

RESEARCH

Open Access



# Predicting humoral responses to primary and booster SARS-CoV-2 mRNA vaccination in people living with HIV: a machine learning approach

Giorgio Montesi<sup>1†</sup>, Matteo Augello<sup>2†</sup>, Jacopo Polvere<sup>1</sup>, Giulia Marchetti<sup>2</sup>, Donata Medaglini<sup>1\*</sup> and Annalisa Ciabattini<sup>1\*</sup> 

## Abstract

**Background** SARS-CoV-2 mRNA vaccines are highly immunogenic in people living with HIV (PLWH) on effective antiretroviral therapy (ART). However, whether viro-immunologic parameters or other factors affect immune responses to vaccination is debated. This study aimed to develop a machine learning-based model able to predict the humoral response to mRNA vaccines in PLWH and to assess the impact of demographic and clinical variables on antibody production over time.

**Methods** Different machine learning algorithms have been compared in the setting of a longitudinal observational study involving 497 PLWH, after primary and booster SARS-CoV-2 mRNA vaccination. Both Generalized Linear Models and non-linear Models (Tree Regression and Random Forest) were trained and tested.

**Results** Non-linear algorithms showed better ability to predict vaccine-elicited humoral responses. The best-performing Random Forest model identified a few variables as more influential, within 39 clinical, demographic, and immunological factors. In particular, previous SARS-CoV-2 infection, BMI, CD4 T-cell count and CD4/CD8 ratio were positively associated with the primary cycle immunogenicity, yet their predictive value diminished with the administration of booster doses.

**Conclusions** In the present work we have built a non-linear Random Forest model capable of accurately predicting humoral responses to SARS-CoV-2 mRNA vaccination, and identifying relevant factors that influence the vaccine response in PLWH. In clinical contexts, the application of this model provides promising opportunities for predicting individual vaccine responses, thus facilitating the development of vaccination strategies tailored for PLWH.

**Keywords** Machine learning, SARS-CoV-2, HIV, Statistical modeling, Vaccines, mRNA, Antibodies, Immune response, ImmunoVirology

<sup>†</sup>Giorgio Montesi and Matteo Augello equally contributed.

\*Correspondence:

Donata Medaglini  
donata.medaglini@unisi.it  
Annalisa Ciabattini  
annalisa.ciabattini@unisi.it

Full list of author information is available at the end of the article



## Background

Since the beginning of the SARS-CoV-2 pandemic, people living with HIV (PLWH) have been considered at higher risk of serious illness and severe outcomes from COVID-19. Despite conflicting data emerged from preliminary analyses conducted in small cohorts [1, 2], subsequent larger observational studies confirmed that PLWH may suffer worse COVID-19 outcomes compared to the general population, especially in the presence of scarce immune reconstitution despite antiretroviral therapy (ART) and in case of unsuppressed HIV replication [3–8].

Owing to such vulnerability, PLWH were prioritized for SARS-CoV-2 vaccine administration since the early phases of the vaccination campaign. Research conducted to date agrees that, overall, PLWH mount immune responses to the primary cycle of SARS-CoV-2 vaccine which are comparable to those developed by HIV-negative people [3, 9]. However, when assessing HIV-specific factors typically related to adverse outcomes, such as low CD4 T-cell counts, inverted CD4/CD8 ratio, and uncontrolled HIV viremia, they invariably appeared associated to impaired cellular and humoral responses [3, 9–12], suggesting that PLWH with poor immune restoration and/or ongoing HIV replication should receive booster doses. An additional vaccine dose has been shown to substantially improve humoral responses in PLWH with hyporesponse after primary cycle [13–15]. However, whether HIV-related viro-immunological parameters or other factors may have an impact on immune responses to booster vaccination in PLWH is unclear [13, 16–18], yet it would be of utmost importance to personalize boosting strategies in the current phase of shifting from the pandemic to the endemic stage of COVID-19.

Generally, in biological contexts where regression analysis is required to study associations between variables, linear regression models alongside various feature selection strategies are commonly used [19, 20]. In recent years, advancements in machine learning strategies have enabled the quantification of both linear and non-linear associations in an unbiased manner and provided a comprehensive characterization of more intricate and complex interactions among predictor variables of a certain outcome [21]. Such approaches have been employed to identify key clinical factors associated with antibody responses and to predict vaccine immunogenicity in fragile and immunosuppressed populations such as organ transplant recipients [22, 23]. However, the utility of these algorithms in predicting immune responses to SARS-CoV-2 vaccines in PLWH has not been fully explored.

In the present study, we compared different machine learning algorithms in the setting of a large observational

study involving 497 PLWH after primary and booster SARS-CoV-2 mRNA vaccination, to develop a model able to accurately predict vaccine-elicited humoral immunity and identify relevant factors that influence vaccine response over time in this vulnerable population.

## Methods

### Study design

The San Paolo Infectious Diseases HIV-Vax (SPID-HIV-Vax) is a prospective observational study which was established in March 2021 that enrolled 800 PLWH who received the anti-SARS-CoV-2 Spikevax™ mRNA vaccine (Moderna) at the Clinic of Infectious Diseases and Tropical Medicine, San Paolo Hospital, ASST Santi Paolo e Carlo, Department of Health Sciences, University of Milan, Milan, Italy. Data concerning demographic characteristics, comorbidities, HIV-related features and self-reported previous SARS-CoV-2 infection were collected at enrollment using RedCap electronic data capture tools [24]. PLWH within SPID-HIV-Vax cohort were eligible to participate in the present study if they met the following criteria: availability of at least one post-vaccine anti-S IgG determination between March 2021 and January 2023, and availability of all baseline demographic and clinical variables included in the original database.

