EXPLAINABLE MACHINE LEARNING IDENTIFICATION OF BIOLOGICAL PROFILES ASSOCIATED WITH POLIOVIRUS NEUTRALIZING ANTIBODY SERONEGATIVITY: A POPULATION-BASED NHANES STUDY

[1]Biplab Paul, [2]Suman Guchait, [3]Trisha Mondal, [4]Kaushik Bhattacharyya, [5]Ruchita Sen
[1]Academic Researcher, [2]Academic Researcher, [3]Academic Researcher, [4]Academic Researcher, [5]Academic Researcher
[1]University of Kalyani, [2]Maulana Abul Kalam Azad University of Technology, [3]Maulana Abul Kalam Azad University of Technology, [4]University of Kalyani, [5]University of Calcutta, Kolkata, India

Abstract: Poliovirus immunity can differ across population groups and serotypes. Serological evidence can reveal immunity gaps which are not shown by vaccination coverage alone. This study examined demographic and biological factors connected with poliovirus seronegativity. The study aimed to identify factors associated with poliovirus neutralising-antibody seronegativity. It also assessed whether biological variables improved prediction beyond demographic factors. A cross-sectional method is conducted using NHANES 2017- March 2020 data. Five datasets are linked using the participant sequence number and the final analysis included 5859 participants aged 6-59 years old. Type 3 seronegativity is the primary basic outcome. Type 1, Type 2 and any-serotype seronegativity are secondary outcomes. Demographic, anthropometric, haematological and inflammatory variables are examined. Survey-weighted logistic regression derived adjusted associations. Explainable machine-learning models compared demographic, biological and combined predictors. Type 3 showed the highest seronegativity among individual serotypes. Older age is the clearest independent factor associated with Type 3 seronegativity. Differences are also observed for BMI, neutrophils, platelets, haemoglobin, hs-CRP and PLR. Most biological measures are not independently associated after adjustment. The combined machine-learning model achieved the strongest discrimination. It also showed the highest sensitivity for identifying seronegative participants. However, specificity and overall accuracy remained limited. Age is the dominant predictor in the decision-tree analysis. PLR shows a smaller additional contribution. Poliovirus seronegativity remains an important population immunity concern. Age appears to be the most consistent associated factor. Biological variables provide some additional predictive information when combined with demographic characteristics. However, their overall predictive contribution remains modest. Explainable machine learning can help identify population profiles linked with seronegativity. Further longitudinal studies are needed to examine antibody persistence and changes over time.

Keywords: seronegativity, serosurveillance, Machine learning, poliovirus

1. INTRODUCTION

1.1 Background

Poliovirus continues to be a significant concern for disease prevention on a global level, since its transmission could be possible whenever there are gaps in immunity. Vaccination using oral poliovirus vaccine (OPV) and inactivated poliovirus vaccine (IPV) has helped reduce poliomyelitis, but the disease needs constant immunity for its eradication. Thus, serological surveillance continues to be important because vaccination coverage does not always show whether individuals have developed antibodies. It was shown in a systematic review of studies from Italy that seroprevalence of poliovirus neutralizing antibodies varies among different studies and serotypes, so serosurveillance must continue in countries that have been without polio for a long time (Tafuri *et al.*, 2023). Serosurveys are also done to address such issues in areas that are prone to outbreaks to identify any remaining susceptibility and organize vaccination activities (Halbrook *et al.*, 2024).

Neutralizing antibodies are well-known indicators of humoral protection against poliovirus infections. Neutralizing antibodies can interfere with viral binding and entry into the cell and therefore are used as indicators of the presence of antibodies capable of preventing the infection. Measurement of neutralizing antibodies in all poliovirus types including 1, 2 and 3, is necessary since immunity can be different among serotypes. Recent research in Japan showed that it is possible to measure neutralizing antibodies in a high throughput way and that the results correlate very well with the conventional neutralization tests for all three serotypes

(Arita and Iwai-Itamochi, 2022). Therefore, the use of large serological data is feasible in studying the differences in poliovirus immunity.

Data from recent population-based surveys also show that seronegativity can persist even after vaccination efforts. In Liberia, after two campaigns using new oral poliovirus vaccine type 2, serological protection in the population had increased but not fully (Kennedy *et al*., 2023). Also, in the Democratic Republic of Congo, significant portions of young children did not have neutralizing titres that are protective against each of the three serotypes, despite the ongoing vaccination activities and outreach programs (Halbrook *et al*., 2024). Thus, biological susceptibility may persist despite active vaccination programs.

Poliovirus immunity depends also on the interplay between systemic and mucosal responses. Neutralizing antibodies in serum are particularly important indicators of protection from paralytic disease, whereas mucosal immunity can affect the replication in intestine and virus shedding. Serum neutralizing antibodies measurement is therefore just one way of estimating population immunity. The present research uses neutralizing-antibody seronegativity as an indicator that can be studied together with demographic and biological factors. This makes the basis of analysing whether routine laboratory measurements are able to predict seronegativity of poliovirus neutralizing antibodies.

1.2 Research Problem

Although considerable advances have been made in vaccinating and eradicating poliovirus, the existence of seronegativity and lack of detectable neutralizing antibodies has been observed in vaccinated or partially vaccinated people. Such information is important, as even a small number of seronegative individuals becomes crucial in cases where poliovirus is being introduced or where vaccine strains of the virus circulate. Serosurveys conducted recently show that the protection from the virus is not complete for all serotypes and demographic groups, proving the disjunction between coverage and immunity (Kennedy *et al*., 2023; Halbrook *et al*., 2024). Moreover, the problem is relevant for type-specific immunity as well, as a person may be positive for one type of poliovirus but seronegative against another one.

The reasons for seronegativity are likely to be complex, as age may indicate different vaccine schedule, historical exposure and time since vaccination, and sex and socioeconomic status may also play a role. The results of the studies in the field of serology show that antibody levels may vary depending on the age group and may persist even after vaccination (Arita and Iwai-Itamochi, 2022). Other biological features, such as nutrition, inflammation, and variability in blood-cell profile may also be relevant for immune competence. However, most of the studies on poliovirus serology focused on the history of vaccination, exposure to campaign, geography and prevalence of seropositivity, and did not consider the combination of biological features. Thus, it is unknown whether routinely collected measures, such as white blood cell level, neutrophils, lymphocytes, platelets, haemoglobin level, body mass index, and C-reactive protein, would be able to differentiate seronegative people. Another limitation of the traditional analyses of relationships in question is the possible nonlinearity of relationship and interplay of several factors. This creates a need for an approach that can examine multiple characteristics together while still producing interpretable findings.

1.3 Machine Learning in Population Immunology

Machine learning is frequently used in the field of immunology due to the existence of various factors that affect immune outcomes in complex and nonlinear ways. Instead of assuming that there is a stable relationship between one predictor and outcome, machine learning can identify combinations of characteristics that improve the classification and prediction process. It is relevant for population immunology where various demographic, anthropometric and laboratory measures may jointly reflect immune response variability. Current vaccine studies provide an example of application of machine learning in the area. Wistuba-Hamprecht *et al*. (2024) employed supervised machine-learning methods to analyse the patterns of antibodies in relation to the efficacy of malaria vaccines. In turn, Montesi *et al*. (2024) used machine learning to predict humoral responses to SARS-CoV-2 vaccine and compared multiple algorithms in terms of identifying informative predictors. Machine learning may be useful in situations where it is impossible to represent antibody response using a few predictors.

