Prediction of Antibody Responses to a Newly Authorized Vaccine

Minghe Wang, Zebang Zhang, Xuanyu Guo

2025-03-27

1. Introduction

Understanding variation in antibody response after vaccination is key to identifying individuals who may be less protected. In this study, we analyzed data from 5,000 participants who received a new vaccine to examine how demographic and clinical factors relate to antibody levels. The outcome of interest was log-transformed antibody concentration. We aimed to build predictive models and assess their ability to generalize to new data. Three approaches were compared: LASSO regression, Generalized Additive Models (GAMs), and Multivariate Adaptive Regression Splines (MARS). Each model was trained on one dataset and evaluated on an independent test set to ensure robustness.

2. Exploratory Data Analysis (EDA)

We first verified the structure and completeness of the training (`dat1`) and test (`dat2`) datasets. Both contained 14 variables with no missing values and identical column structures. The outcome variable, `log_antibody`, was continuous and approximately symmetric, making it suitable for linear or semi-parametric modeling.

2.1 Continuous Variables

Summary statistics and visualizations for continuous predictors (e.g., age, BMI, SBP, LDL, time since vaccination) revealed varied distributions. Age, BMI, and SBP were roughly normal, while LDL and time were right-skewed with some outliers. Scatterplots with LOESS smoothing indicated potential non-linear relationships between BMI, weight, and time with the outcome, suggesting the need for flexible models like GAM or MARS.

2.2 Categorical Variables

Categorical variables (gender, race, smoking, diabetes, hypertension) were converted to labeled factors. Gender was balanced, but most participants were White and never smokers. Diabetes was uncommon, while hypertension was moderately present. Boxplots showed modest group-level differences in `log_antibody`, with lower levels observed among current smokers and males.

2.3 Correlation Analysis

Strong correlations were observed between height, weight, and BMI (Pearson $r > 0.7$). VIF analysis confirmed multicollinearity among these variables (VIF > 10). Since BMI is both derived from height and weight and more clinically meaningful, we retained BMI and excluded the other two from modeling to improve interpretability and model stability.

3. Model Training

To identify key predictors and build a robust predictive model of antibody response, we trained and evaluated three modeling strategies: LASSO regression, Generalized Additive Models (GAMs), and Multivariate Adaptive Regression Splines (MARS). All models were trained on `dat1` using 10-fold cross-validation for hyperparameter tuning. Based on VIF analysis, we excluded height and weight due to multicollinearity and retained BMI as a clinically interpretable composite metric.

3.1 LASSO Regression

We began with LASSO regression to reduce dimensionality and mitigate multicollinearity. Using the `glmnet` package with standardized predictors, cross-validation identified an optimal penalty (lambda.min = 0.0067). The final model selected a sparse subset of predictors: age, gender (Male), smoking (Current), BMI, time since vaccination, and race (Hispanic). The root mean squared error (RMSE) of the LASSO model on the training set was 0.5518, while its performance on the test set yielded an RMSE of 0.5749. These results highlight LASSO's limitations in capturing complex, non-linear patterns, despite its value for feature selection.

3.2 Generalized Additive Models (GAMs)

To better accommodate potential non-linear effects observed during exploratory analysis, we employed a Generalized Additive Model using the `mgcv` package. The model specification allowed smooth terms for continuous variables including age, SBP, LDL, BMI, time, while categorical variables such as gender, race, smoking, diabetes, and hypertension were modeled linearly. The GAM demonstrated moderate predictive capacity, achieving an adjusted R-squared of 0.22 and explaining 22.4% of the total deviance. Importantly, several smooth terms, including those for BMI and time, were statistically significant, indicating clear non-linear relationships. For instance, antibody levels peaked at moderate BMI values then declined at higher BMI levels. The GAM's training RMSE was approximately 0.528, with test RMSE increasing to 0.5701, suggesting potential overfitting or sensitivity to shifts in predictor distributions across datasets.

3.3 Multivariate Adaptive Regression Splines (MARS)

Finally, we implemented MARS using the `earth` package, which automatically captures non-linearities and interaction effects via piecewise linear basis functions. A grid search over model parameters identified the optimal configuration with degree = 1 (indicating no interaction terms) and nine basis functions (nprune = 9). The final model included hinge functions on BMI, age, and time, alongside categorical indicators for gender and smoking. Compared to other models, MARS demonstrated the best balance of flexibility and generalizability, achieving a training RMSE of 0.528 and a test RMSE of 0.5328. The inclusion of threshold-based transformations allowed MARS to capture inflection points—such as a decline in antibody levels beyond a BMI of approximately 23.7 or post-vaccination time beyond 57 days—that were difficult for other models to approximate.

4. Results

The MARS model was selected as the final model due to its superior balance of performance and interpretability.

4.1 Final Model Interpretation (MARS)

MARS identified key non-linear effects: antibody levels peaked at about BMI 24 and declined beyond; responses increased before ~57 days post-vaccination but then decreased afterwards. Response decreased slightly as age increased. Males and current smokers had lower predicted antibody levels. Partial dependence and LOESS plots on `dat2` confirmed these trends.

4.2 Performance Metrics

Model	CV RMSE (Training)	Test RMSE (dat2)
MARS	0.5283	0.5328
GAM	0.5279	0.5701
LASSO	0.5518	0.5749

MARS achieved the lowest RMSE across datasets. Although GAM performed well during training, its test performance dropped, likely due to overfitting or distributional shifts (e.g., time range differences).

4.3 Demographic and Clinical Effects

Across demographic and clinical factors, consistent trend of partial dependence on response were observed. MARS predictions on `dat2` showed lower antibody levels in males and current smokers as expected. Age and BMI against antibody behave differently from trained model when using dat2, but can be explained by its different distribution. So we can still consider MARS model robust and well-generalized.

4.4 Temporal Trends

In `dat2`, antibody levels declined steadily with time, unlike the rise-then-fall seen in `dat1`. This reflects a lack of early time points (< 61 days) in `dat2`, highlighting the importance of time coverage for modeling temporal trends.

5. Discussion

Non-linear models outperformed linear ones in predicting antibody response. While LASSO served as a useful baseline, it missed complex trends. GAM captured smooth non-linear effects, but its test performance dropped. MARS showed the best balance of accuracy and interpretability, generalizing well across datasets attributed to its variable selection feature and flexibility in detecting and fitting pattern when our data has sharp turn. MARS also performs well in understanding partial factors' effects. And GAM is useful when understanding subgroup effect on time decaying behavior within factors. Key predictors included BMI, time, etc, aligning with known biological patterns. These findings support using flexible models and highlight the importance of full time coverage for capturing antibody dynamics.

6. Conclusion

We developed predictive models of vaccine-induced antibody response using demographic and clinical data. MARS outperformed other models, capturing key non-linear trends and generalizing well to new data. Factors like BMI, time, gender, and smoking were strong predictors of immune response. These results can help identify groups who may benefit from closer monitoring or earlier booster doses. Future work could expand the model by adding biomarker or longitudinal data to improve prediction.