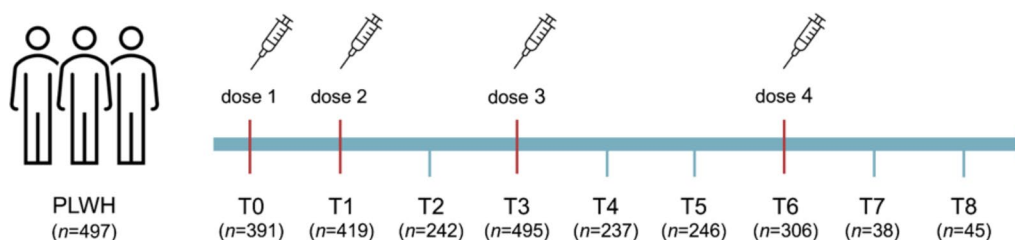
The study was approved by the local Ethical Committee and written informed consent was obtained from each participant.

### Biological samples collection and antibody quantification

From each participant, venous peripheral blood samples were collected at the following time points: day of first dose (T0); 1 month after first dose—coinciding with the day of second dose—(T1); 1 month after second dose (T2); 6 months after second dose—coinciding with third dose administration—(T3); 1 month after third dose (T4); 6 months after third dose (T5); 12 months after third dose—coinciding with fourth dose administration—(T6); 1 month after fourth dose (T7); 6 months after fourth dose (T8) (Fig. 1). Anti-trimeric Spike (S) IgG antibodies were quantitatively determined in serum samples by the LIAISON® SARS-CoV-2 TrimericS IgG assay (DiaSorin, Italy), and concentration expressed as binding antibody units per milliliter (BAU/mL).

### Statistical and machine learning models

To build a model capable of predicting antibody response to vaccination based on available demographic and clinical parameters both linear and non-linear regression methods were employed and compared. Variables were normalized using z-score transformation when numeric



**Fig. 1** Clinical study design. Blood samples were collected at each longitudinal time-point (T) from PLWH receiving multiple doses of mRNA SARS-CoV-2 vaccines. Number of available anti-S IgG samples are reported for each time-point. PLWH: people living with HIV; T0: day of the first dose; T1: 1 month after the first dose, coinciding with the day of the second dose; T2: 1 month after the second dose; T3: 6 months after the second dose, coinciding with the third dose administration; T4: 1 month after the third dose; T5: 6 months after the third dose; T6: 12 months after the third dose, coinciding with the fourth dose administration; T7: 1 month after the fourth dose; T8: 6 months after the fourth dose

and converted into dummy variables when categorical. The independent variables used as predictors encompassed demographic characteristics, comorbidities, viro-immunological HIV-related parameters, HIV epidemiology, and ART, totaling 39 variables. A temporal variable accounting for the days elapsed since the first vaccine dose administration was also included to build all algorithms. All models employed a train-and-test strategy, with the 80% of the dataset used as the training set, while the remaining 20% as the test set for evaluating model-performances. The metrics used to assess model quality were calculated as the mean values obtained through a fivefold cross-validation (CV): R-squared ( $R^2$ ) and Root Mean Squared Error (RMSE) for both linear and non-linear models. Only for linear models, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were computed during the training phase.

#### Generalized linear models

For the construction of linear models, the *glm* function from the *stats* package v3.6.2 via R statistical software v4.3.1 was employed. Feature selection was performed by Stepwise and Multi-Model inference to mitigate high dimensionality and multicollinearity issues.

Stepwise variable selection was executed employing both forward (SF) and backward (SB) approaches by means of the pre-built-in *stepAIC* function in R. Multi-model (MM) inference was conducted using the *glmulti* package v1.0.8 by generating all possible linear combinations of the independent variables and testing each model with the aim of AIC minimization. Due to the impracticality of testing all possible linear combinations with all the variables, dummy variables related to comorbidities (12 variables) were excluded. Thus, over 4 million combinations were tested, with each iteration fitting a *glm* via an exhaustive screening method.

#### Non-linear models

Tree Regression and Random Forest analyses were conducted using the *rpart* v4.1.23 and the *randomForest* packages v4.7–1.1 in R, respectively. Regarding Tree Regression, an analysis of variance (ANOVA) was chosen as the criterion for assessing node split quality, while default settings were retained for all other parameters. For Random Forest, hyperparameter tuning led to the selection of 300 trees as the optimal parameter within the range of 10–10,000. Variable importances were computed using two measures: *%IncMSE*, assessing the rise in MSE resulting from the removal of the variable; *IncNodePurity*, quantifying the increase in residual sum of squares attributable to the exclusion of the variable.

#### Statistical analysis

Continuous variables were expressed as median (interquartile range, IQR), and categorical variables as number, *n* (percentage, %). Correlation analyses employed the Spearman correlation coefficient, by the *ggpubr* package v0.6.0 in R. A *p*-value  $\leq 0.05$  was considered significant. Visualizations were generated using *ggplot2* and *plotmo* libraries in R v4.3.