When using machine learning for biomedical research, interpretability of results becomes especially crucial. A model may achieve excellent predictive performance without providing sufficient information about the reasons for different classification of certain individuals. Explainable approaches solve the problem by demonstrating the main variables that influence predictions and, in some cases, whether their role is to increase or decrease risk. For example, Borole and Rajan (2024) have demonstrated the usefulness of interpretable explanation in the context of immune response prediction, whereas Boushehri *et al*. (2023) developed an explainable machine-learning framework to study immunological and antibody-related features. At the same time, Gardiner *et al*. (2022) demonstrated that explainable models can use demographic and biological information to determine key features in heterogeneous immune-related data.

For population studies, the combination of prediction and interpretability of results is relevant since identification of profiles is more important than the accuracy itself. Decision trees allow finding combinations of predictors, and feature importance can identify those characteristics that contribute to classification the most. At the same time, the usage of machine learning alongside regression can provide complementary results. Hence, the present study uses explainable machine learning to see whether demographic, anthropometric, haematological and inflammatory characteristics are combined to form patterns of poliovirus neutralising-antibody seronegativity.

1.4 Research Gap and Novelty

There exist several knowledge gaps where machine learning could intersect with poliovirus serology, population health data and seronegativity in particular. Firstly, recent research into poliovirus primarily revolves around calculating seroprevalence, analysing

effectiveness of vaccination campaigns, exploring serotype-specific immunity and evaluating risk of outbreaks (Halbrook *et al*., 2024; Kennedy *et al*., 2023; Tafuri *et al*., 2023). These are essential studies from an epidemiological point of view, yet they do not show if common biological variables can distinguish those who lack neutralising-antibody protection. Secondly, machine learning has increasingly been applied to analyse immune response, vaccine effectiveness and antibody outcomes, but the focus was usually limited to a particular disease, molecular markers or clinical samples. Thus, Wistuba-Hamprecht *et al*. (2024) used antibody profiling to forecast malaria vaccine efficacy, whereas Montesi *et al*. (2024) employed clinical and immunological markers to predict humoral immune response to SARS-CoV-2 vaccination. Cuperlovic-Culf *et al*. (2024) further demonstrated that machine-learning algorithms and multivariate analysis are capable of identifying correlations between demographic markers, especially age and sex, and variation in antibody response. Still, none of these applications examined poliovirus neutralising-antibody seronegativity using a wide population survey. Thirdly, explainable machine learning is underused in population-level immunological research. There already are examples of how interpretable techniques can identify relevant biological characteristics and add transparency to complex predictive models (Borole and Rajan, 2024; Boushehri *et al*., 2023). Moreover, research using routine haematological biomarkers shows that explainable methods can be helpful in identifying informative blood-based predictors even when no black box approach is involved. Nevertheless, this methodology has rarely been used in regard to poliovirus serology outcomes.

In this respect, the present study tries to bridge the existing gaps by using newly available poliovirus antibody NHANES 2017-March 2020 data and correlating serological outcomes with demographic, anthropometric, haematological and inflammatory markers. The central novelty of this study lies in the explicit comparison of demographic-only, biological-only and combined predictive models which allows testing whether routine biological information provides any predictive power beyond basic demographic data. At the same time, explainable modelling strategy allows for the identification of the most influential variables, instead of just measuring prediction performance. The novelty of this study does not lie in the use of machine learning or serological surveillance, but in their combination for a different population and outcome context. In addition, by focusing on type 3 seronegativity as the main outcome and expanding the scope to types 1 and 2 and any-serotype seronegativity, the study takes into account serotype specificity of poliovirus immunity. The findings may help generate hypotheses about biological profiles linked to absent neutralising antibodies while remaining appropriately cautious about causality.

1.5 Aim, Objectives and Research Questions

Research Aim

To identify demographic and biological profiles associated with poliovirus neutralizing-antibody seronegativity using explainable machine-learning techniques in the NHANES 2017-March 2020 population.

Research Objectives

- **RO1:** To examine the prevalence and demographic and biological characteristics associated with poliovirus neutralizing-antibody seronegativity among NHANES participants.
- **RO2:** To determine whether hematological, inflammatory, anthropometric, and demographic variables can predict poliovirus neutralizing-antibody seronegativity using interpretable machine-learning models.
- **RO3:** To compare the predictive performance of demographic-only, biological-only, and combined models and identify the most important predictors of poliovirus seronegativity.

Research Questions

- **RQ1:** What demographic and biological characteristics are associated with poliovirus neutralizing-antibody seronegativity among NHANES participants?
- **RQ2:** To what extent can demographic, haematological, inflammatory, and anthropometric variables predict poliovirus neutralizing-antibody seronegativity using explainable machine learning?
- **RQ3:** Do biological variables improve the prediction of poliovirus neutralizing-antibody seronegativity beyond demographic variables, and which predictors are most important?

2. MATERIALS AND METHODS

2.1 Study Design

In this study, cross-sectional secondary data is analysed from the National Health and Nutrition Examination Survey (NHANES) (Akinbani *et al*.2022). Datasets cover the time period from 2017 March to 2020 pre-pandemic cycle. The study is based on newly accessible poliovirus serological information together with demographic, anthropometric, haematological and inflammatory measurements to identify features associated with poliovirus neutralizing antibody seronegativity. The end-to-end analytical framework aggregated traditional statistical methods with explainable machine learning techniques (Yu *et al*.2022). This design is outlined as the research questions require both identification of variables associated with seronegativity and assessment of whether biological characteristics improve prediction beyond demographic information. The NHANES dataset is population based, whereas the cross-sectional design allows examination of relationships between derived biological characteristics and antibody status at the time of assessment. The study was therefore designed to identify predictive and associative patterns rather than establish temporal relationships or causality.

2.2 Study Population

The study population considered NHANES participants aged from 6 years to 59 years who were eligible for poliovirus antibody assessment during the time period 2017 to 2020(Cdc,2017). This time is considered a pre-pandemic survey period. The initial

analytical population is defined using participants with valid poliovirus neutralizing antibody measurements in the P_SSPOLI dataset and relevant information for the demographic and biological variables included in the analysis. The study systematically follows this inclusion criteria. On the other hand, participants are excluded where the primary poliovirus outcome was not available. The study does not include the population where the information required for the relevant analysis could not be obtained from the linked NHANES datasets. It also excludes participants without a valid survey weight from analyses requiring population representative inference. The initial P_SSPOLI dataset contains 6707 participants with eligible poliovirus antibody results. Hence, the final analytical sample is smaller following linkage with the CBC, hs-CRP, BMI and demographic datasets and assessment of missing observations (Hemade *et al*.2025).

