Does Cohort Selection Affect Machine Learning from Clinical Data?

Atefehsadat Haghighathoseini[1], Janusz Wojtusiak[1], Hua Min[1], Timothy Leslie[1], Cara Frankenfeld[2], Nirup M Menon[1,3] (On behalf of the N3C Consortium[1])

[1]George Mason University, Fairfax, VA, USA; [2]MaineHealth Institute for Research, Scarborough, ME, USA
[3]On behalf of the N3C Consortium

Abstract
This study investigates cohort selection and its effects on the quality of machine learning (ML) models trained on clinical data, focusing on measurements taken within the first 48 hours of hospital admission. It discusses the potential repercussions of making arbitrary decisions during data processing prior to applying ML methods. Experiments are performed within the framework of the National COVID Cohort Collaborative (N3C) dataset. The research aims to unravel biases and assess the fairness of machine learning models used to predict outcomes for hospitalized patients. Detailed discussions cover the data, decision-making processes, and the resulting impact on model predictions regarding patient outcomes. An experiment is conducted in which four arbitrary decisions are made, resulting in 16 distinct datasets characterized by varying sizes and properties. The findings demonstrate significant differences in the obtained datasets and indicate a high potential for bias based on inclusion or exclusion decisions. The results also confirm significant differences in the performance of models constructed on different cohorts, especially when cross-compared between ones based on different inclusion criteria. The study specifically chose to analyze gender, race, and ethnicity as these social determinants of health played a significant role in COVID-19 outcomes.

Keywords— *Prediction, Selection Bias, Data Processing, Machine Learning, National COVID Cohort Collaborative (N3C)*

Introduction

In health data processing, decisions regarding cohort and variable selection often lack standardization or reproducibility, relying on the intuition of researchers and existing literature. This common practice creates a significant gap in the assessment of the impact of these choices on data analysis models. Healthcare stakeholders must be vigilant in conducting unbiased data analyses and recognize the potential limitations stemming from arbitrary decisions. Overlooking seemingly minor choices during data processing can result in selection bias and alterations to the dataset. Strict inclusion/exclusion criteria may inadvertently exclude important subpopulations, affecting generalizability. Conversely, lenient inclusion/exclusion criteria may increase the likelihood of introducing noise into the data. A thorough evaluation of these decisions is crucial during data processing.

Selection bias arises when the chosen cohort inadequately represents the broader population under study, resulting in diminished external validity or generalizability. This occurs when individuals, groups, or data are selected for analysis without proper randomization, leading to a sample that may not accurately reflect the population [1]. James Heckman's "Varieties of Selection Bias" in econometrics explores the consequences of selection bias on estimating the effects of specific variables [2]. Furthermore, selection bias can inflate effect sizes and yield inaccurate results, undermining the reliability of statistical tests and the validity of research findings [3].

Electronic medical records (EMR) are valuable data sources for assessing health outcomes. However, using EMRs and other health data in clinical settings introduces several biases that can significantly impact data analysis results, subsequent healthcare policies, and patient outcomes. These biases may arise from arbitrary decisions in cohort criteria and variable selection, patient selection, and data completeness [7]. Significant efforts have focused on examining and mitigating selection bias in conventional statistics and experimental research. For example, a study on participants' arbitrary choices between two equally favored options demonstrated that such choices could significantly affect their preferences [4]. This highlights the substantial impact that arbitrary decisions can have on study outcomes. The consequences of bias extend across various domains, including clinical practice, research, and decision-making. In healthcare, implicit bias has been recognized as a significant contributor to disparities, ultimately influencing patient

[1] Authorship was determined using ICMJE recommendations. The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave covid.cd2h.org/enclave and supported by CD2H - The National COVID Cohort Collaborative (N3C) IDeA CTR Collaboration 3U24TR002306-04S2 NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data from participating organizations (covid.cd2h.org/dtas) and the organizations and scientists (covid.cd2h.org/duas) who have contributed to the on-going development of this community resource. See Hanendel et al 2021 for details.

outcomes [13]. In outcome prediction, selection bias can introduce systematic errors into predictions generated by machine learning (ML) models. For example, models trained on nonrepresentative data may yield inaccurate predictions when applied to a broader population [7]. Suri [8] and MacNamee [9] highlight the critical role of bias in ML systems for outcome prediction, especially in cardiovascular disease risk assessment and anticoagulant drug therapy. Researchers have utilized various data sources, including electronic health records (EHRs), claims data, and registries, to analyze and evaluate biases. This comprehensive approach enhances the understanding of factors affecting the quality and reliability of clinical data. Recognizing and addressing these biases is essential for improving healthcare policies and patient outcomes.

This study delves into the influence of cohort selection on the efficacy of ML models trained using clinical data. It explores how arbitrary decisions made during data processing can affect the quality and fairness of ML models within the National COVID Cohort Collaborative (N3C) dataset. By examining biases and evaluating the fairness of ML models in predicting outcomes for hospitalized patients, particularly in the context of N3C, this research sheds light on the importance of thoughtful cohort selection. The study specifically chose to analyze gender, race, and ethnicity as these social determinants of health played a significant role in COVID-19 outcomes. The study conducted an experiment involving four arbitrary decisions, resulting in 16 distinct datasets with varying sizes and characteristics. Detailed discussions cover the data, decision-making processes, and the resulting impact on model predictions regarding patient outcomes.

Design of the Experiment

This study explores the impact of data preprocessing decisions, particularly cohort selection, on the quality of machine learning models. Traditionally, machine learning projects for predicting patient outcomes follow a standard pipeline (Figure 1 top). After cohort selection, input and outcome variables are generated, and the data is split into training and testing sets. The training set is used to develop and fine-tune models to improve the AUC, while the testing set is reserved for the final evaluation of the models. However, this approach often involves making data processing decisions, such as cohort selection, before splitting the data. This practice limits the ability to assess how well the model generalizes beyond the specific training and testing sets.

This study generates multiple versions of the final dataset to investigate model generalizability, analyzing each version from the point of data splitting into training and testing sets. A matched training/testing split was used to ensure that the same patient did not appear in the training and testing sets across cohorts. Models are trained and tested uniformly, and results are compared to highlight cross-testing performance across cohorts (Figure 1 bottom).