## Results

#### Study population

A total of 497 PLWH within the SPID-HIV-Vax cohort met the inclusion criteria for the present study. The vaccination schedule and blood sample collection are reported in Fig. 1. Baseline characteristics of the study participants are reported in detail in Table 1. Briefly, median age was 54 (IQR: 44–59) years, and 408 (82.1%) were males. Median time from HIV diagnosis was 12 (IQR: 7–22) years. Median CD4 T-cell nadir was 220 (IQR: 81–370) cells/ $\mu$ L; current CD4 T-cell count was 701 (IQR: 512–934) cells/ $\mu$ L with a median CD4/CD8 ratio of 0.81 (IQR: 0.56–1.14). All participants have been on ART for a median of 9 (IQR: 5–15) years, and 483

**Table 1** Demographic and HIV-related characteristics of PLWH at baseline

| Characteristics                           | PLWH (n = 497)      |
|---|---------------------|
| Age, years, median (IQR)                  | 54 (44–59)          |
| Sex at birth, n (%)                       |                     |
| Male                                      | 408 (82.1)          |
| Female                                    | 89 (17.9)           |
| Ethnicity, n (%)                          |                     |
| Caucasian                                 | 428 (86.12)         |
| Arab                                      | 9 (1.81)            |
| Latin                                     | 41 (8.25)           |
| African                                   | 6 (1.21)            |
| Asian                                     | 13 (2.61)           |
| Epidemiology, n (%)                       |                     |
| MSM                                       | 254 (51.11)         |
| MSW/WSM                                   | 148 (29.79)         |
| Transgender                               | 1 (0.2)             |
| IDU                                       | 71 (14.29)          |
| Vertical transmission                     | 5 (1.01)            |
| Other/unknown                             | 18 (3.6)            |
| Comorbidities, n (%)                      |                     |
| None                                      | 279 (56.14)         |
| Hypertension                              | 101 (20.32)         |
| Cardiovascular disease                    | 52 (10.46)          |
| Cerebrovascular disease                   | 20 (4.02)           |
| Chronic kidney disease                    | 30 (6.05)           |
| COPD/Asthma                               | 33 (6.64)           |
| Peptic ulcer                              | 12 (2.41)           |
| Hemiplegy                                 | 9 (1.81)            |
| Previous leukemia/lymphoma                | 19 (3.82)           |
| Previous solid cancer                     | 35 (3.82)           |
| Metastatic solid cancer                   | 9 (1.81)            |
| Autoimmune diseases                       | 48 (9.66)           |
| Neurologic disease                        | 30 (6.05)           |
| Charlson Comorbidity Index, median (IQR)  | 0 (0–1)             |
| BMI, median (IQR)                         | 24.62 (22.66–27.46) |
| BMI strata, n (%)                         |                     |
| < 18.5                                    | 19 (3.8)            |
| 18.5–24.99                                | 249 (50.1)          |
| 25–29.9                                   | 174 (35)            |
| ≥ 30                                      | 55 (11.1)           |
| Viro-immunologic parameters, median (IQR) |                     |
| CD4 T-cell nadir, cells/μL                | 220 (81–370)        |
| CD4 T-cell count, cells/μL                | 701 (512–934)       |
| CD4 T-cell percentage                     | 32 (25–39)          |
| CD8 T-cell count, cells/μL                | 885 (640–1156)      |
| CD8 T-cell percentage                     | 40 (33–47)          |
| CD4/CD8 ratio                             | 0.81 (0.56–1.14)    |
| CD4 T-cell count strata, n (%)            |                     |
| < 200 cells/μL                            | 20 (4.02)           |
| 200–500 cells/μL                          | 95 (19.12)          |

**Table 1** (continued)

| Characteristics  | PLWH (n = 497) |
|--|----------------|
| > 500 cells/μL   | 382 (76.86)    |
| CD4/CD8 ratio ≥ 1, n (%)                                       | 174 (18.91)    |
| Undetectable HIV-RNA (< 50 copies/mL), n (%)                   | 483 (97.18)    |
| Previous AIDS diagnosis, n (%)                                 | 150 (30.18)    |
| Time from HIV diagnosis, years, median (IQR)                   | 12 (7–22)      |
| Current ART regimen, n (%)                                     |                |
| INSTI-based  | 361 (72.64)    |
| PI-based   | 34 (6.84)      |
| INSTI + PI-based   | 8 (1.61)       |
| NNRTI-based  | 94 (18.91)     |
| Duration of ART, years, median (IQR)                           | 9 (5–15)       |
| Previous SARS-CoV-2 infection, n (%)                           | 37 (7.44)      |
| Time between prime and booster vaccination, days, median (IQR) | 176 (162–190)  |

IQR interquartile range, MSM men who have sex with men, MSW men who have sex with women, WSM women who have sex with men, IDU injective drug use, COPD chronic obstructive pulmonary disease, BMI body mass index, AIDS acquired immunodeficiency syndrome, INSTI integrase strand-transfer inhibitor, PI protease inhibitor, NNRTI non-nucleoside reverse transcriptase inhibitor

(97.18%) had undetectable plasma HIV-RNA. Thirty-seven (7.44%) PLWH reported SARS-CoV-2 infection prior to vaccination.

**Regression models selection**

To develop a predictive model for the antibody concentrations at each sequential time point following the administration of SARS-CoV-2 vaccine, linear and non-linear models were constructed as described above.

**Generalized linear models**

All three GLM models—derived from SE, SB and MM feature selection approaches—identified a positive relationship of antibody concentrations with the presence of a prior SARS-CoV-2 infection, previous AIDS events, and duration of ART, whilst a negative association with CD8 T-cell percentage (Additional file 1: Tables S1–S3). Additionally, SF and SB models revealed a positive dependence of vaccine-elicited anti-S IgG on sex at birth (female) and heterosexual behavior, while a negative correlation with Caucasian ethnicity. Specifically, SF model also identified a negative association with time between primary and booster vaccination, as well as with HIV-RNA copies in plasma, while SB model found a significant positive dependence on BMI. Lastly, the MM inference approach uncovered a positive relationship with BMI and a negative one with CD4 T-cell nadir. Overall, these three linear models showed a generally low cross-validated R<sup>2</sup> and elevated cross-validated RMSE

indicative of suboptimal performances on the test dataset (Table 2). Consequently, a shift towards non-linear methodologies was employed to capture the intricate relationships among variables more accurately.