2.3 Data Sources

Five NHANES data components are combined in the study for the analysis. P_SSPOLI served quantitative and interpreted neutralizing antibody measurements for poliovirus serotypes 1,2 and 3 and the specific surplus specimen survey weight (WTSSMMPP) (Alleman *et al*.2023). P_CBC provided complete blood count and separate measurements including white blood cell (WBC), neutrophil, lymphocyte, monocyte and platelet counts, haemoglobin and red cell distribution width (RDW). P_HSCRP provided high sensitivity C-reactive protein (hs-CRP), while P_BMX supplied body measurement information including body mass index (BMI). P_DEMO highlighted demographic characteristics and survey design variables including age, sex, information (Olson *et al*.2023).

2.4 Outcome Variables

The key result is the presence or absence of the neutralizing antibody to poliovirus type 3 based on the CDC interpretation variable SSPV3I. Individuals were classified into seropositive and seronegative categories using the CDC interpretation, where individuals with neutralizing-antibody titre below the specified cut off were considered seronegative. The original interpretation variable was retained and the analytical outcome was encoded with a binary outcome coding, such that 1 represented seropositive and 0 represented seronegative. Type 3 was selected as the key outcome as it exhibited the highest percentage of seronegatives of all three serotypes, resulting in a better distribution of events for modeling. Secondary outcomes comprised type 1 and type 2 seronegativity and an any-serotype seronegativity measure, defined as seronegativity for at least one of the three serotypes. A supplementary 0–3 seronegativity score was also considered, representing the number of serotypes for which an individual was seronegative.

2.5 Predictor Variables

In this study, predictor variables are organised into demographic and biological domains. Demographic variables included age, sex, race and family income-to-poverty ratio. Biological predictors included BMI, WBC, neutrophil count, lymphocyte count, monocyte count, platelet count, haemoglobin, RDW and hs-CRP. There are derived three inflammatory indices such as neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR) and monocyte-to-lymphocyte ratio (MLR) (Buonacera *et al*.2022). This separation of predictors enabled direct comparison of demographic-only and biological-only information and assessment of whether biological variables improved prediction when added to demographic characteristics (Zinellu *et al*.2024).

2.6 Data Preparation

The five NHANES datasets were imported and merged using SEQN as the participant identifier. Following merging, duplicate identifiers, variable coding and the completeness of the linked records were checked at the time of data preparation (Chen *et al*.2023). Missing observations were quantified for all study variables, and participants lacking essential outcome information were excluded from the relevant analyses. Biological measurements were not automatically replaced with mean values. The original categorical serological interpretation variables were recorded into binary analytical outcomes while preserving the original NHANES variables. NLR, PLR and MLR were calculated from the corresponding absolute blood-cell counts, with invalid or missing denominators treated as missing rather than generating artificial ratio values.

2.7 Statistical Analysis

In this research concerning explainable machine learning identification of biological profiles associated with poliovirus seronegativity, descriptive analyses are conducted. The motive is to summarise participant characteristics and the prevalence of poliovirus seronegativity. Categorical variables were examined using Chi-square tests, while continuous variables were compared between seropositive and seronegative groups using independent samples t tests when distributional assumptions were appropriate or Mann Whitney U tests for non-normally distributed variables (Wuensch *et al*.2025). Multivariable survey weighted logistic regression was used to identify independent predictors of type 3 seronegativity. Adjusted odds ratios (ORs), 95% confidence intervals (CIs) and p-values were reported. NHANES survey design was incorporated using WTSSMMPP together with the relevant strata and primary sampling unit variables to support population-representative statistical inference.

2.8 Machine Learning Analysis

In this study, Explainable machine-learning models were developed to predict type 3 seronegativity. The dataset was divided into approximately 70% training and 30% testing subsets, with attention to maintaining the distribution of the outcome between subsets. Three predictor configurations were evaluated. These are sequentially Model A, containing demographic variables, Model B, containing biological variables and Model C, containing demographic and biological variables together (Zhang *et al*.2025). The study focused on CHAID and CRT decision tree approaches as they provide interpretable predictor structures and identifiable combinations of characteristics (Yang *et al*.2023). Predictor importance and the resulting tree structures were examined to identify variables contributing most strongly to classification. Variable selection or model learning was restricted to the training data to reduce the risk of information leakage.

2.9 Model Evaluation

Predictive performance was assessed using the testing data. The principal evaluation measures were receiver operating characteristic area under the curve (ROC-AUC), accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Performance was compared through the demographic only, biological only and combined models to determine whether incorporation of biological information improved classification beyond demographic characteristics (Oehr *et al*.2025). Sensitivity analyses were accordingly planned for poliovirus serotypes 1 and 2 and for any serotype seronegativity to examine the consistency of the principal findings across alternative outcome definitions.

3. RESULTS

3.1 Participants Characteristics

The eligible dataset initially included 6707 participants with valid Type 3 serology, where the final analytical dataset contains 5859 participants. Thus, 848 participants are excluded from the complete case analysis. It is observed that most missing information came from the income-to-poverty ratio. It is missing for 11.48% of eligible participants. BMI was missing for 0.95%, while missingness in blood measures is minimal. The mean age of participants is 30.28 years. The median age is 29 years and mean BMI is 27.39 kg/m^2. Average of White Blood Cells (WBC) is 7.22 × 10^3 cells/µL. Mean neutrophils, lymphocytes and monocytes are 4.06, 2.35 and 0.56 × 103 cells/µL. Mean platelet count is 263.59 × 103 cells/µL , while haemoglobin and Red Blood Cell Distribution Width (RDW) are 13.87 g/dL and 13.67% respectively. Median High Sensitive C Reactive Protein (hs-CRP) is 1.27 mg/L. Median NLR, MLR and PLR are 1.68, 0.24 and 113.93. So, the participant profile presented substantial variation in age and biological measures (Xanthopoulos *et al*.2022). This variation supported further comparison of serostatus groups.

Variables	**Value**	**Variables**	**Value**
Demographic Characteristics		**Anthropometric**	
Age, years	30.28 ± 16.46	BMI, kg/m^2	27.39 ± 8.29
Sex, %	Male 49.6, Female 50.4	Income-to-poverty ratio	2.46 ± 1.66
Race/ethnicity, %	Mexican American 11.1; Other Hispanic 8.0, NH White 57.9, NH Black 12.1, NH Asian 5.6, Other/Multiracial 5.3		
Haematological measures			
WBC, × 10^3 cells/µL	7.22 ± 2.16	Neutrophils, × 10^3 cells/µL	4.06 ± 1.71
Lymphocytes × 10^3 cells/µL	2.35 ± 0.77	Monocytes, × 10^3 cells/µL	0.56 ± 0.18
Platelets, × 103 cells/µL	263.59 ± 65.64	Haemoglobin, g/dL	13.87 ± 1.48
RDW, %	13.67 ± 1.24	hs-CRP, mg/L	3.29 ± 7.38
Derived Inflammatory Indices			
NLR	1.88 ± 1.13	MLR	0.257 ± 0.116
PLR	121.81 ± 47.69		

Table 3.1: Demographic and Biological Characteristics of the study population

3.2 Prevalence of Poliovirus Seronegativity

From the inferential analysis, it is shown that 10.8% of the population is seronegative for Type 1 and the confidence interval is 95%, runs from 9.65% to 12.05%. Similarly, 7.18% of the population is seronegative for Type 2 and the range lies from 6.27% to 8.21%. Surprisingly, Type 3 seronegativity is higher at 21.02% maintaining the range for parameter from 19.19% to 22.98%. Lastly, Any-serotype is 28.39% which is dominating and the confidence interval is 95% for all the prevalence. The parameter range runs from 26.41 to 30.46% for Any-serotype. Hence, Type 3 showed the greatest seronegativity among individual serotypes. The larger any-

serotype estimation suggests that a considerable proportion lacked detectable immunity to at least one serotype. This pattern aligns the use of Type 3 as the primary outcome (Greenland *et al.*2022). The importance of examining serotype specific immunity is explored by these analyses.

