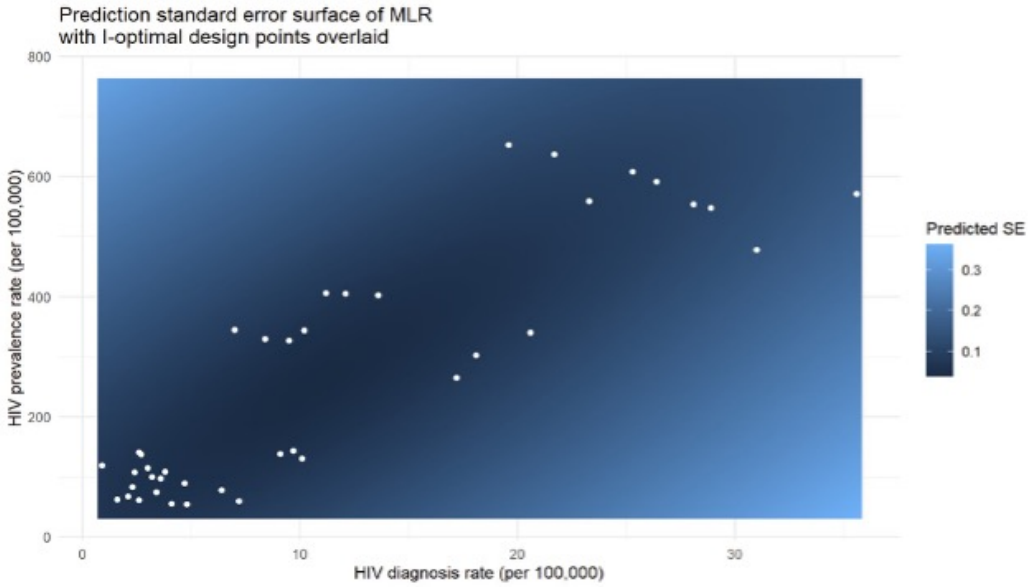


Figure 5: Prediction standard error surface from the multiple linear regression model with I-optimal design points overlaid. The I-optimal points concentrate in regions with higher prediction uncertainty, demonstrating how the design minimizes the average prediction variance across the predictor space.

but produced substantial reductions in predictive performance for nonlinear methods such as Random Forest (RF), Gradient Boosting Machines (GBM), XGBoost, and Support Vector Regression (SVR). These nonlinear algorithms rely on larger, more diverse training samples to capture interaction effects and complex nonlinear relationships that a 40-point I-optimal subset cannot represent. SVR, the strongest-performing model on the full dataset, experienced the most pronounced degradation under I-optimal subsampling, underscoring its sensitivity to training sample diversity. These results align with epidemiologic and statistical theory: linear models effectively summarize broad population-level associations and therefore remain relatively robust to variance-optimized subsampling, whereas flexible machine learning models require denser and more heterogeneous datasets to model nonlinearities accurately. Thus, while I-optimality is a powerful tool for experimental efficiency in linear modeling, its utility diminishes for nonlinear, interaction-heavy machine learning approaches that thrive on larger design spaces. These observations are consistent with prior work showing that I-optimal designs minimize prediction variance for linear models, whereas nonlinear models require richer data coverage to capture curvature and interaction structures effectively. [20, 21]

4.5 Implications

The findings underscore critical considerations for designing datasets tailored to predictive modeling studies. In settings where models such as multiple linear regression (MLR) are sufficient, or where computational resources or data collection capacity are limited, I-optimal design remains a valuable strategy because it preserves essential linear information while substantially reducing dataset size. Although the I-optimal subset did not improve MLR performance in our cross-validated analysis, it maintained comparable predictive accuracy to the full dataset, demonstrating its practical utility in scenarios requiring efficient, cost-conscious data collection without major loss of interpretability or predictive capability.

In contrast, the results clearly show that advanced nonlinear machine learning models—including Random Forests (RF), Gradient Boosting Machines (GBM), XGBoost, and Support Vector Regression (SVR)—require

the richness and diversity of the full dataset to capture complex interactions and nonlinear relationships inherent in epidemiological processes. As these models exhibited substantial reductions in predictive accuracy when trained on the I-optimal subset, our findings suggest that researchers and public health practitioners intending to use flexible learning algorithms should prioritize broad, comprehensive data collection strategies.

Consequently, the implications of this study emphasize the importance of aligning data design methodologies with analytical objectives. I-optimal designs may be advantageous when linear modeling suffices or when resources constrain data acquisition, whereas nonlinear machine learning applications benefit from larger, more heterogeneous datasets that adequately represent underlying complexity. Balancing efficiency with the demands of the chosen modeling framework is therefore essential for producing reliable and informative predictive analyses in public health settings.

5. CONCLUSION

The comparison between the full dataset and the I-optimal design subset demonstrates the important role of data design in shaping predictive performance across different modeling approaches. Our results show that while I-optimality preserves the essential structure needed for linear prediction, it does not improve the performance of Multiple Linear Regression (MLR) under cross-validation. Instead, it provides a more compact training set that maintains comparable accuracy, making it a practical option in settings where data collection is costly or sample sizes are inherently limited. In contrast, nonlinear machine learning methods—including Random Forests (RF), Gradient Boosting Machines (GBM), XGBoost, and Support Vector Regression (SVR)—exhibited substantial reductions in predictive accuracy when trained on the I-optimal subset. These models rely on rich and diverse training data to capture nonlinearities and interactions, and thus benefit far more from the full dataset.