### Non-linear models

The Tree Regression model exhibited enhanced performances compared to its linear counterparts ( $CV-R^2=0.795$ ,  $CV-RMSE=0.451$ ) (Table 2). The optimal tree configuration identified 12 variables, including previous SARS-CoV-2 infection, BMI, plasma HIV-RNA and HIV-related viro-immunological parameters (Additional file 1: Fig S1). Random Forest outperformed the preceding methodologies in terms of both highest  $R^2$  and lowest RMSE ( $CV-R^2=0.845$ ,  $CV-RMSE=0.412$ ) (Table 2), leading to its selection as the optimal model for subsequent analysis.

### Antibody prediction through random forest regression

The selected Random Forest regression approach identified, among 39 different clinical, demographic, and immunological factors, several variables as more influential in predicting the antibody response, such as demographic variables (age and Caucasian ethnicity), BMI, previous SARS-CoV-2 infection, the number of days between the second and third vaccine doses, HIV-related viro-immunological parameters, time since HIV infection acquisition, and duration of ART (Fig. 2A). The top 5 variables, in terms of both %IncMSE and IncNodePurity, in order of importance, were previous SARS-CoV-2 infection, BMI, CD4 T-cell count, days between prime and booster vaccination, and CD4/CD8 ratio. To investigate the role of these variables in predicting the antibody response over time, the variable “days between prime and booster vaccination” was excluded since it could not impact time points preceding the third dose. The temporal influence of each of the selected 4 variables on the prediction of the antibody response within the Random Forest model was investigated using 3D partial

dependence plots. The graphs were obtained by pairing each variable of interest with the temporal one, while setting all other variables (background variables) to their median value (for continuous variables) or mode (for categorical variables) (Fig. 2B–E).

A pronounced dependence of the predicted antibody response was observed at earlier time points after the administration of the primary vaccine cycle for all selected variables. This dependence gradually diminished over time with the consequent administration of vaccine doses, culminating in a uniform humoral response for each value assumed by the examined variable. Specifically, a positive relationship was observed with the presence of a previous SARS-CoV-2 infection before vaccine administration (Fig. 2B), while reduced antibody concentrations were noted for lower values of BMI, CD4 T-cell count, and CD4/CD8 ratio (Fig. 2C–E).

In quantitative terms, the declining impact over time of the top 5 variables on the antibody response was corroborated by a Spearman correlation analysis conducted across the entire study population at each time point (Additional file 1: Fig S2). Notably, similar trends to those highlighted by applying the Random Forest model were observed, since the significance of the correlation showed a diminishing trend with the increasing number of vaccine doses received over time. This alignment between the correlation analysis and the Random Forest model outcomes reinforces the temporal evolution of variables’ influence on the antibody response, emphasizing the consistency of observations across analytical methodologies.

### Discussion

In this study, we leveraged data from a longitudinal observational study to train and test different machine learning algorithms to develop a predictive model of immune responses to SARS-CoV-2 mRNA primary and booster vaccination in PLWH. The specific aims were to forecast vaccine-elicited humoral responses in this vulnerable population based on several demographic and clinical information that may be easily retrieved from electronic charts in clinical practice settings, and to simultaneously analyze the impact of these variables on antibody production over time.