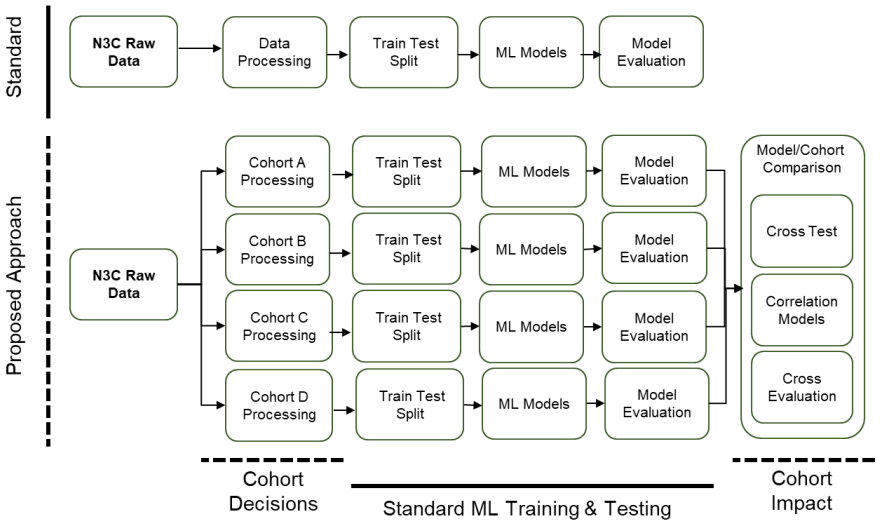

Figure 1. Standard (top) and proposed (bottom) design of experiments.

Data Source:

The National COVID Cohort Collaborative (N3C) is an indispensable tool in the study of COVID-19, offering a rich dataset for detailed analysis through its specialized N3C Data Enclave [10]. This initiative capitalizes on integrating diverse health records from multiple medical institutions, establishing a unified structure that aids researchers in accessing reliable and comprehensive information. More specifically, the presented study used a Limited Data Set ("LDS"), providing access to individual patient-level data with removed patient identifiers.

The focus of this study revolves around predicting outcomes for individuals admitted to hospitals with COVID-19. The primary examination unit is the patient's hospitalization, and the goal is to forecast the associated outcome. The N3C data, structured across various tables within a relational database, adhere to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standard. The necessary transformation involves converting this data into a format where each hospitalization corresponds to a single record (flat table, sometimes referred to as an analytic file). This format is standard for applying both ML methods and conventional statistical tools.

Data Preprocessing Decisions

The initial and critical step in data analysis for constructing machine learning models is cohort creation. This study employed a systematic approach to build cohorts from approximately 21 million N3C patients. As shown in Figure 2, this approach involved identifying COVID-19 cases, recognizing inpatient hospitalization records, determining COVID-19-related hospitalizations, and excluding records missing specific admission time data. While other preprocessing decisions were made, they are not explored in detail in this work for simplicity. For example, the data were filtered to include only records with a first COVID-19 diagnosis between August 1, 2020, and December 31, 2021. This timeframe was selected as the study target period, with data prior to August 2020 excluded due to unreliable early pandemic reporting and data post-2021 removed to maintain a fixed cohort timeframe. While reasonable, the choice of these specific dates does not have a precise clinical justification. Hospitalization records were merged with other data using a unique identifier and analyzed for features such as gender, date of birth, race, ethnicity, and age at death. Non-adult patients were excluded, with the study focusing on individuals aged 18 and older, though there is no specific clinical rationale for choosing 18 over other ages like 21. The application of these four preprocessing decisions results in 16 potential datasets. The optimal choice among these datasets for final analysis remains uncertain, as different options may be argued for at each decision level. Detailed descriptions of these decisions are available in[11]. Although seemingly arbitrary, these choices lead to 16 distinct datasets characterized by varying sizes and properties. This study analyzes the selection of four cohorts (A, B, C, and D) but holds relevance for all potential datasets. These decisions, along with corresponding annotations, are depicted in Figure 2.

The subsequent process involves extracting all pertinent measurements from the measurement table for patients within specific cohorts of the N3C database. These measurements are then meticulously categorized based on their concept IDs, which are distinctly selected from the concept set member's table. The concept IDs encompass a wide range of medical measurements, including but not limited to Respiratory Rate, GFR (Glomerular Filtration Rate), Cardiac Troponin T, Sodium, ABG indices, Prothrombin Time, Alanine Transaminase, Calcium, Systolic BP (Blood Pressure), Heart Rate, CD3 CD8 T cells, Albumin, Erythrocyte Sedimentation Rate, Venous Lactate, BMI (Body Mass Index), CBC with PLT (Complete Blood Count with Platelets), CD3 CD4 T cells, Cardiac Troponin I, Interleukin 6, FiO2, Blood Urea Nitrogen, Temperature, Bilirubin, D-Dimer, Alkaline Phosphatase, fibrinogen, Creatinine, Chloride,

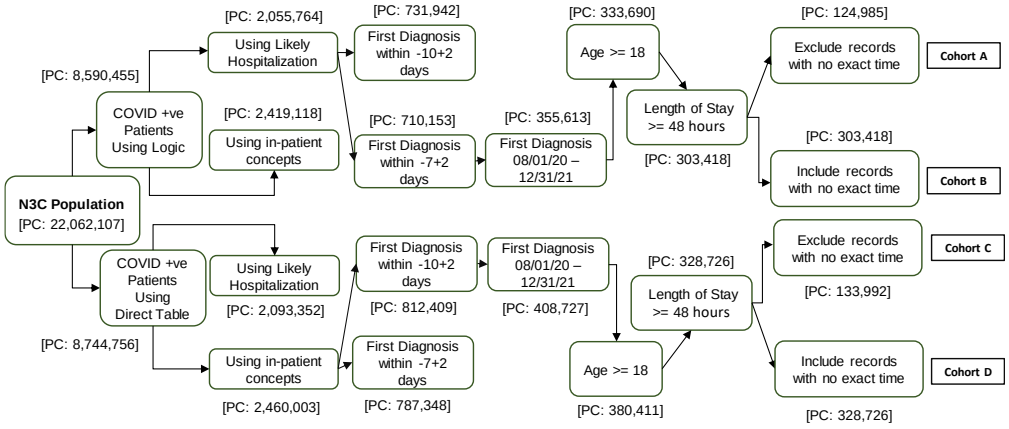

Figure 2. Partial inclusion/exclusion tree outlining extreme choices for the four decisions. PC - patient counts.

Glucose, C-Reactive Protein, Urine Output, NT-ProBNP, Diastolic BP, Weight, Ferritin, and Interleukin 10. This comprehensive extraction and categorization are focused on the critical initial 48 hours of hospital admission, targeting the acute phase of the patient's stay. This approach is designed to facilitate clinical decision-making and support research efforts by providing a structured dataset of key patient measurements during this pivotal period of hospitalization.