The use of the Fedorov exchange algorithm to construct I-optimal designs highlights how optimal design theory can strategically reduce dataset size while retaining key predictive information for linear models. This approach has particular relevance in resource-constrained public health environments, where robust inference must sometimes be drawn from limited or selectively collected data. However, the diminished performance of nonlinear models under I-optimal subsampling underscores that optimal design is not universally advantageous—its utility depends critically on both the modeling framework and the complexity of the underlying data-generating process.

Overall, the findings emphasize the need to align data design methodology with the analytical goals and model class. While I-optimality offers efficiency and stability for linear modeling, flexible machine learning models require larger, more heterogeneous datasets to achieve high predictive accuracy. Future research may explore hybrid strategies that integrate optimal design principles with modern machine learning workflows, enabling researchers to balance computational efficiency with the need to preserve the structural richness required for complex predictive tasks.

REFERENCES

[1] Centers for Disease Control and Prevention (CDC), "Advancing HIV prevention: new strategies for a changing epidemic–United States, 2003," *MMWR. Morbidity and Mortality Weekly Report*, vol. 52, no. 15, pp. 329–332, 2003.

[2] Centers for Disease Control and Prevention (CDC), *Compendium of HIV prevention interventions with evidence of effectiveness*, 1999.

[3] T. Sanchez, T. Finlayson, A. Drake, S. Behel, M. Cribbin, E. Dinenno, T. Hall, S. Kramer, A. Lansky, and Centers for Disease Control and Prevention (CDC), "Human immunodeficiency virus (HIV) risk,

prevention, and testing behaviors–United States, National HIV Behavioral Surveillance System: men who have sex with men, November 2003-April 2005," *Morbidity and Mortality Weekly Report. Surveillance Summaries*, vol. 55, no. 6, pp. 1–16, 2006.

[4] H. I. Hall, R. Song, P. Rhodes, J. Prejean, Q. An, L. M. Lee, J. Karon, R. Brookmeyer, E. H. Kaplan, M. T. McKenna, and R. S. Janssen, "Estimation of HIV incidence in the United States," *JAMA*, vol. 300, no. 5, pp. 520–529, 2008.

[5] X. Tian, J. Chen, X. Wang, Y. Xie, X. Zhang, D. Han, H. Fu, W. Yin, and N. Wu, "Global, regional, and national HIV/AIDS disease burden levels and trends in 1990-2019: A systematic analysis for the global burden of disease 2019 study," *Frontiers in Public Health*, vol. 11, 1068664, 2023.

[6] P. Whittle, "Some general points in the theory of optimal experimental design," *Journal of the Royal Statistical Society, Series B*, vol. 35, pp. 123–130, 1973.

[7] S. Mandal and B. Torsney, "Construction of optimal designs using a clustering approach," *Journal of Statistical Planning and Inference*, vol. 136, pp. 1120–1134, 2006.

[8] S. Mandal, B. Torsney, and M. Chowdhury, "Optimal designs for minimising covariances among parameter estimators in a linear model," *Australian & New Zealand Journal of Statistics*, vol. 59, pp. 255–273, 2017.

[9] J. Kiefer, "General Equivalence theory for optimum designs (approximate theory)," *Annals of Statistics*, vol. 2, pp. 849–879, 1974.

[10] V. V. Fedorov, *Theory of optimal experiments*. Academic Press (Elsevier), 2013.

[11] J. Lopez-Fidalgo, *Optimal Experimental Design*. Springer, 2023.

[12] Centers for Disease Control and Prevention (CDC), "NCHHSTP AtlasPlus: HIV Data Tables," 2024. [Online]. Available: https://gis.cdc.gov/grasp/nchhstpatlas/tables.html

[13] A. Schneider, G. Hommel, and M. Blettner, "Linear regression analysis," *Dtsch Arztebl Int*, vol. 107, no. 44, pp. 776–782, 2010.

[14] C. Michelle and E. H. Kathryn, "Chapter 27 - regression, translational radiation oncology," in *Handbook for designing and conducting clinical and translational research*, pp. 145–152, 2023.

[15] G. Debopriya and C. Javier, "Enriched random forest for high dimensional genomic data," *IEEE/ACM Trans Computat Biol Bioinform*, vol. 19, p. 5, 2022.

[16] H. Friedman, "Stochastic Gradient Boosting," *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[17] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science and Business Media, 2010.

[19] J. Fan, R. Li, C.-H. Zhang, and H. Zou, *Statistical foundations of data science*. Chapman and Hall/CRC, 2020.

[20] L. Yiou and D. Xinwei, "An efficient algorithm for Elastic I-optimal design of generalized linear models," *Canadian Journal of Statistics*, vol. 49, no. 2, pp. 438–470, 2020.

[21] M. P. F. Berger and F. E. S. Tan, "Robust Designs for Linear Mixed Effects Models," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 53, no. 4, pp. 569–581, 2004.