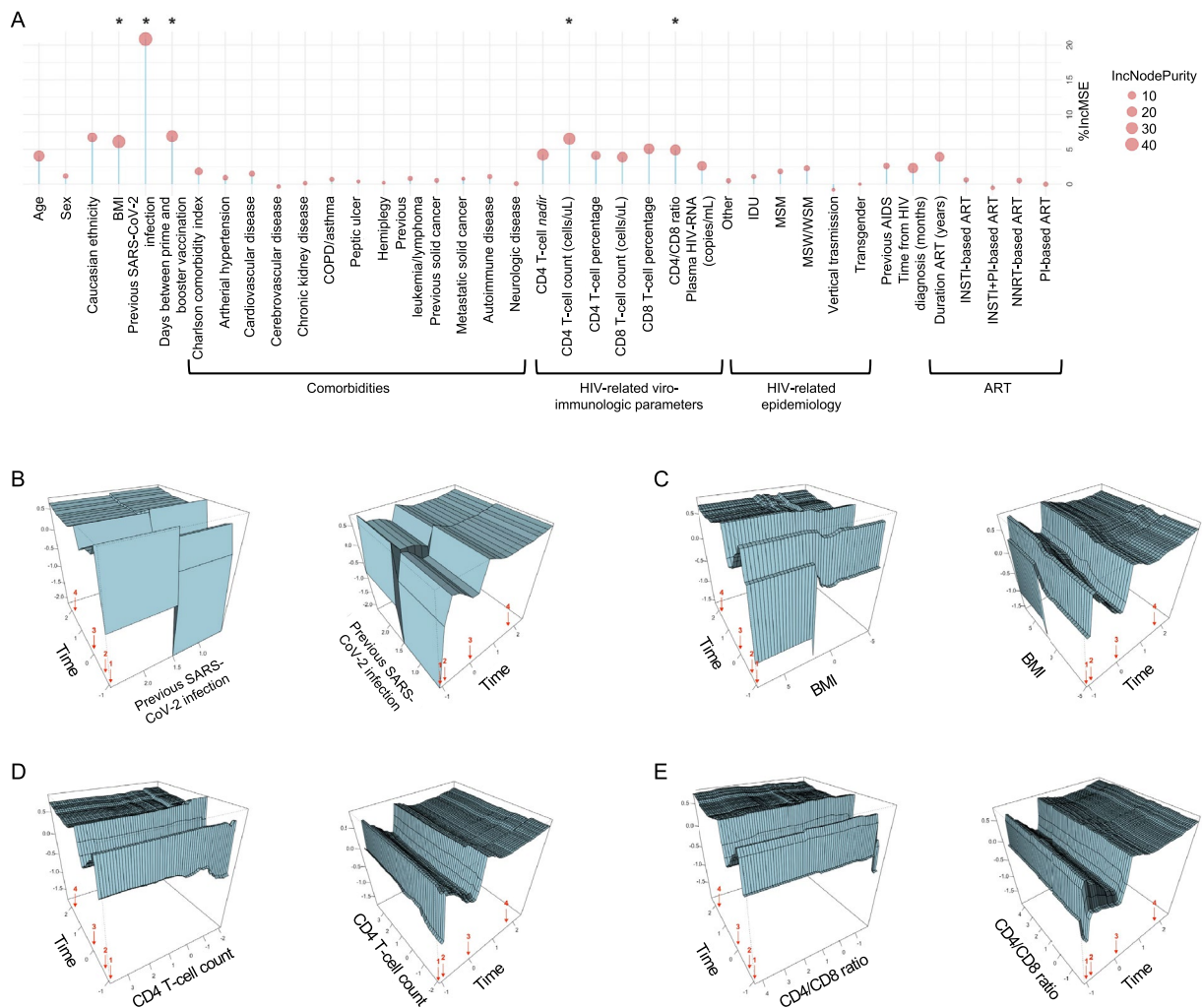
We found that, while commonly used linear regression models show suboptimal performances, non-linear methodologies display a significantly better ability to capture the intricate relationships among variables. In particular, Random Forest regression resulted as the best performing algorithm in predicting vaccine-induced antibody response. This is likely attributable to the fact that the various feature selection strategies employed in linear models often lead to the exclusion of important variables that, when considered individually, may have

**Table 2** Linear and non-linear model performances

| Name              | Model               | AIC    | BIC    | CV R <sup>2</sup> | CV RMSE |
|-------------------|---------------------|--------|--------|-------------------|---------|
| Stepwise-forward  | <i>glm</i>          | 4047.2 | 4123.2 | 0.357             | 0.801   |
| Stepwise-backward | <i>glm</i>          | 4072.1 | 4169.8 | 0.358             | 0.801   |
| glmulti           | <i>glm</i>          | 4079.9 | 4134.2 | 0.351             | 0.804   |
| Tree regression   | <i>rpart</i>        | –      | –      | 0.795             | 0.451   |
| Random forest     | <i>randomForest</i> | –      | –      | 0.845             | 0.412   |

R<sup>2</sup> (CV-R<sup>2</sup>) and Root Mean Squared Error (CV-RMSE) metrics have been computed on test set using a fivefold cross-validation approach

AIC Akaike Information Criterion obtained on the train set, BIC Bayesian Information Criterion computed on the train set



**Fig. 2** Random Forest regression analysis. **A** Variable importance resulting from Random Forest model. Variable importances are expressed in terms of the rise in MSE resulting from the removal of the variable (namely %IncMSE, y-axis) and the increase in residual sum of squares attributable to the exclusion of the variable (namely IncNodePurity, sphere dimension). Top 5 selected important variables are marked with an asterisk (\*). **B-E** Interaction between top important variables and time since first vaccine dose within the Random Forest model. 3D partial dependence plots are generated by plotting the predicted response [anti-S IgG expressed in binding antibody units per milliliter (BAU/mL)] on the z-axis as two variables are changed (all other variables held at their median/mode values). Numerical variables are normalized using z-score. Red arrows indicate the number of vaccine doses received. For each pair of parameters two different visualizations were obtained by rotating the plot around the z-axis with an angle of 80°. The interaction of time since first dose administration, reported on the x-axis, versus previous SARS-CoV-2 infection (**B**), BMI (**C**), CD4 T-cell count (**D**), and CD4/CD8 ratio (**E**). *BMI* body mass index, *COPD* chronic obstructive pulmonary disease, *IDU* injective drug use, *MSM* men who have sex with men, *MSW* men who have sex with women, *WSM* women who have sex with men, *AIDS* acquired immunodeficiency syndrome, *ART* antiretroviral therapy, *INSTI* integrase strand-transfer inhibitor, *PI* protease inhibitor, *NNRTI* non-nucleoside reverse transcriptase inhibitor

a minor role, whereas in a multi-variable context they would assume a stronger predictive role [25, 26].