Cohort Characteristics

There is a clear difference in the size of the cohorts, ranging from 124,985 to 328,726 patients, indicating an almost three-fold increase in population size. The most significant impact on the cohort size results from the last decision, which is purely data-driven (see Figure 2) and accounts for approximately 60% of the reduction in cohorts. The inclusive cohorts, determined by the timestamp decision (Cohorts B & D), show higher proportions of males, Asians, and Hispanics/Latinos or individuals with unknown ethnicity than their counterparts. Descriptive statistics for gender, race, and ethnicity across the four cohorts reveal notable disparities. For instance, Cohort A has the lowest number of males (64,953), while Cohort D has the highest number of females (161,188). Examining the racial distribution across cohorts, we observe Cohort D generally has the highest counts across all racial categories, indicating a larger overall patient population in that cohort. Cohort A consistently has the lowest counts across most racial categories. The distribution of races significantly differs among the cohorts, as indicated by the low p-values from the Chi-squared test. This suggests that the racial composition varies considerably between the cohorts. Cohort D also has the highest

Table 1. Cohorts and combinations of specialized test datasets, along with their gender, race, and ethnicity subgroups

		Gender*			***Ethnicity****			***Race****					
Cohort ID	***Patients count***	***Male***	***Female***	***Unknown***	***Hispanic or Latino***	***Not Hispanic or Latino***	***Unknown***	***White***	***Black***	***Asian***	***Native Hawaiian***	***Multiple races***	***Unknown***
A	124,985	64,953 (52%)	60,020 (48%)	<20 (0%)	22,425 (18%)	97,227 (78%)	5,333 (4%)	73,539 (59%)	22,081 (18%)	4,729 (4%)	311 (0%)	130 (0%)	24,195 (19%)
		10,544** (52%)	9,709 (48%)	<20 (0%)	3,751 (19%)	15,645 (77%)	858 (4%)	11,628 (57%)	3,651 (18%)	740 (4%)	43 (0%)	26 (0%)	4,166 (21%)
B	303,418	155,354 (51%)	147,800 (49%)	264 (0%)	38,549 (13%)	244,810 (81%)	20,059 (7%)	191,356 (63%)	55,018 (18%)	8,714 (3%)	1,452 (0%)	3,276 (1%)	43,602 (14%)
		23,374 (52%)	21,997 (48%)	<20 (0%)	5,811 (13%)	37,737 (83%)	1,827 (4%)	28,479 (63%)	8,888 (20%)	1,120 (2%)	68 (0%)	612 (1%)	6,208 (14%)
C	133,992	69,114 (52%)	64,866 (48%)	<20 (0%)	24,425 (18%)	103,891 (78%)	5,676 (4%)	78,357 (58%)	23,324 (17%)	5,584 (4%)	329 (0%)	137 (0%)	26,261 (20%)
		11,082 (52%)	10,439 (48%)	<20 (0%)	4,018 (19%)	16,604 (77%)	900 (4%)	12,362 (57%)	3,851 (18%)	782 (4%)	44 (0%)	27 (0%)	4,456 (21%)
D	328,726	167,259 (51%)	161,188 (49%)	279 (0%)	42,364 (13%)	264,819 (81%)	21,543 (7%)	207,695 (63%)	58,792 (18%)	9,869 (3%)	1,508 (0%)	3,603 (1%)	47,259 (14%)
		25,017 (51%)	23,864 (49%)	<20 (0%)	6,241 (13%)	40,676 (83%)	1,968 (4%)	30,924 (63%)	9,362 (19%)	1,211 (2%)	69 (0%)	692 (1%)	6,627 (14%)
B-C	25,014	12,762 (51%)	12,249 (49%)	<20 (0%)	2,036 (8%)	22,013 (88%)	965 (4%)	16,798 (67%)	5,227 (21%)	379 (2%)	24 (0%)	586 (2%)	2,000 (8%)
B-A	24,984	12,744 (51%)	12,237 (49%)	<20 (0%)	2,032 (8%)	21,987 (88%)	965 (4%)	16,786 (67%)	5,222 (21%)	373 (1%)	24 (0%)	586 (2%)	1993 (8%)
D-A	28,297	14,291 (51%)	14,003 (49%)	<20 (0%)	2,445 (9%)	24,751 (87%)	1,101 (4%)	19,100 (68%)	5,650 (20%)	460 (2%)	25 (0%)	666 (2%)	2396 (8%)
D-C	27.220	13,846 (51%)	13,371 (49%)	<20 (0%)	2,195 (8%)	23,961 (88%)	1,064 (4%)	18,494 (68%)	5,494 (20%)	421 (2%)	24 (0%)	665 (2%)	2,122 (8%)
D-B	3,020	1,385 (46%)	1,635 (54%)	<20 (0%)	390 (13%)	2,502 (83%)	128 (4%)	2,113 (70%)	374 (12%)	82 (3%)	<20 (0%)	65 (2%)	385 (13%)

* All P-values < 0.001 using Pearson's Chi-squared test

** Each cell shaded in grey, contains information from the test dataset

number of patients across all categories of ethnicity, which is consistent with it being the largest cohort. Cohort A has the lowest counts of Hispanic/Latino and NOT Hispanic/Latino patients and the lowest count of patients with unknown ethnicity.

Understanding and addressing potential selection bias based on gender, race, and ethnicity is crucial for developing ML-based models within the N3C. Adjusting for these demographic factors using appropriate techniques is essential to ensure that predictive models are not biased and can be effectively generalized across diverse cohorts. The analysis

indicated that the cohorts are statistically different ($p<0.001$) in the composition of age, gender, race, ethnicity, and length of stay, as tested with the Mann-Whitney *U* Test and Chi-squared test. The different parts of the cohorts are of most importance, denoted as B-C, D-A, etc. These patients are present in larger cohorts but not in smaller ones, and ones for which the generalizability of models needs to be carefully assessed. Table 1 presents the numbers of patients in these sets split by gender, race, and ethnicity.

For each cohort, the results focus on the measurements taken within the first 48 hours of hospital admission, categorized by the survival status, which indicates whether the patient survived or died. In Cohort A, 86.95% of patients survived, while 13.05% did not. Cohort B shows a slightly different outcome, with 84.5% of patients surviving and 15.5% not surviving. Cohort C's results are like Cohort A's, with 87.12% of patients alive and 12.88% having expired. Lastly, Cohort D's data reveals that 84.77% of patients survived, and 15.23% did not. These figures provide insights into the survival rates of patients within each cohort during the acute phase of their hospital stay.

Machine Learning Models

Model Construction Methods

Standard supervised machine models were trained on the training set, with the predicted outcome being survival at the end of hospitalization. Random Forests (RF) [12], Gradient Boosting (GB) [12], and L2-regularized Logistic Regression (LR) [13] models were trained. RF, GB, and LR are all supervised learning algorithms for classification and regression tasks. RF and GB are ensemble methods that combine multiple models to improve performance, while LR uses regularization to prevent overfitting. RF is known for its robustness and ability to handle diverse datasets, GB for optimizing arbitrary loss functions, and LR for its regularization capabilities to handle multicollinearity and non-independent features.