Poliovirus outcome	Seronegative n*	Weighted prevalence (%)	95% CI
Type 1	631	10.80	9.65–12.05
Type 2	418	7.18	6.27–8.21
Type 3	1,114	21.02	19.19–22.98
Any serotype	1,575	28.39	26.41–30.46

Table 3.2: Weighted Prevalence of Poliovirus Seronegativity

3.3 Biological Differences by Serostatus

Several Characteristics distinguished between Seropositive and Seronegative participants. Age showed the clearest difference where Seronegative participants were a median age of 34.5 years. It is compared with 26 years among seropositive participants where p value seemed to be less than 0.01. This identifies a strong age gradient in Type 3 serostatus. Body Mass Index (BMI) is a bit higher among seronegative participants, median BMI is 28.0 versus 25.7 kg/m^2 where p value is less than 0.01. These findings interpret that body composition can be relevant to serostatus. Neutrophils also differed but the difference is small. Median values are 3.9 and 3.8 × 103 cells/µL, in this comparison, p value is 0.027. In this study, platelet related analysis also followed and platelet count is lower among seronegative participants. Median values are found to be 248 and 259 × 103 cells/µL and p value is lower than threshold (0.01). Haemoglobin is higher in the seronegative group as median values are 14.1 and 13.8 g/dL. It is inspected that hs-CRP is also higher among seronegative participants. Its median is 1.59 versus 1.20 mg/L and p value is less than level of significance (Kotronoulas *et al.*2023).

Consequently, age and BMI showed the strongest group differences. Some blood and inflammatory calculations also differed. Apart from that, these unadjusted findings do not construct independent associations.

Variable	Seropositive, median (IQR)	Seronegative, median (IQR)	p-value
BMI, kg/m²	25.7 (21.0–31.5)	**28.0 (23.6–33.3)**	**<0.001**
WBC, ×10³ cells/µL	6.9 (5.7–8.4)	7.0 (5.8–8.5)	0.144
Neutrophils, ×10³ cells/µL	3.8 (2.9–4.9)	3.9 (2.9–5.0)	0.027
Lymphocytes, ×10³ cells/µL	2.3 (1.8–2.8)	2.3 (1.9–2.7)	0.811
Monocytes, ×10³ cells/µL	0.5 (0.4–0.7)	0.5 (0.4–0.6)	0.614
Platelets, ×10³ cells/µL	259 (220–304)	**248 (212–292)**	**<0.001**
Haemoglobin, g/dL	13.8 (12.9–14.8)	**14.1 (13.0–15.1)**	**<0.001**
RDW, %	13.4 (13.0–14.0)	13.4 (12.9–14.0)	0.611
hs-CRP, mg/L	1.20 (0.47–3.54)	**1.59 (0.64–4.03)**	**<0.001**
NLR	1.67 (1.22–2.25)	1.71 (1.28–2.26)	0.071
MLR	0.235 (0.187–0.300)	0.235 (0.185–0.296)	0.419
PLR	114.58 (91.95–143.33)	110.61 (88.45–140.36)	0.003

Table 3.3: Biological Characteristics by Type 3 Serostatus

4.4 Multivariable Logistic Regression

In this research, logistic regression is implemented which is survey weighted including 5859 participants. The overall model is statistically significant where metrics such as adjusted F score is 69.40 and p value less than 0.001 (Koch *et al.*2022). In this analysis, age remained positively associated with Type 3 seronegativity after adjustment. The adjusted odd ratio is 1.015 where the value

range lies from 1.006 to 1.024 holding 95% confidence interval. Thus, each additional year of age corresponded to a small increase in the odds of seronegativity. It is found that the second sex category had lower odds than the reference category. Its adjusted odd ratio is 0.690 and the true population parameters run from 0.502 to 0.948 and corresponding p value is 0.0301. BMI portrays a positive association but it is borderline where the odd ratio (OR) is 1.016 and the range is from 0.999 to 1.032. The OR tells us how the odds of being seronegative change when a predictor increases or changes category, while the other variables in the model are held constant (Higuichi *et al.*2026). The remaining demographic, haematological and inflammatory measures are not statistically significant. These findings suggest that age is the clearest independent predictor. They also suggest that several biological differences inspected in past analysis are reduced after adjustment.

Predictor	**Adjusted OR**	**95% CI**	**p-value**
Age	**1.015**	1.006–1.024	**0.009**
Female vs male	**0.690**	0.502–0.948	**0.030**
Mexican American vs reference	1.332	0.952–1.864	0.080
Other Hispanic vs reference	0.928	0.598–1.440	0.679
Non-Hispanic Black vs reference	0.739	0.498–1.097	0.106
Non-Hispanic Asian vs reference	1.333	0.907–1.960	0.113
Other race/multiracial vs reference	0.918	0.520–1.622	0.716
Income-to-poverty ratio	0.955	0.884–1.033	0.192
BMI	1.016	0.999–1.032	0.060
WBC	1.160	0.590–2.282	0.596
Neutrophils	0.753	0.358–1.584	0.371
Lymphocytes	0.819	0.437–1.537	0.452
Monocytes	1.749	0.180–16.970	0.555
Platelets	1.000	0.995–1.005	0.965
Haemoglobin	0.938	0.812–1.084	0.309
RDW	1.034	0.876–1.221	0.622
hs-CRP	1.007	0.988–1.026	0.387
NLR	1.241	0.840–1.834	0.214
MLR	0.071	0.002–2.632	0.119
PLR	0.997	0.989–1.005	0.396

Table 4.4: Multivariable Predictors of Type 3 Seronegativity

4.5 Machine Learning Results

In this study, Machine Learning is used to evaluate the performance in predicting poliovirus serotypes based on unseen data of centers of disease control (CDC). Three models are developed using demographic, biological and combined predictors. The demographic model and the biological model achieved accuracy of 78.45%. Where the combined model achieved 78.1% overall accuracy. The initial decision tree models produced accuracy values near 78% but their sensitivity is lower than expectation. The demographic and biological models identified no seronegative participants at the default threshold. The combined model detected 11.8%. This result shows why accuracy is not only the answer or metrics, the outcome is imbalance toward seropositivity. As a result, the high initial accuracy precisely reflected correct classification of the majority group. To improve the performance and generalise the model, cross-validation and threshold optimisation is brought in the field. The combined CRT reached an ROC-AUC of 0.623 (Mosquera *et al.*2024). Its sensitivity increased to 80.9% where specificity is 33.3%, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are 25% and 86.4% respectively. Age variables dominated the combined CRT as its tree importance score is 0.374. Platelet-to-lymphocyte ratio (PLR) followed at 0.231, while RDW developed 0.036. Permutation

importance showed the same pattern as age reduced ROC-AUC by 0.148 when permuted. PLR produced 0.0048 percent reduction, while other variables's importance is little in this model (Lu *et al*.2022).