Notably, the key clinical factors influencing the vaccine humoral immunogenicity that were identified by the Random Forest model were: previous SARS-CoV-2 infection, CD4 T-cell count, CD4/CD8 ratio, BMI, and time between primary vaccination cycle and booster dose. In detail, SARS-CoV-2 infection before vaccine administration appeared to positively influence the vaccine-elicited

antibody levels. By contrast, low CD4 T-cell counts, CD4/CD8 ratio and BMI values were associated with reduced antibody responses to the vaccine. Lastly, increasing time between primary vaccination cycle and booster dose was associated with higher antibody levels. Remarkably, these key clinical factors, identified through the Random Forest machine learning approach, are congruent with clinical and laboratory observations [3, 9, 10, 12, 27,

28]. This corroborates and supports the validity of our model, confirming it as an important tool for future prediction studies. Indeed, hybrid immunity, derived from a combination of both natural infection and vaccination, has been shown to ensure immune protection which is higher in both magnitude and durability than that provided by either vaccine or infection alone [29]. Furthermore, poor immune recovery despite ART has been distinctly associated with reduced humoral and T-cell responses to SARS-CoV-2 vaccines in PLWH [3, 9–12]. Similarly to obese people, underweight ones develop weaker immune responses to SARS-CoV-2 vaccines [28], due to a severe impairment of the immune system [30]. Lastly, several studies demonstrated that an extended interval between SARS-CoV-2 mRNA vaccine doses results in stronger humoral responses [31–34], owing to a decline in antibody levels which limits the Fc-mediated clearance of vaccine-encoded antigens, thus allowing de novo priming of B cells [34].

In this context, our machine learning approach expands the knowledge of the modeling strategies to be employed in studies aiming to predict outcomes involving complex biological mechanisms. Indeed, we demonstrated that such associations are not linear and thus more nuanced than previously believed, due to the reciprocal interactions between such factors in influencing vaccine-induced humoral responses.

Additionally, while previous research mainly focused on the primary vaccine cycle, the present study extends the knowledge on immune responses to booster doses in this vulnerable population. In this respect, a reduction of the model dependence from the identified predictors over time was observed, revealing that while the aforementioned factors may play a critical role in dictating humoral immunogenicity to the primary vaccine cycle in PLWH, the importance of their role significantly wane over time, so that antibody responses to booster shots are uniform across the entire population regardless demographic and clinical features.

Some limitations need to be acknowledged in this study. Firstly, the model herein presented was developed using data derived from individuals vaccinated with the Spikevax™ mRNA vaccine (Moderna), and thus may not directly translate to PLWH receiving other SARS-CoV-2 vaccines or heterologous vaccine combinations. Additionally, while previous SARS-CoV-2 infection was recorded at baseline, data on breakthrough infections during the follow-up period were not available. Lastly, the sample size was relatively small, especially for latest time points. Expanding the scope of the investigation to encompass different vaccine platforms and heterologous prime-boost combinations, alongside with data from other cohorts of PLWH

and other fragile populations and vaccine antigens, will strengthen such findings, providing valuable insights for the design of future vaccination strategies.

## Conclusions

This study showcases that machine learning algorithms capable of quantifying non-linear associations allow to accurately predict humoral responses to SARS-CoV-2 mRNA primary and booster vaccination in PLWH by employing clinical, demographic and HIV-related variables commonly available in medical charts. While low CD4 T-cell counts, CD4/CD8 ratio and BMI are associated with poor immunogenicity of the primary vaccine cycle, the administration of additional doses overcome the negative influence of these factors, suggesting that further booster doses should be offered to PLWH. Moreover, the application of this model in clinical contexts holds potential promise for public health strategies, empowering clinicians to predict individual humoral responses to vaccination, simply using demographic and clinical information that may be easily retrieved in medical practice settings. As a result, our model not only contributes to a deeper understanding of vaccine responsiveness but also offers practical guidance for implementing effective and targeted vaccination strategies in PLWH that can be particularly helpful for improving possible epidemic or pandemic vaccination policies.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12967-024-05147-1>.

**Additional file 1: Table S1.** Summary of Forward model regression analysis. **Table S2.** Summary of Backward model regression analysis. **Table S3.** Summary of Multi-Model regression analysis. **Figure S1.** Tree Regression model importance plot. Variables selected as most important and used to build the final optimal tree of the model. Variable importances, calculated by rpart package, are reported on both x-axis and sphere radius. ART: antiretroviral therapy; BMI: body mass index. **Figure S2.** Spearman correlation analysis. Correlation analysis performed between the top 5 important variables selected from the Random Forest model (y-axis) and anti-S IgG (binding antibody units per milliliter (BAU/mL)) at each time point (x-axis). R<sup>2</sup> values are expressed as colour gradient ranging from violet to yellow. Numerical P-values of each pairwise comparison are reported within each box. BMI: body mass index; NA: not applicable; T1: 1 month after the first dose, coinciding with the day of the second dose; T2: 1 month after the second dose; T3: 6 months after the second dose, coinciding with the third dose administration; T4: 1 month after the third dose; T5: 6 months after the third dose; T6: 12 months after the third dose, coinciding with the fourth dose administration; T7: 1 month after the fourth dose; T8: 6 months after the fourth dose.

## Acknowledgements

We thank all the SPID-HIV-Vax participants. Our special thanks go to all the physicians and nurses at the Clinic of Infectious Diseases and Tropical Medicine at San Paolo Hospital in Milan who continuously helped in patients' care.

### Author contributions

Conceptualization of the study: A.C., G.Ma., D.M. Clinical study design: M.A., G.Ma.; Acquisition of data: M.A., G.Ma.; Statistical analysis and modeling: G.Mo. Interpretation of the data and figure design: G.Mo., M.A., J.P.; Draft of the article: M.A., G.Mo., J.P.; Review of the article and critical revision of important intellectual content: A.C., G.Ma., D.M..

### Funding

This research was supported by the Department of Medical Biotechnologies of the University of Siena (D.M.), by Project 2021–4236 “LLC Network”, Fondazione Cariplo (G.Ma.) and by the PNRR PE13 INF\_ACT—CUP B63C22001400007 project.

### Availability of data and materials

The datasets generated during the current study are not publicly available because they contain sensitive data to be treated under data protection laws and regulations. Appropriate forms of data sharing can be arranged after a reasonable request to the corresponding author.