This experiment used a holdout training set to construct three ML models - LR, RF, and GB- using their default parameters. The primary goal was to compare the performance of these models across different cohorts rather than training the best possible models. To account for potential variations in the data, the dataset was divided into four distinct cohorts: A, B, C, and D. GridSearchCV was then employed to optimize each cohort's hyperparameters of the LR, RF, and GB models separately. This approach allowed us to fine-tune the hyperparameters specifically for each cohort, potentially improving the overall performance and accuracy of the predictions while maintaining a consistent baseline for comparison across the four cohorts.

For Cohort A, the Logistic Regression model was configured with a maximum iteration of 1000 and balanced class weights. The Random Forest Classifier utilized 1000 estimators, a maximum depth of 12, and balanced class weights. The Gradient Boosting Classifier had 150 estimators, a learning rate of 0.05, and a maximum depth of 5. In Cohort B, the Logistic Regression model maintained the same settings as in Cohort A. The Random Forest Classifier retained the same configuration as well. However, the Gradient Boosting Classifier was slightly modified, with 160 estimators and a maximum depth of 6, while keeping the learning rate at 0.05. For Cohort C, the Logistic Regression and Random Forest Classifier models were identical to those used in Cohort A. The Gradient Boosting Classifier had 150 estimators, a learning rate of 0.05, and a maximum depth of 7, varying from the previous cohorts. Lastly, in Cohort D, the Logistic Regression and Random Forest Classifier models remained unchanged from Cohorts A, B, and C. The Gradient Boosting Classifier had 160 estimators, a learning rate of 0.07 (slightly higher than the other cohorts), and a maximum depth of 6. This approach allowed us to fine-tune the hyperparameters of the models for each cohort separately, potentially improving the overall performance and accuracy of the predictions across the four cohorts.

Model Evaluation Methods

The presented work followed the general evaluation framework described by Wojtusiak [14 15]. Evaluation metrics are crucial for assessing the performance of ML models. They provide insights into how well a model's predictions match the data. This study utilizes several metrics, including AUC (Area Under the Receiver-Operator Curve), Precision, Recall, and F1 Score, as well as measures such as model calibration. Assessing machine learning models across subpopulations delineated by gender, race, and ethnicity is imperative to test for fairness and reliability. Subpopulations include categories such as male and female for gender and white, black, Asian, Native Hawaiian, and multiple races for racial classification. Ethnicity is distinguished by Hispanic and non-Hispanic categories. Proper evaluation across these diverse groups is critical to mitigating biases and ensuring equitable outcomes. To ensure algorithmic fairness, we assess the performance of machine learning models to predict the outcome of survival at the end of hospitalization by evaluating the AUC. This evaluation allows us to audit fairness across the entire dataset and within subgroups.

Model Comparison

For each model— LR, RF, and GB—there is a standard train-test evaluation where the model is trained on a cohort's data (e.g., cohort A) and then tested on that same cohort's test data (A). This provides a baseline understanding of the model's performance on data similar to what it was trained on. However, the evaluation goes further by testing each model on data from other cohorts. For example, a model trained on cohort A's data is also tested on cohorts B, C, and D's test data.

This is done by creating specialized test datasets for each combination, ensuring that individuals from the training set are not included in the test sets to avoid data leakage and to ensure a fair assessment of the model's predictive power on entirely new data. This cross-comparison is repeated for all possible combinations, such as training on cohort B, testing on A, C, and D, and so on for the other cohorts.

Moreover, we have specialized test datasets. These specialized datasets are created by combining test data from different cohorts while ensuring no individual is included in more than one test dataset, leading to combinations such as B-A, B-C, D-A, D-B, and D-C. For the D-A test dataset, as an example, all individuals from cohort D's test data are included except for those who also appear in cohort A's test data.

Comparison of Feature Importance

In ML, identifying which input variables most significantly predict the outcome is crucial, and this is where feature importance becomes essential. Methods for determining feature importance include statistical validation, Shapley value analysis, and notably, the Random Forest Feature Importance (RFFI) method [16].

RFFI, embedded within the Random Forest algorithm, assesses feature significance by measuring each feature's contribution to reducing the model's impurity, with common metrics such as Gini impurity or mean squared error being used for this impurity-based importance [17]. In this study, we have implemented the RFFI method to evaluate the significance of features across cohorts A, B, C, and D.

In this method, we have observed that 'age' occupies a larger portion of the bar plot. Therefore, we have decided to exclude it from our calculation. Consequently, we implemented the RFFI without taking 'age' into consideration. Each value represents the calculated importance score for a specific feature within a cohort, with higher scores indicating a greater impact on the predictive model.

Results

Accuracy comparison

The accuracy comparison is centered around the AUC metric. This metric is used to evaluate the performance of classification models, providing an aggregate measure of their ability to classify positive and negative cases across different thresholds 18 correctly. Based on the AUC scores, which measure the model's ability to distinguish between classes, we determine the best model by comparing the AUC values across three models. At first, this comparison is conducted across all cohorts (A, B, C, and D) using the entire dataset (see Table 2).

Table 2 illustrates the consistency and variation in model performance across different cohort combinations, showing that models tend to perform similarly within the same cohort combinations (e.g., Train = A and Test = A vs. Train = C and Test = C).

In contrast, performance can vary when models are trained and tested on different cohorts. This evaluation helps assess the models' robustness across different data distributions.

The table effectively assesses how well the models generalize to unseen data from different cohorts by comparing AUC values across various cohort combinations. These results indicate that the similarity of the training and testing

Table 2. The AUC values for three models were calculated across all cohorts using the entire dataset

		Test								
		A	B	C	D	B-A	B-C	D-A	D-B	D-C
Train	A	0.75, 0.75, 0.76*	0.71, 0.71, 0.72	0.75, 0.75, 0.76	0.71, 0.71, 0.73	0.69, 0.68, 0.70		0.70, 0.69, 0.71		
	B	0.74, 0.74, 0.74	0.73, 0.75, 0.75	0.74, 0.74, 0.74	0.74, 0.75, 0.75	0.73, 0.74, 0.74	0.73, 0.74, 0.74		0.79, 0.78, 0.79	
	C	0.75, 0.75, 0.76	0.71, 0.71, 0.72	0.75, 0.76, 0.76	0.71, 0.71, 0.73		0.69, 0.68, 0.70			0.69, 0.68, 0.70
	D	0.74, 0.74, 0.73	0.73, 0.75, 0.75	0.74, 0.74, 0.73	0.74, 0.75, 0.75			0.73, 0.75, 0.75	0.79, 0.78, 0.79	0.73, 0.74, 0.75

*The sequence of numbers in each cell, listed from left to right, represents the AUC values for LR, RF, and GB models.

data affects the model's ability to generalize, with the best performance seen when both are closely aligned. This underscores the importance of training on diverse and representative data for better model generalization.