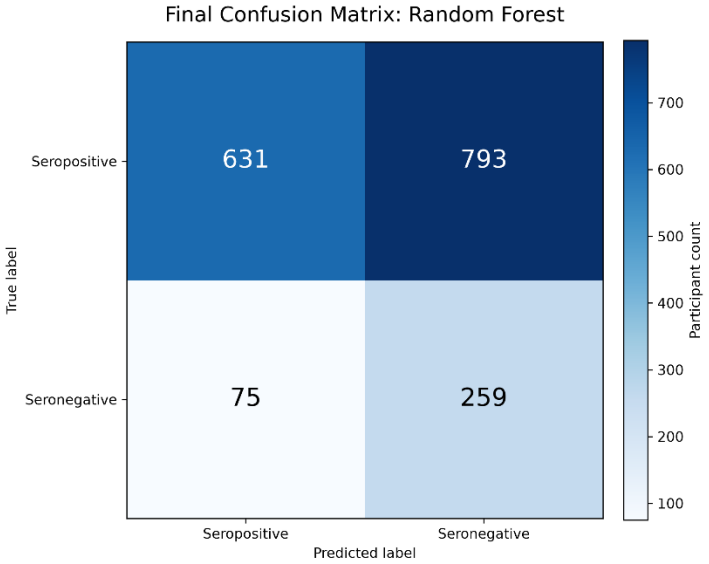

Figure 4.5: Confusion Matrix

(Source: Python)

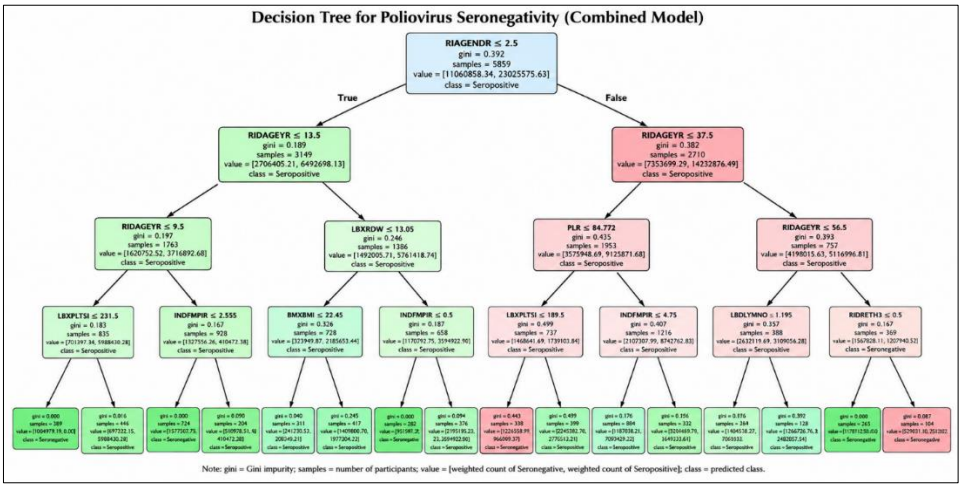

Figure 4.6: CRT Decision Tree for Poliovirus Seronegativity (Combined Model)

(Source: Self-Developed)

4.6 Model Comparison

For comparative analysis and parallel learning, Random Forest and tuning are focused (Gemechu *et al*.2026). The tuned demographic model achieved an ROC-AUC score of 0.593 and sensitivity is 65.2%, specificity is 49.9% and accuracy is 78.68%. The random forest classifier has a ROC-AUC score of 0.685 and sensitivity is 69.2% and accuracy of 78.82%. The biological model aligned perfectly with the tuned CRT and random forest. Consequently, discrimination surpassing demographics has been improved by adding biological predictors. The improvement is valuable mainly for identifying seronegative cases. It is not associated with strong overall classification performance. The ROC curves support this conclusion that the combined model remained above the other models across much of the curve. All three curves are relatively close to the diagonal line.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**Accuracy**	**Interpretation**
Demographic	**0.593**	**65.2%**	**49.9%**	**53.2%**	Modest discrimination with balanced classification
Biological	**0.566**	**69.2%**	**39.7%**	**46.0%**	Lower discrimination, but better detection of seronegative cases

Combined	**0.623**	**80.9%**	**33.3%**	**43.5%**	Best discrimination and sensitivity, but lowest specificity and accuracy

Table 4.6: Comparison of Predictive Performance by Type 3 Poliovirus Seronegativity

4.7 Sensitivity Analysis

In this study, sensitivity analysis is focused on precision scores that means among all the people who are actually seronegative, how many did the model correctly identify (Michaud *et al.*2023). Type 1, Type 2 and any serotype seronegativity outcomes are successfully derived. Their weighted prevalence estimates are also calculated. In addition, 'Age' is identified as the most consistent predictor of Type 3 seronegativity through the primary analysis. Biological measures contributed additional information, especially within combined tree models. Despite modest prediction, the overall predictive model remains robust. The findings support an association-based interpretation rather than a causal explanation.

5. DISCUSSION

5.1 Principal Findings

In this study, the initial references are information of 5859 NHANES participants. Among the four classifications of seronegativity, Type 3 showed the highest seronegativity with weighted prevalence of 21.02%. Prevalence in this study indicates how the amount of seronegativities are distributed in the studied population. Any-serotype seronegativity reached 28.39% and older age participants are the most independent correlate of Type 3 seronegativity. Female participants presented lower adjusted odds than the reference group. BMI, platelets, neutrophils, haemoglobin, hs-CRP and PLR varied between serostatus groups. The combined ML model developed the highest ROC-AUC at 0.623, and it achieved highest sensitivity at 80.9%. These findings perfectly align with the proposed research objectives.

5.2 Biological Interpretation

From the biological perspective, it can be stated that Seronegative participants have higher BMI and hs-CRP than seropositive participants. They also have lower platelet counts and higher haemoglobin values. Consequently, neutrophil counts express a minor group difference. PLR is lower among seronegative participants and differences in metabolic or inflammatory states can be influenced by these patterns. Despite these dynamic performances, the observed differences are not consistent across all markers. WBC, lymphocytes, monocytes, RDW, NLR and MLR give evidence of no significant group differences. Biological markers can explain wider health variation. The variable 'Age' showed a stronger and more consistent relationship across the complete analyses. Age can detect differences in vaccination or natural exposure. These interpretations remain valid as the study is cross-sectional.