### Declarations

#### Ethics approval and consent to participate

The study was approved by the local Ethical Committee and written informed consent was obtained from each participant.

#### Consent for publication

All authors have read and agreed to the submitted version of the manuscript.

#### Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential competing interests.

#### Author details

<sup>1</sup>Laboratory of Molecular Microbiology and Biotechnology, Department of Medical Biotechnologies, University of Siena, Siena, Italy. <sup>2</sup>Clinic of Infectious Diseases and Tropical Medicine, Department of Health Sciences, San Paolo Hospital, ASST Santi Paolo e Carlo, University of Milan, Milan, Italy.

Received: 24 January 2024 Accepted: 31 March 2024

Published online: 07 May 2024

### References

- Vizcarra P, Pérez-Eliás MJ, Quereda C, Moreno A, Vivancos MJ, Dronda F, et al. Description of COVID-19 in HIV-infected individuals: a single-centre, prospective cohort. *Lancet HIV*. 2020;7:e554–64.
- Del Amo J, Polo R, Moreno S, Díaz A, Martínez E Jr, et al. Incidence and severity of COVID-19 in HIV-positive persons receiving antiretroviral therapy: a cohort study. *Ann Intern Med*. 2020. <https://doi.org/10.7326/M20-3689>.
- Augello M, Bono V, Rovito R, Tincati C, Marchetti G. Immunologic interplay between HIV/AIDS and COVID-19: adding fuel to the flames? *Curr HIV/AIDS Rep*. 2023. <https://doi.org/10.1007/s11904-023-00647-z>.
- Geretti A, Stockdale A, Kelly S, Cevik M, Collins S, L W, et al. Outcomes of coronavirus disease 2019 (COVID-19) related hospitalization among people with human immunodeficiency virus (HIV) in the ISARIC world health organization (WHO) clinical characterization protocol (UK): a prospective observational study. *Clin Infect Dis*. 2021;73:e2095–106.
- Giacomelli A, Gagliardini R, Tavelli A, Benedittis SD, Mazzotta V, Rizzardini G, et al. Risk of COVID-19 in-hospital mortality in people living with HIV compared to general population according to age and CD4 strata: data from the ICONA network. *Int J Infect Dis*. 2023;136:127–35.
- Western Cape Department of Health in collaboration with the National Institute for Communicable Diseases, South Africa. Risk Factors for Coronavirus Disease 2019 (COVID-19) Death in a Population Cohort Study from the Western Cape Province, South Africa. *Clin Infect Dis*. 2021. <https://doi.org/10.1093/cid/ciaa1198>
- Yang X, Sun J, Patel R, Zhang J, Guo S, Q Z, et al. Associations between HIV infection and clinical spectrum of COVID-19: a population level analysis based on US National COVID cohort collaborative (N3C) data. *Lancet HIV*. 2021. [https://doi.org/10.1016/S2352-3018\(21\)00239-3](https://doi.org/10.1016/S2352-3018(21)00239-3).
- Augello M, Bono V, Rovito R, Tincati C, Bianchi S, Taramasso L, et al. Association between SARS-CoV-2 RNAemia, skewed T cell responses, inflammation, and severity in hospitalized COVID-19 people living with HIV. *iScience*. 2024. <https://doi.org/10.1016/j.isci.2023.108673>.
- Augello M, Bono V, Rovito R, Tincati C, d'Arminio Monforte A, Marchetti G. Six-month immune responses to mRNA-1273 vaccine in combination antiretroviral therapy treated late presenter people with HIV according to previous SARS-CoV-2 infection. *AIDS*. 2023. <https://doi.org/10.1097/QAD.0000000000003585>.
- Antinori A, Cicalini S, Meschi S, Bordoni V, Lorenzini P, Vergori A, et al. Humoral and cellular immune response elicited by mRNA vaccination against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in people living with human immunodeficiency virus receiving antiretroviral therapy based on current CD4 T-lymphocyte count. *Clin Infect Dis*. 2022;75:e552.
- Polvere J, Fabbiani M, Pastore G, Rancan I, Rossetti B, Medagliani D, et al. B cell response after SARS-CoV-2 mRNA vaccination in people living with HIV. *Commun Med*. 2023. <https://doi.org/10.1038/s43856-023-00245-5>.
- Benet S, Blanch-Lombarte O, Ainsua-Enrich E, Pedreño-Lopez N, Muñoz-Basagoiti J, Raich-Regué D, et al. Limited humoral and specific T-cell responses after SARS-CoV-2 vaccination in PWH with poor immune reconstitution. *J Infect Dis*. 2022. <https://doi.org/10.1093/infdis/jiac406>.
- Jongkees M, Geers D, Hensley K, Huisman W, GeurtsvanKessel C, Bogers S, et al. Immunogenicity of an additional mRNA-1273 SARS-CoV-2 vaccination in people with HIV with hyporesponse after primary vaccination. *J Infect Dis*. 2023. <https://doi.org/10.1093/infdis/jiac451>.
- Vergori A, Lepri AC, Cicalini S, Matusali G, Bordoni V, Lanini S, et al. Immunogenicity to COVID-19 mRNA vaccine third dose in people living with HIV. *Nat Commun*. 2022. <https://doi.org/10.1038/s41467-022-32263-7>.
- Vergori A, Tavelli A, Matusali G, Azzini A, Augello M, Mazzotta V, et al. SARS-CoV-2 mRNA vaccine response in people living with HIV according to CD4 count and CD4/CD8 ratio. *Vaccines*. 2023. <https://doi.org/10.3390/vaccines11111664>.
- Lapointe H, Mwimanzu F, Cheung P, Sang Y, Yaseen F, Speckmaier S, et al. Antibody response durability following three-dose coronavirus disease 2019 vaccination in people with HIV receiving suppressive antiretroviral therapy. *AIDS*. 2023. <https://doi.org/10.1097/QAD.0000000000003469>.
- Tau L, Hagin D, Freund T, Halperin T, Adler A, Marom R, et al. Humoral and cellular immune responses of people living with human immunodeficiency virus after 3 doses of messenger RNA BNT162b2 severe acute respiratory syndrome coronavirus 2 vaccine: a prospective cohort study. *Open Forum Infect Dis*. 2023. <https://doi.org/10.1093/ofid/ofad347>.
- Heftdal LD, Pérez-Alós L, Hasselbalch RB, Hansen CB, Hamm SR, Møller DL, et al. Humoral and cellular immune responses eleven months after the third dose of BNT162b2 an mRNA-based COVID-19 vaccine in people with HIV—a prospective observational cohort study. *eBioMedicine*. 2023. <https://doi.org/10.1016/j.ebiom.2023.104661>.
- Kageyama T, Ikeda K, Tanaka S, Taniguchi T, Igari H, Onouchi Y, et al. Antibody responses to BNT162b2 mRNA COVID-19 vaccine and their predictors among healthcare workers in a tertiary referral hospital in Japan. *Clin Microbiol Infect*. 2021. <https://doi.org/10.1016/j.cmi.2021.07.042>.
- Noe S, Ochana N, Wiese C, Schabaz F, Von Krosigk A, Heldwein S, et al. Humoral response to SARS-CoV-2 vaccines in people living with HIV. *Infection*. 2022. <https://doi.org/10.1007/s15010-021-01721-7>.
- Wiens J, Shenoy E. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis*. 2018. <https://doi.org/10.1093/cid/cix731>.
- Alejo J, Mitchell J, Chiang T, Chang A, Abedon A, Wa W, et al. Predicting a positive antibody response after 2 SARS-CoV-2 mRNA vaccines in transplant recipients: a machine learning approach with external validation. *Transplantation*. 2022. <https://doi.org/10.1097/TP.0000000000004259>.
- Giannella M, Huth M, Righi E, Hasenauer J, Marconi L, Konnova A, et al. Using machine learning to predict antibody response to SARS-CoV-2 vaccination in solid organ transplant recipients: the multicentre ORCHESTRA cohort. *Clin Microbiol Infect*. 2023. <https://doi.org/10.1016/j.cmi.2023.04.027>.