Table 3. The AUC values for three models were calculated across all cohorts using the gender dataset

		Test (Gender)								
		A	B	C	D	B-A	B-C	D-A	D-B	D-C
Train	A	0.74, 0.75, 0.76* 0.75, 0.76, 0.76	0.71, 0.70, 0.72 0.71, 0.71, 0.73	0.74, 0.75, 0.75 0.75, 0.76, 0.76	0.71, 0.70, 0.72 0.71, 0.71, 0.73	0.69, 0.67, 0.70 0.69, 0.69, 0.70		0.69, 0.67,0.70 0.70, 0.69, 0.71		
	B	0.74, 0.74, 0.74 0.74, 0.74, 0.74	0.73, 0.75, 0.75 0.73, 0.74, 0.75	0.73, 0.73, 0.73 0.74, 0.75, 0.75	0.73, 0.74, 0.75 0.74, 0.75, 0.75	0.73, 0.74, 0.74 0.72, 0.74, 0.74	0.73, 0.74, 0.74 0.72, 0.74, 0.74		0.73, 0.72, 0.74 0.83, 0.83, 0.83	
	C	0.74, 0.75, 0.76 0.75, 0.76, 0.76	0.71, 0.70, 0.72 0.71, 0.71, 0.72	0.74, 0.75, 0.75 0.75, 0.76, 0.76	0.71, 0.70, 0.72 0.71, 0.72, 0.73		0.69, 0.68, 0.70 0.68, 0.68, 0.70			0.69, 0.68, 0.70 0.69, 0.69, 0.70
	D	0.74, 0.74, 0.73 0.74, 0.74, 0.74	0.73, 0.75, 0.75 0.73, 0.74, 0.75	0.73, 0.73, 0.73 0.74, 0.75, 0.74	0.73, 0.75,0.75 0.74, 0.75, 0.75			0.73, 0.74, 0.75 0.73, 0.75, 0.75	0.73, 0.73, 0.74 0.83, 0.83, 0.83	0.73, 0.74, 0.75 0.73, 0.74, 0.75

* The sequence of numbers in each cell, listed from left to right, represents the AUC values for LR, RF, and GB models. Also, the sequence of three numbers represents first the Male and then the Female.

The performance analysis across gender subgroups reveals consistent results for the LR, RF, and GB models (Table 3). Minimal differences in AUC values are observed for both training and testing sets within each gender category (male and female), indicating stability in model performance across genders. These findings suggest that there is no significant gender bias present in the AUC metrics, highlighting the models' consistent performance regardless of gender distinctions.

Table 4 highlights the variability in model performance across different training and test cohorts, emphasizing the impact of cohort selection on model generalizability. The GB model generally performs better across different racial groups, indicating its effectiveness in distinguishing between cohorts. However, challenges arise in the Native Hawaiian group, suggesting potential overfitting during training and poor generalization ability. Overall, the table

Table 5. The AUC values for three models were calculated across all cohorts using the race dataset

		Test (Race)								
		A	B	C	D	B-A	B-C	D-A	D-B	D-C
Train	A	0.71, 0.72, 0.72* 0.74, 0.75, 0.76 0.76, 0.78, 0.77 0.86, 0.78, 0.82 0.88, 0.92, 0.88	0.69, 0.69, 0.70 0.70, 0.70, 0.72 0.76, 0.76, 0.77 0.75, 0.62, 0.67 0.72, 0.71, 0.72	0.71, 0.72, 0.72 0.74, 0.75, 0.75 0.76, 0.78, 0.77 0.86, 0.79, 0.82 0.88, 0.92, 0.88	0.69, 0.69, 0.71 0.70, 0.70, 0.72 0.76, 0.77, 0.77 0.75, 0.62, 0.67 0.72, 0.71, 0.72	0.68, 0.67, 0.69 0.68, 0.68, 0.70 0.76, 0.67, 0.78 0.48, 0.30, 0.42 0.71, 0.70, 0.72		0.69, 0.68, 0.69 0.69, 0.68, 0.70 0.77, 0.77, 0.79 0.48, 0.24, 0.42 0.72, 0.70, 0.71		
	B	0.70, 0.71, 0.71 0.73, 0.73, 0.73 0.76, 0.75, 0.73 0.86, 0.80, 0.78 0.88, 0.88, 0.92	0.72, 0.73, 0.73 0.72, 0.74, 0.74 0.78, 0.78, 0.78 0.75, 0.65, 0.64 0.75, 0.76, 0.75	0.70, 0.71, 0.71 0.73, 0.73, 0.73 0.76, 0.75, 0.73 0.86, 0.81, 0.78 0.88, 0.88, 0.92	0.72, 0.73, 0.73 0.72, 0.74, 0.74 0.78, 0.79, 0.79 0.75, 0.65, 0.64 0.75, 0.75, 0.75	0.72, 0.74, 0.74 0.71, 0.73, 0.73 0.81, 0.81, 0.81 0.41, 0.41, 0.41 0.74, 0.75, 0.75	0.72, 0.74, 0.74 0.71, 0.73, 0.73 0.81, 0.81, 0.81 0.41, 0.43, 0.45 0.74, 0.75, 0.75		0.75, 0.67, 0.75 0.77, 0.76, 0.78 0.89, 0.81, 0.90 Nan, Nan, Nan 0.84, 0.77, 0.76	
	C	0.71, 0.72, 0.72 0.74, 0.75, 0.76 0.76, 0.78, 0.78 0.86, 0.80, 0.79 0.88, 0.88, 0.88	0.69, 0.69, 0.70 0.70, 0.70, 0.72 0.76, 0.76, 0.76 0.75, 0.62, 0.67 0.71, 0.71, 0.71	0.71, 0.72, 0.72 0.74, 0.75, 0.76 0.76, 0.78, 0.78 0.86, 0.80, 0.79 0.88, 0.88, 0.88	0.69, 0.66, 0.70 0.70, 0.70, 0.72 0.76, 0.77, 0.76 0.75, 0.62, 0.67 0.72, 0.71, 0.70		0.68, 0.67, 0.69 0.68, 0.68, 0.70 0.76, 0.76, 0.75 0.45, 0.30, 0.51 0.71, 0.70, 0.71			0.68, 0.68, 0.69 0.68, 0.68, 0.70 0.77, 0.77, 0.77 0.45, 0.23, 0.51 0.72, 0.71, 0.70
	D	0.70, 0.71, 0.71 0.73, 0.73, 0.72 0.76,0.75,0.71 0.86, 0.81,0.77 0.88,0.88, 0.92	0.72, 0.73, 0.74 0.72, 0.74, 0.74 0.78, 0.78, 0.78 0.74, 0.65, 0.63 0.75, 0.76, 0.75	0.70, 0.71, 0.71 0.73, 0.73, 0.72 0.76, 0.75, 0.71 0.86, 0.82, 0.77 0.88, 0.88, 0.92	0.72,0.73, 0.74 0.72, 0.74, 0.74 0.78, 0.79, 0.79 0.75, 0.65, 0.64 0.75, 0.76, 0.75			0.73, 0.74, 0.75 0.72, 0.74, 0.74 0.82, 0.82, 0.81 0.41, 0.46, 0.43 0.74, 0.75, 0.75	0.75, 0.75, 0.75 0.78, 0.77, 0.78 0.89, 0.83, 0.91 Nan, Nan, Nan 0.84, 0.77, 0.74	0.73, 0.74, 0.75 0.72, 0.73, 0.74 0.82, 0.82, 0.81 0.41, 0.39, 0.43 0.74, 0.75, 0.75