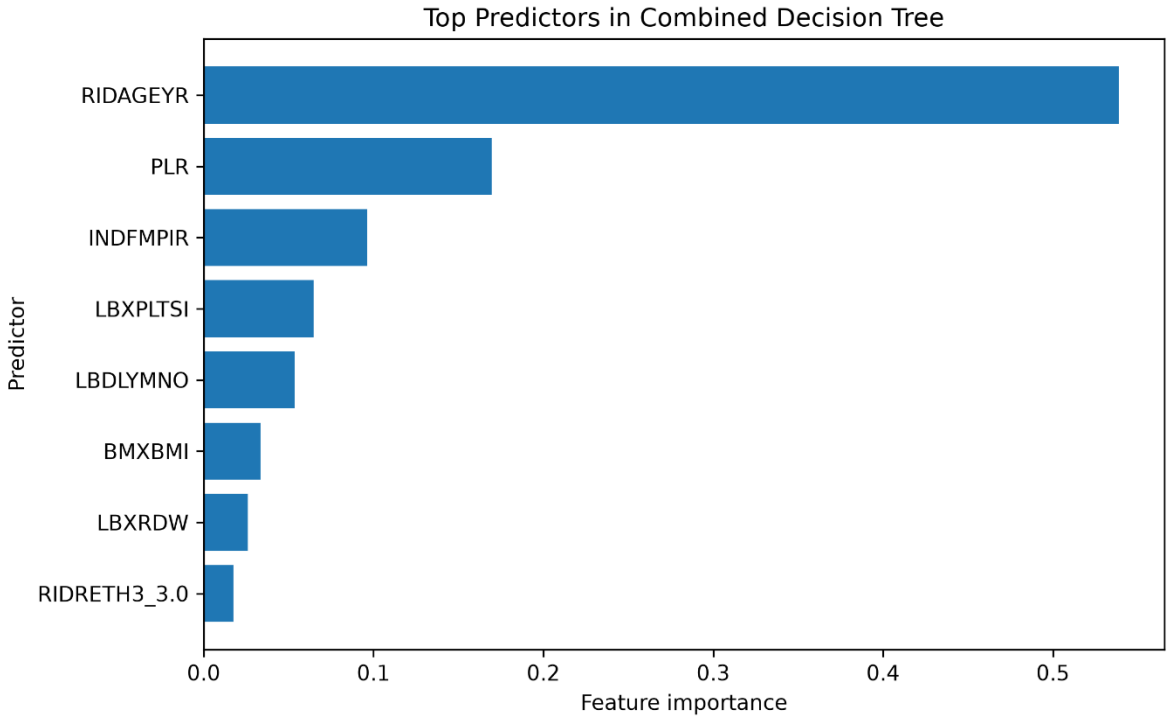

Figure 5.2: Top Predictors in Combined Decision Tree

(Source: Python)

5.3 Comparison with Previous Studies

In this study, the findings bring earlier evidence evaluating variances in poliovirus seroprotection. Comparative studies pay attention to Vaccine Derived Polioviruses (VDPV) strains in the adult population. It studied the level of poliovirus neutralizing antibodies in adults implementing a panel of immunologically diverse challenge strains of type 1 poliovirus. The motive is to know the factors of protection against VDPV strains (Kouiavskaia *et al.*2023). The study tested post-vaccination sera obtained from the clinical study objects to better explain the breadth of protection against polio, provided by Inactivated Poliovirus Vaccine (IPV). From the study

of Liu *et al*. (2022), it is stated that neutralizing antibody titer of a serum from an immunised patient is the most promising factor for assessing the proactive efficacy of polio vaccines, carrying new polio vaccines such as S-IPV.

5.4 Contribution of Machine Learning

A different view from traditional regression is obtained with Machine Learning. 'Age' is the most important factor in the combined decision tree and importance analysis. PLR showed a smaller contribution whereas other biological variables influenced little to overall discrimination. The combined model has higher ROC-AUC than the demographic model. Its AUC expanded in such a way from 0.593 to 0.623. Its sensitivity also developed from 65.2% to 80.9%. The improved random forest classifier achieved 78.82% accuracy and the combined tuned CRT model got 78.1 % for predicting the seronegativity types. These findings suggest several additional predictive values from biological information. Hence, the gain in the curve is modest and comes with more 'false-positive' classifications. So, biological variables should not be described as strong predictors. The metrics defining the variables appear greater for identifying possible seronegative profiles. This supports the study's attention on explainability instead of black box prediction.

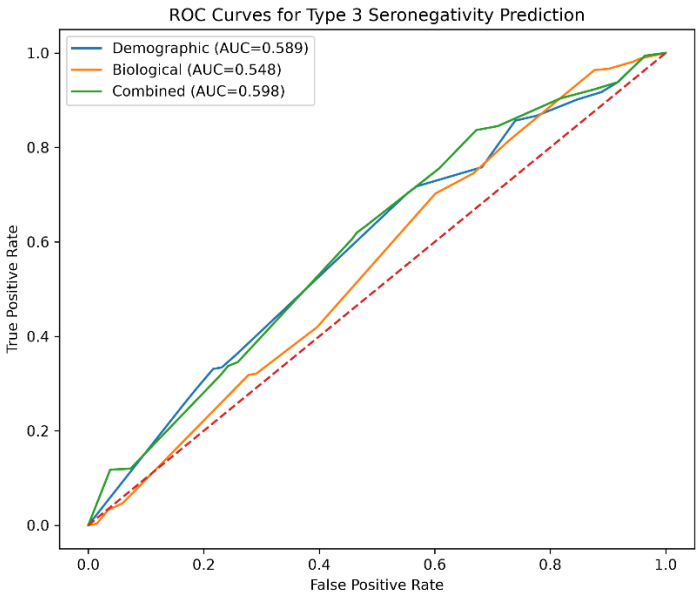

Figure 5.4: ROC Curves for Type 3 Seronegativity Prediction
(Source: Python)

5.5 Public Health Implications

The present study concerning 'Explainable machine learning identification of biological profiles associated with poliovirus neutralizing antibody seronegativity' has several implications in real-world cases. The findings have relevance for population immunity monitoring where a considerable minority presented no detectable Type 3 antibodies. It is recognised that more than one quarter are seronegative for at least one serotype. This indicates that vaccination coverage alone cannot explain complete immunity. Serological surveillance can distribute additional information about population sensitivity. This type of study is highly recommended for varying age groups to identify the pattern of polioviruses and the categorical antibodies and vaccines. Healthcare departments and Medical AI domains can refer to this study for tuning advanced models with image detection of patients and viruses (Yang *et al*.2025).

5.6 Strengths and Limitation

This study is effective and strong as it chooses NHANES dataset as a large population. It keeps notes on newly available poliovirus serological information. Several biological measures which are linked with poliovirus serotype serve a broad profile for analysis. Another major strength is combining regression with explainable machine learning (belle *et al*.2021). Despites these strengths, there are several limitations in the study. The cross-sectional design prevents temporal interpretation and observed associations cannot develop causality. This research does not directly measure longitudinal antibody decline. Besides that, vaccination history can be unavailable or incomplete for some participants. Complete case analysis can have brought selection bias. Vaccine-specific immune responses cannot be captured by routine biological markers. In addition, modest model performance creates boundaries in direct clinical prediction.