24. Harris P, Taylor R, Minor B, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform.* 2019. <https://doi.org/10.1016/j.jbi.2019.103208>.
25. Olusegun AM, Dikko HG, Gulumbe SU. Identifying the limitation of stepwise selection for variable selection in regression analysis. *Am J Theor Appl Stat.* 2015;4:414–9.
26. Steyerberg E, Eijkemans M, Habbema J. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol.* 1999. [https://doi.org/10.1016/s0895-4356\(99\)00103-1](https://doi.org/10.1016/s0895-4356(99)00103-1).
27. Chammartin F, Kusejko K, Pasin C, Trkola A, Briel M, P A, et al. Determinants of antibody response to severe acute respiratory syndrome coronavirus 2 mRNA vaccines in people with HIV. *AIDS.* 2022. <https://doi.org/10.1097/QAD.0000000000003246>.
28. Piernas C, Patone M, Astbury N, Gao M, Sheikh A, Khunti K, et al. Associations of BMI with COVID-19 vaccine uptake, vaccine effectiveness, and risk of severe COVID-19 outcomes after vaccination in England: a population-based cohort study. *Lancet Diabetes Endocrinol.* 2022. [https://doi.org/10.1016/S2213-8587\(22\)00158-9](https://doi.org/10.1016/S2213-8587(22)00158-9).
29. Lasrado N, Barouch D. SARS-CoV-2 hybrid immunity: the best of both worlds. *J Infect Dis.* 2023. <https://doi.org/10.1093/infdis/jiad353>.
30. Dobner J, Kaser S. Body mass index and the risk of infection—from underweight to obesity. *Clin Microbiol Infect.* 2018. <https://doi.org/10.1016/j.cmi.2017.02.013>.
31. Tauzin A, Gong S, Beaudoin-Bussi eres G, V ezina D, Gasser R, Nault L, et al. Strong humoral immune responses against SARS-CoV-2 Spike after BNT162b2 mRNA vaccination with a 16-week interval between doses. *Cell Host Microbe.* 2022. <https://doi.org/10.1016/j.chom.2021.12.004>.
32. Nicolas A, Sannier G, Dub e M, Nayrac M, Tauzin A, Mm P, et al. An extended SARS-CoV-2 mRNA vaccine prime-boost interval enhances B cell immunity with limited impact on T cells. *IScience.* 2023. <https://doi.org/10.1016/j.isci.2022.105904>.
33. Hall V, Ferreira V, Wood H, Ierullo M, Majchrzak-Kita B, Manguiat K, et al. Delayed-interval BNT162b2 mRNA COVID-19 vaccination enhances humoral immunity and induces robust T cell responses. *Nat Immunol.* 2022. <https://doi.org/10.1038/s41590-021-01126-6>.
34. Dangi T, Sanchez S, Lew M, Visvabharathy L, Richner J, Koralnik I, et al. Pre-existing immunity modulates responses to mRNA boosters. *Cell Rep.* 2023. <https://doi.org/10.1016/j.celrep.2023.112167>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.