* The sequence of numbers in each cell, listed from left to right, represents the AUC values for LR, RF, and GB models. Also, the sequence of three numbers represents first the Black and then the White, Asian, Native Hawaiian, Multi Race .

Table 4. The AUC values for three models were calculated across all cohorts using the ethnicity dataset

		Test (Ethnicity)								
		A	B	C	D	B-A	B-C	D-A	D-B	D-C
Train	A	0.77, 0.78, 0.78* 0.74, 0.74, 0.75	0.73, 0.71, 0.75 0.70, 0.70, 0.72	0.77, 0.78, 0.79 0.73, 0.74, 0.75	0.73, 0.71, 0.75 0.70, 0.70, 0.72	0.66, 0.59, 0.68 0.68, 0.65, 0.70		0.66, 0.60, 0.69 0.69, 0.65, 0.71		
	B	0.77, 0.78, 0.77 0.73, 0.73, 0.73	0.76, 0.76, 0.77 0.72, 0.72, 0.74	0.77, 0.78, 0.77 0.73, 0.73, 0.73	0.76, 0.76, 0.77 0.73, 0.74, 0.74	0.73, 0.74, 0.74 0.72, 0.70, 0.73	0.72, 0.73, 0.74 0.72, 0.70, 0.73		0.71, 0.71, 0.75 0.78, 0.75, 0.78	
	C	0.77, 0.78, 0.78 0.73, 0.74, 0.75	0.73, 0.71, 0.74 0.70, 0.70, 0.72	0.77, 0.78, 0.79 0.73, 0.74, 0.75	0.73, 0.71, 0.75 0.70, 0.70, 0.72		0.65, 0.59, 0.68 0.68, 0.65, 0.70			0.65, 0.59, 0.68 0.69, 0.65, 0.71
	D	0.77, 0.78, 0.78 0.73, 0.73, 0.72	0.76, 0.76 0.77 0.72, 0.74, 0.74	0.77, 0.78, 0.78 0.73, 0.73, 0.72	0.75, 0.76, 0.77 0.73, 0.74, 0.74			0.73, 0.74, 0.75 0.73, 0.70, 0.73	0.71, 0.71, 0.76 0.78, 0.75, 0.78	0.72, 0.74, 0.75 0.72, 0.70, 0.73

* The sequence of numbers in each cell, listed from left to right, represents the AUC values for LR, RF, and GB models. Also, the sequence of three numbers represents first the Hispanic and then the Not Hispanic.

underscores the importance of training on diverse and representative data for better model generalization and fairness across different racial groups.

Table 5 emphasizes training on diverse and representative data for better model generalization and fairness across different ethnic groups. The GB model consistently demonstrates superior performance across various ethnic groups, indicating its effectiveness in distinguishing between cohorts. However, it encounters challenges in the Native Hawaiian group, suggesting potential overfitting during training and a diminished ability to generalize effectively.

Comparison of Feature Importance

The RFFI analysis across cohorts A, B, C, and D, excluding 'age', shows 'ABG indices', 'Respiratory Rate', and 'Chloride' as top predictors in Cohort A, with 'Cardiac Troponin I', 'Heart Rate', and 'Creatinine' also notable. In Cohort B, 'ABG indices', 'Cardiac Troponin I', and 'Respiratory Rate' are most influential, alongside 'Not Hispanic or Latino', 'Hispanic or Latino', and 'FiO2'. Cohort C highlights 'ABG indices', 'Respiratory Rate', and 'Chloride', with 'Cardiac Troponin I', 'Heart Rate', and 'Creatinine' also significant. Cohort D emphasizes 'ABG indices', 'Cardiac Troponin I', and 'Not Hispanic or Latino', with 'Hispanic or Latino', 'CBC with PLT', and 'MSRMNTS_FiO2' also important. This analysis underscores the need for cohort-specific feature consideration in healthcare ML models, with 'ABG indices', 'Respiratory Rate', and 'Cardiac Troponin I' consistently important, guiding feature selection and model optimization. The analysis also reveals the importance of cohort-specific features, such as ethnicity indicators in Cohorts B and D, suggesting that tailored models considering these unique characteristics could improve predictive accuracy. This insight underscores the need for a nuanced approach to feature selection and model development, considering both universally important features and those specific to individual cohorts.

Correlation between models trained on the same cohorts

We conducted correlation analyses using both Pearson's and Spearman's methods to assess the performance of three different ML models across multiple cohorts. Each cohort was provided with its own training and testing datasets, and we examined the correlations between these datasets when the models were trained and tested under various combinations. You can observe Pearson's and Spearman's correlation results in Table 6, illustrating the correlations between LR and RF, LR and GB, and RF and GB. Table 6 highlights a strong positive correlation between different ML models, particularly between LR and GB, across all cohorts. This indicates a high degree of agreement in their predictions. However, some variability in correlation strength between other model pairs, such as LR-RF and RF-GB, suggests nuances in data processing. Spearman's correlation consistently yields higher scores than Pearson's, indicating its effectiveness in capturing the relationship between model predictions.