6. CONCLUSION

After analysing and researching about the biological profiles associated with poliovirus neutralizing antibody seronegativity, several key decisions can be made. It referred data and information from NHANES 2017 to 2020 pre pandemic data. Type 3 has the highest individual seronegativity prevalence. The findings show that poliovirus seronegativity remains an important population immunity issue. Seronegativity varied across the three poliovirus serotypes. The broader any-serotype measure also presented remaining immunity gaps. These findings support continuous attention to serotype-specific population immunity.

Age appeared as the most consistent factor across the analyses as Type 3 seronegativity got most frequency among older participants. Sex also interpreted an adjusted association with serostatus. Several biological measures differed between the two serostatus groups. These include BMI, neutrophils, platelets, haemoglobin, hs-CRP and PLR. Most biological measurements do not remain independently significant after adjustment. The ML workflow provided an additional perspective. The combined predictor model performed better than the separate models. Biological metrics added some predictive value beyond demographic information. In supervised learning such as decision trees, Age is the dominant predictor and Platelet-to-lymphocyte ratio (PLR) gives a smaller additional contribution. Different laboratory measures showed limited importance within the final predictive structure. These findings indicate that demographic characteristics can provide the strongest baseline information. Biological variables can perform classification tasks within particular participant profiles. In this study, healthcare professionals get facilities to understand population-level poliovirus immunity. It explores the value of merging traditional statistics with explainable ML. This method can recognise patterns that can be less visible through single-variable analysis. It can align future research and advanced analysis on groups with reduced detectable antibody protection. From a public health surveillance, the findings helped in serological disparities. Vaccine coverage individually cannot explain population immunity, so to recover that, serological information can help identify possible gaps in protection. Routine biological measures can assist future risk profiling but it should not replace direct antibody assessment or vaccination records. Several limitations should be checked with thorough interpretation and cross-sectional design. The study cannot directly measure changes in antibody levels over time. Complete vaccination histories are not always available for all participants. Missing information can have influenced the final analytical sample so further longitudinal studies should examine antibody persistence more directly. As a whole combination, the study identifies age as the strongest consistent predictor of seronegativity. It also justifies that biological information provides some additional predictive value. However, overall prediction remains limited and the findings are therefore best viewed as evidence of associated patterns. They also provide a basis for future studies of poliovirus immunity and antibody persistence.

REFERENCES

Akinbami, L., Chen, T.C., Davy, O., Ogden, C., Fink, S., Clark, J., Riddles, M. and Mohadjer, L., 2022. National Health and Nutrition Examination Survey, 2017–March 2020 prepandemic file: sample design, estimation, and analytic guidelines. https://dx.doi.org/10.15620/cdc:115434

Alleman, M.M., Jorba, J., Riziki, Y., Henderson, E., Mwehu, A., Seakamela, L., Howard, W., Mbule, A.K., Nsamba, R.N., Djawe, K. and Yapi, M.D., 2023. Vaccine-derived poliovirus serotype 2 outbreaks and response in the Democratic Republic of the Congo, 2017–2021. *Vaccine*, *41*, pp.A35-A47. https://doi.org/10.1016/j.vaccine.2023.02.042

Arita, M. and Iwai-Itamochi, M., 2022. High-throughput analysis of anti-poliovirus neutralization antibody titre in human serum by the pseudovirus neutralization test. Scientific Reports, [online] 12(1). https://doi.org/10.1038/s41598-022-20544-6.

Belle, V. and Papantonis, I., 2021. Principles and practice of explainable machine learning. *Frontiers in big Data*, *4*, p.688969. https://doi.org/10.3389/fdata.2021.688969

Borole, P. and Rajan, A., 2024. Building trust in deep learning-based immune response predictors with interpretable explanations. Communications Biology, [online] 7(1). https://doi.org/10.1038/s42003-024-05968-2.

Buonacera, A., Stancanelli, B., Colaci, M. and Malatino, L., 2022. Neutrophil to lymphocyte ratio: an emerging marker of the relationships between the immune system and diseases. *International journal of molecular sciences*, *23*(7), p.3636. https://doi.org/10.3390/ijms23073636

Cdc.gov. (2017). *P_sspoli*. [online] Available at: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_SSPOLI.htm [Accessed 3 Sept. 2026].

Chen, Y., Ding, Z., Alowain, L., Chen, X. and Wagner, D., 2023, October. Diversevul: A new vulnerable source code dataset for deep learning based vulnerability detection. In *Proceedings of the 26th international symposium on research in attacks, intrusions and defenses* (pp. 654-668). https://doi.org/10.1145/3607199.3607242

Cuperlovic-Culf, M., Bennett, S.A.L., Galipeau, Y., McCluskie, P.S., Arnold, C., Bagheri, S., Cooper, C.L., Langlois, M.-A., Fritz, J.H., Piccirillo, C.A. and Crawley, A.M., 2024. Multivariate analyses and machine learning link sex and age with antibody responses to SARS-CoV-2 and vaccination. iScience, [online] 27(8), p.110484. https://doi.org/10.1016/j.isci.2024.110484.

Gardiner, L.-J., Carrieri, A.P., Bingham, K., Macluskie, G., Bunton, D., McNeil, M. and Pyzer-Knapp, E.O., 2022. Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease. PLOS ONE, [online] 17(2), p.e0263248. https://doi.org/10.1371/journal.pone.0263248.

Gemechu, H., Biru, G., Gebremeskel, E., Teshome, F., Nemomssa, H.D., Dese, K., Wakjira, E., Tefera, M., Alayu, M., Ayana, G. and Kong, J., 2026. Machine learning framework for early detection of polio outbreaks from acute flaccid paralysis surveillance data. *Virology*, p.110863. https://doi.org/10.1016/j.virol.2026.110863

Greenland, S., Mansournia, M.A. and Joffe, M., 2022. To curb research misreporting, replace significance and confidence by compatibility: a Preventive Medicine Golden Jubilee article. *Preventive medicine*, *164*, p.107127. https://doi.org/10.1016/j.ypmed.2022.107127

Halbrook, M., Gadoth, A., Mukadi, P., Hoff, N.A., Musene, K., Dzogang, C., Sinai, C.S., Spencer, D., Ngoie-Mwamba, G., Tangney, S., Salet, F., Nyembwe, M., Kambamba Nzaji, M., Tambu, M., Mbala, P., Fuller, T., Gerber, S.K., Kaba, D., Muyembe-Tamfum, J.J. and Rimoin, A.W., 2024. Poliovirus-Neutralizing Antibody Seroprevalence and Vaccine Habits in a Vaccine-Derived Poliovirus Outbreak Region in the Democratic Republic of Congo in 2018: The Impact on the Global Eradication Initiative. Vaccines, [online] 12(3), p.246. https://doi.org/10.3390/vaccines12030246.