Table 6. Correlation Coefficient between two models on the same cohorts

		Test								
		A	B	C	D	B-A	B-C	D-A	D-B	D-C
Train	A	0.90, 0.88, 0.87* 0.91, 0.95, 0.95	0.87, 0.88, 0.86 0.87, 0.93, 0.90	0.90, 0.88, 0.87 0.91, 0.96, 0.95	0.87, 0.88, 0.86 0.87, 0.93, 0.90	0.78, 0.91, 0.78 0.86, 0.92, 0.88		0.77, 0.91, 0.77 0.86, 0.92, 0.88		
	B	0.90, 0.87, 0.86 0.89, 0.93, 0.93	0.91, 0.90, 0.89 0.92, 0.95, 0.96	0.90, 0.87, 0.86 0.90, 0.93, 0.94	0.91, 0.90, 0.89 0.92, 0.93, 0.96	0.86, 0.93, 0.82 0.91, 0.95, 0.96	0.91, 0.90, 0.89 0.91, 0.95, 0.96		0.86, 0.93, 0.83 0.93, 0.96, 0.96	
	C	0.91, 0.84, 0.85 0.91, 0.93, 0.95	0.88, 0.81, 0.82 0.88, 0.88, 0.89	0.91, 0.84, 0.85 0.91, 0.93, 0.95	0.88, 0.81, 0.82 0.88, 0.88, 0.89		0.86, 0.78, 0.77 0.86, 0.84, 0.82			0.77, 0.91, 0.78 0.87, 0.85, 0.84
	D	0.90, 0.84, 0.85 0.89, 0.91, 0.93	0.91, 0.88, 0.89 0.92, 0.94, 0.96	0.90, 0.84, 0.85 0.90, 0.91, 0.93	0.92, 0.88, 0.89 0.92, 0.95, 0.96			0.86, 0.93, 0.83 0.92, 0.94, 0.96	0.87, 0.93, 0.83 0.94, 0.96, 0.97	0.86, 0.93, 0.83 0.92, 0.94, 0.96

* The sequence of numbers in each cell, listed from left to right, represents the correlations between LR and RF, LR and GB, and RF and GB. Also, the sequence of three numbers represents first the Pearson's and then the Spearman's correlation.

Table 7. Correlation Coefficient between two models on the different cohorts

		Test								
Train	Train	A	B	C	D	B-A	B-C	D-A	D-B	D-C
A	B	0.98, 0.95, 0.87 0.98,0.95,0.94	0.90, 0.86, 0.87 0.90,0.85,0.91	0.98, 0.95, 0.87 0.98,0.95,0.94	0.90, 0.86, 0.87 0.90,0.85,0.91	0.85, 0.79, 0.83 0.85,0.79,0.87				
	C	1.00, 1.00, 0.97 1.00, 1.00, 0.98	1.00, 1.00, 0.94 1.00, 1.00, 0.97	1.00, 1.00, 0.97 1.00, 1.00, 0.98	1.00, 1.00, 0.94 1.00, 1.00, 0.97					
	D	0.98, 0.95, 0.83 0.98,0.95,0.93	0.90, 0.86, 0.86 0.90,0.86,0.91	0.98, 0.95, 0.84 0.98,0.95,0.93	0.90, 0.86, 0.86 0.90,0.86, 0.91			0.86, 0.80, 0.82 0.86,0.80,0.87		
B	C	0.98, 0.95, 0.84 0.98,0.95,0.92	0.89, 0.86, 0.83 0.89,0.85,0.89	0.98, 0.95, 0.84 0.98,0.95,0.93	0.89, 0.86, 0.83 0.89,0.85,0.89		0.84, 0.79, 0.78 0.84,0.79, 0.86		1.00, 1.00, 0.99 1.00, 1.00, 0.99	
	D	1.00, 1.00, 0.98 1.00, 1.00, 0.99	1.00, 1.00, 0.99 1.00, 1.00, 0.99	1.00, 1.00, 0.98 1.00, 1.00, 0.99	1.00, 1.00, 0.99 1.00, 1.00, 0.99					
C	D	0.98, 0.95, 0.81 0.98,0.95,0.91	0.90, 0.86, 0.82 0.90,0.86,0.89	0.98, 0.95, 0.82 0.98,0.96,0.91	0.89, 0.86, 0.82 0.89,0.86,0.89					0.84, 0.80, 0.77 0.84,0.79,0.85

* The sequence of numbers in each cell, listed from left to right, represents the correlations between LR and LR, RF and RF, and GB and GB on the different cohorts. Also, the sequence of three numbers represents first the Pearson's and then the Spearman's correlation.

Correlation between models trained on the different cohorts

Table 7 underscores the strong positive correlations between ML models trained on different cohorts, indicating a high degree of agreement in their predictions. This is particularly evident in the high correlation coefficients for the LR model, which demonstrates consistent correlations across different cohorts. Despite the overall strong positive correlation, there is some variability in the strength of the correlation between different pairs of models. For instance, the RF model shows moderate variability, while the Gradient Boosting GB model exhibits significant fluctuations when generalizing across cohorts. Table 7 includes Pearson's and Spearman's correlation coefficients, with Spearman's correlation generally yielding higher scores. This indicates that Spearman's correlation captures the relationship between the models' predictions more effectively, underscoring the robust correlation between different ML models across different cohorts. This emphasizes the importance of considering different correlation measures to understand the relationship between models and their generalization ability fully.

Model Calibration

Model calibration indicates how far "scores" outputted from an ML model are from the actual probabilities. As one can see in Figure 3, Gradient Boost is the best calibrated model, yet it does not output any scores above about 0.7. Logistic Regression and Random Forest models significantly overestimate predictions. More importantly, the calibration worsens when models are applied to data from outside the cohort. This is the most prominent model trained on Cohort A, which is applied to test data from Cohort D (right side of the figure).

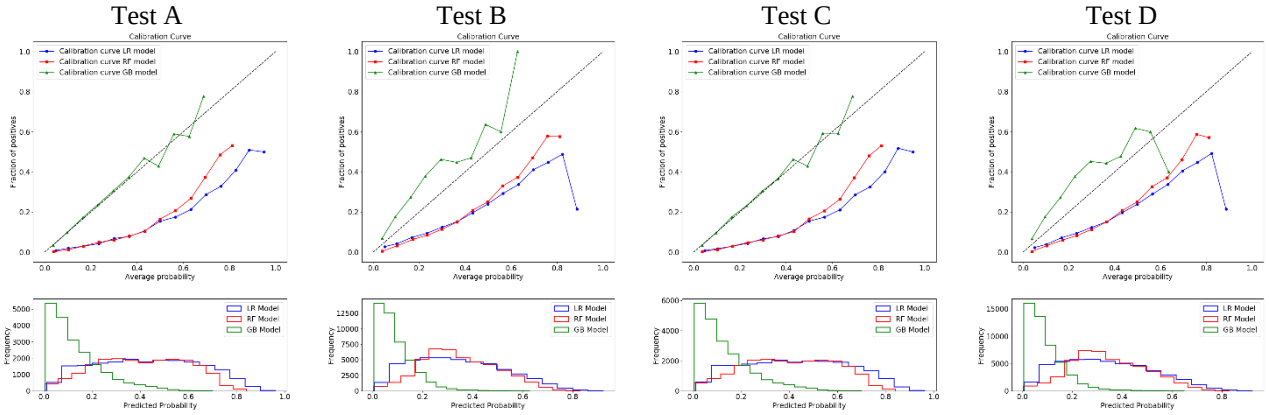

Figure 3: Example calibration plots for models trained on Cohort A and tested on Cohorts A, B, C and D.

Conclusion

This study meticulously underscores the profound influence that decisions made during data preprocessing exert on the composition and scale of cohorts. Such decisions can introduce selection bias, potentially compromising the integrity of research findings. The inclination towards constructing more exclusive cohorts necessitates a judicious approach, as such exclusions could diminish the representativeness of the data and raise concerns regarding the fairness of the outcomes. Moreover, comparing feature importance across cohorts illuminates the nuanced interplay between input variables and predictive outcomes, emphasizing the need for tailored approaches to model development that account for cohort-specific characteristics. Furthermore, correlation analyses between models trained on the same and different cohorts reveal varying degrees of agreement, underscoring the complexity of model generalization and the importance of cohort-specific considerations. The research presented herein highlights the imperative for ongoing inquiry into the ramifications of preprocessing choices on biases within ML models, advocating for the development of bespoke strategies to mitigate these biases. Our team's research endeavors are dedicated to evaluating the impact of these preprocessing decisions on the quality of ML models derived from the data. Our current findings definitively establish that the cohorts differ, and these differences significantly affect ML outcomes. Future endeavors will concentrate on domains where biases manifest more prominently, scrutinizing the effects of diverse preprocessing methodologies on the generalizability of research. These efforts are anticipated to yield critical insights that will fortify the external validity of studies. Adopting a discerning and informed methodology in data preprocessing is paramount to enhancing the reliability and equity of ML applications across a spectrum of fields. Social determinants of health, such as Gender, Race, and Ethnicity, have been shown to play a significant role in COVID-19 outcomes, further emphasizing the need for careful cohort selection and bias mitigation in ML models to ensure fair and accurate predictions.

Discussion

The study highlights the critical importance of thoughtful cohort selection in health data processing, emphasizing that arbitrary decisions can introduce biases affecting the representativeness, fairness, and generalizability of ML models in healthcare. Standardized and transparent selection criteria are crucial to ensure data representativeness, avoiding exclusion of important subpopulations or introducing noise. Cross-cohort validation is needed to ensure model robustness and reliability across different clinical settings. The insight that certain features are consistently important across cohorts while others vary can guide feature selection and model optimization. Implementing rigorous data preprocessing protocols, including sensitivity analyses and considering ensemble methods or meta-learning, can improve generalization. Ongoing inquiry into the ramifications of preprocessing choices on biases within ML models is essential, particularly in domains where biases are prominent. Given the significant role of social determinants of health in COVID-19 outcomes, it is imperative to consider these factors in ML model development to ensure fair and accurate predictions.

It is important to acknowledge certain limitations within the scope of this study. Primarily, the analysis is confined to data from August 1, 2020, to December 31, 2021, which may not encompass subsequent evolutions in COVID-19 healthcare practices, data documentation, or patient demographics. Additionally, not all preprocessing decisions outlined in the decision tree have been thoroughly examined, i.e., the selection of patients aged 18 and over remains a point of contention, as alternative age thresholds such as 21 could be equally justifiable.

References

1. Heckman, J. J. Selection Bias. in *Encyclopedia of Social Measurement* (ed. Kempf-Leonard, K.) 463–468 (Elsevier, New York, 2005). doi:10.1016/B0-12-369398-5/00115-8.
2. Heckman, J. Varieties of Selection Bias. *The American Economic Review* 80, 313–318 (1990).
3. Lu, H., Cole, S. R., Howe, C. J. & Westreich, D. Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology* 33, 699–706 (2022).
4. Nakamura, K. & Kawabata, H. I Choose, Therefore I Like: Preference for Faces Induced by Arbitrary Choice. *PLoS One* 8, e72071 (2013).
5. Quick Safety Issue 23: Implicit bias in health care | The Joint Commission. https://www.jointcommission.org/resources/news-and-multimedia/newsletters/newsletters/quick-safety/quick-safety-issue-23-implicit-bias-in-health-care/.
6. Gopal, D. P., Chetty, U., O'Donnell, P., Gajria, C. & Blackadder-Weinstein, J. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthc J* 8, 40–48 (2021).
7. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Mitigating bias in machine learning for medicine. *Commun Med* 1, 1–3 (2021).
8. Suri, J. S. *et al.* Understanding the bias in machine learning systems for cardiovascular disease risk assessment: The first of its kind review. *Comput Biol Med* 142, 105204 (2022).
9. Mac Namee, B., Cunningham, P., Byrne, S. & Corrigan, O. I. The problem of bias in training data in regression problems in medical decision support. *Artif Intell Med* 24, 51–70 (2002).
10. N3C - Home. https://covid.cd2h.org/.
11. Haghighathoseini, A., Qodrati, M., Min, H., Leslie, T.F., Frankenfeld, C.L., Menon, N.M. and Wojtusiak, J. Selection Bias from Data Processing in N3C. *Proceedings of the IEEE 12th International Conference on Healthcare Informatics (ICHI), Orlando, FL, 2024*.
12. Cha, G.-W., Moon, H.-J. & Kim, Y.-C. Comparison of Random Forest and Gradient Boosting Machine Models for Predicting Demolition Waste Based on Small Datasets and Categorical Variables. *Int J Environ Res Public Health* 18, 8530 (2021).
13. Nusinovici, S. *et al.* Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology* 122, 56–69 (2020).
14. Domanski, P. A., Steven Brown, J., Heo, J., Wojtusiak, J. & McLinden, M. O. A thermodynamic analysis of refrigerants: Performance limits of the vapor compression cycle. *International Journal of Refrigeration* 38, 71–79 (2014).
15. Irvin, K. & Wojtusiak, J. Comparison of Classification Learning Methods for Medical Claims Payments. in *AMIA* (2012).
16. Lee, Y. & Seo, J. Suggestion of statistical validation on feature importance of machine learning. *Annu Int Conf IEEE Eng Med Biol Soc* 2023, 1–4 (2023).
17. Algehyne, E. A., Jibril, M. L., Algehainy, N. A., Alamri, O. A. & Alzahrani, A. K. Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia. *BDCC* 6, 13 (2022).
18. Ling, C. X., Huang, J. & Zhang, H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. in *Advances in Artificial Intelligence* (eds. Xiang, Y. & Chaib-draa, B.) 329–341 (Springer, Berlin, Heidelberg, 2003). doi:10.1007/3-540-44886-1_25.

Acknowledgement

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS Contract No. 75N95023D00001, Axle Informatics Subcontract: NCATS-P00438-B. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource [https://doi.org/10.1093/jamia/ocaa196]