Hemade, A. and Salameh, P., 2025. Thrombopoiesis Meets Metabolism—But Not in the CBC: Survey-Weighted Analysis of Platelet Counts, hs-CRP, and Metabolic Risk. https://doi.org/10.21203/rs.3.rs-7925686/v1

Higuchi, T., Tanaka, E., Inoue, E., Sugano, E., Ikari, K. and Harigai, M., 2026. Increasing proportion and more favourable outcomes of seronegative rheumatoid arthritis: analysis of the IORRA cohort from 2000 to 2021. *Clinical Rheumatology*, pp.1-11. https://doi.org/10.1007/s10067-026-08139-2

Kennedy, S.B., Macklin, G.R., Mason Ross, G., Lopez Cavestany, R., Moukom, R.A., Jones, K.A.V., Mainou, B.A., Massaquoi, M.B.F., Kieh, M.W.S. and Mach, O., 2023. Poliovirus antibodies following two rounds of campaigns with a type 2 novel oral poliovirus vaccine in Liberia: a clustered, population-based seroprevalence survey. The Lancet Global Health, [online] 11(6), pp.e917–e923. https://doi.org/10.1016/s2214-109x(23)00116-x.

Koch, G., Casula, E.P., Bonnì, S., Borghi, I., Assogna, M., Minei, M., Pellicciari, M.C., Motta, C., D'Acunto, A., Porrazzini, F. and Maiella, M., 2022. Precuneus magnetic stimulation for Alzheimer's disease: a randomized, sham-controlled trial. *Brain*, *145*(11), pp.3776-3786. https://doi.org/10.1093/brain/awac285

Kotronoulas, G., Miguel, S., Dowling, M., Fernández-Ortega, P., Colomer-Lahiguera, S., Bağçivan, G., Pape, E., Drury, A., Semple, C., Dieperink, K.B. and Papadopoulou, C., 2023, April. An overview of the fundamentals of data management, analysis, and interpretation in quantitative research. In *Seminars in oncology nursing* (Vol. 39, No. 2, p. 151398). WB Saunders. https://doi.org/10.1016/j.soncn.2023.151398

Kouiavskaia, D., Mirochnitchenko, O., Troy, S. and Chumakov, K., 2023. Antigenic diversity of type 1 polioviruses and its implications for the efficacy of polio vaccines. *Vaccine*, *41*(13), pp.2147-2154. https://doi.org/10.1016/j.vaccine.2023.01.031

Liu, S., Lu, W., Ma, S., Guo, H., Zhang, Z. and Li, X., 2022. Comparison of the neutralizing activities of antibodies in clinical sera against both Sabin and wild-type polio pseudoviruses. *Journal of virological methods*, *300*, p.114376. https://doi.org/10.1016/j.jviromet.2021.114376

Lu, Y., Ye, T. and Zheng, J., 2022, August. Decision tree algorithm in machine learning. In *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)* (pp. 1014-1017). IEEE. https://doi.org/10.1109/AEECA55500.2022.9918857

Michaud, E.J., Liu, Z. and Tegmark, M., 2023. Precision machine learning. *Entropy*, *25*(1), p.175. https://doi.org/10.3390/e25010175

Montesi, G., Augello, M., Polvere, J., Marchetti, G., Medaglini, D. and Ciabattini, A., 2024. Predicting humoral responses to primary and booster SARS-CoV-2 mRNA vaccination in people living with HIV: a machine learning approach. Journal of Translational Medicine, [online] 22(1). https://doi.org/10.1186/s12967-024-05147-1.

Mosquera, C., Ferrer, L., Milone, D.H., Luna, D. and Ferrante, E., 2024. Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. *European Radiology*, *34*(12), pp.7895-7903. https://doi.org/10.1007/s00330-024-10834-0

Oehr, P., 2025. Interrelationships among sensitivity, precision, accuracy, specificity and predictive values in bioassays, represented as combined ROC curves with integrated cutoff distribution curves and novel index values. *Diagnostics*, *15*(4), p.410. https://doi.org/10.3390/diagnostics15040410

Olson, M.E., Hornick, M.G., Stefanski, A., Albanna, H.R., Gjoni, A., Hall, G.D., Hart, P.C., Rajab, I.M. and Potempa, L.A., 2023. A biofunctional review of C-reactive protein (CRP) as a mediator of inflammatory and immune responses: differentiating pentameric and modified CRP isoform effects. *Frontiers in immunology*, *14*, p.1264383. https://doi.org/10.3389/fimmu.2023.1264383

Shetab Boushehri, S., Essig, K., Chlis, N.-K., Herter, S., Bacac, M., Theis, F.J., Glasmacher, E., Marr, C. and Schmich, F., 2023. Explainable machine learning for profiling the immunological synapse and functional characterization of therapeutic antibodies. Nature Communications, [online] 14(1). https://doi.org/10.1038/s41467-023-43429-2.

Tafuri, S., Cuscianna, E. and Bianchi, F.P., 2023. Prevalence of poliovirus neutralizing antibodies in Italian population: A systematic review and meta-analysis. Vaccine, [online] 41(27), pp.4057–4063. https://doi.org/10.1016/j.vaccine.2023.04.047.

Wistuba-Hamprecht, J., Reuter, B., Fendel, R., Hoffman, S.L., Campo, J.J., Felgner, P.L., Kremsner, P.G., Mordmüller, B. and Pfeifer, N., 2024. Machine learning prediction of malaria vaccine efficacy based on antibody profiles. PLOS Computational Biology, [online] 20(6), p.e1012131. https://doi.org/10.1371/journal.pcbi.1012131.

Wuensch, K.L., 2025. Chi-square tests. In *International encyclopedia of statistical science* (pp. 466-468). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-69359-9_110

Xanthopoulos, A., Giamouzis, G., Dimos, A., Skoularigki, E., Starling, R.C., Skoularigis, J. and Triposkiadis, F., 2022. Red blood cell distribution width in heart failure: pathophysiology, prognostic role, controversies and dilemmas. *Journal of clinical medicine*, *11*(7), p.1951. https://doi.org/10.3390/jcm11071951

Yang, Q., Chen, J., Sun, Y., Wang, Y. and Tan, T., 2025. Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise. *Quantitative Imaging in Medicine and Surgery*, *15*(6), p.5450. https://doi.org/10.21037/qims-2024-2655

Yang, Y., Yi, F., Deng, C. and Sun, G., 2023. Performance analysis of the CHAID algorithm for accuracy. *Mathematics*, *11*(11), p.2558. https://doi.org/10.3390/math11112558

Yu, C.H.A., 2022. *Data mining and exploration: From traditional statistics to modern data science*. CRC Press. https://doi.org/10.1201/9781003153658

Zhang, Z., Meng, Q., Cui, Z., Yao, M., Shao, Z. and Tao, B., 2025. Machine learning applications in parallel robots: A brief review. *Machines*, *13*(7), p.565. https://doi.org/10.3390/machines13070565

Zinellu, A. and Mangoni, A.A., 2024. The association between the neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, and monocyte-to-lymphocyte ratio and systemic sclerosis and its complications: a systematic review and meta-analysis. *Frontiers in Immunology*, *15*, p.1395993. https://doi.org/10.3389/fimmu.2024.1395993