Prediction of Antibody Responses to a Newly Authorized Vaccine

Minghe Wang, Zebang Zhang, Xuanyu Guo

2025-03-27

1. Introduction

Understanding individual variation in antibody response following vaccination is important for optimizing public health strategies and identifying potentially vulnerable populations. In this analysis, we used a dataset of 5,000 participants who received a newly authorized vaccine to explore predictors of log-transformed antibody levels measured via dried blood spot. Our primary objectives were to:

1. Describe patterns and associations between predictors and antibody levels.
2. Develop and evaluate predictive models of antibody response.
3. Assess the model's robustness using a new, independent dataset.

2. Exploratory Analysis (EDA)

2.1 Continuous Variables

According to the boxplots and density plots of continuous predictors:

- Age, BMI, and SBP are approximately normally distributed, spanning plausible adult ranges.
- LDL and time since vaccination are right-skewed with wider variability and visible outliers, which could influence linear modeling.
- The log-transformed antibody level (our response variable) appears symmetric and reasonably bell-shaped, supporting its use as a continuous outcome in linear or semi-parametric models.

Scatterplots with LOESS smoothing revealed potential non-linear relationships between predictors such as BMI, time, and weight with the response, suggesting that flexible modeling (e.g., GAM or MARS) may be appropriate.

2.2 Categorical Variables

According to the bar plots and boxplots by group:

- Gender is fairly balanced.
- Race distribution is skewed, with a majority identifying as White; minority racial groups are underrepresented.

- Smoking status is skewed toward never smokers.
- Most participants are non-diabetic, while hypertension status shows a more balanced distribution.

Boxplots of log_antibody by each category suggest possible differences, particularly by gender, race, and smoking status, hinting at group-level effects.

2.3 Correlation Analysis

- BMI and weight are highly correlated ($r > 0.7$), indicating redundancy. Multicollinearity diagnostics confirmed this, with high VIF values for height, weight, and BMI in the full linear model.
- log_antibody showed mild inverse correlations with age, BMI, and weight. Associations with LDL, SBP, and time were weak.
- Categorical predictor associations were assessed via chi-squared tests, revealing no major dependencies that would preclude joint inclusion in a model.

3. Model Training

To predict log-transformed antibody levels, we compared multiple modeling approaches, including linear regression, penalized regression (LASSO and Ridge), Generalized Additive Models (GAMs), and Multivariate Adaptive Regression Splines (MARS). The training dataset (dat1) was used to fit all models, and performance was evaluated using 10-fold cross-validation.

3.1 Data Preparation

- The outcome variable was log_antibody.
- Predictors included both continuous and categorical variables. Categorical predictors were converted to factors with appropriate labels.
- The id variable was excluded.
- Due to multicollinearity among height, weight, and BMI, we avoided including all three in the same model unless penalization or dimensionality reduction was applied.

3.2 Model

Model1: LASSO Regression

- LASSO regression was implemented using glmnet with alpha = 1.
- Standardization was applied.
- Cross-validation (cv.glmnet) selected the optimal lambda using the 1-SE rule.
- Final model retained age, gender, BMI, and smoking (current) as predictors.
- RMSE on the test dataset (dat2) was approximately 0.58.

Model2: Ridge Regression

- Ridge regression (alpha = 0) was also fit using glmnet.
- Ridge preserved all predictors but shrank coefficients toward zero.
- Predictive accuracy was similar to LASSO.

Model 3: Generalized Additive Model (GAM)

- GAM was fit using caret::train with method = "gam" and 10-fold cross-validation.
- The model formula included smoothing splines (s()) for continuous variables (age, BMI, SBP, LDL, time, height, weight) and linear terms for categorical variables (gender, race, smoking, diabetes, hypertension).
- The best tuning parameter (select = TRUE) was chosen automatically.
- GAM achieved a cross-validated RMSE around 0.53 and test RMSE of 0.57.

Model 4: Multivariate Adaptive Regression Splines (MARS)

- MARS was implemented using earth via caret::train.
- A grid search explored degrees (1–3) and number of pruning terms (2–15).
- 10-fold cross-validation was used to select the optimal configuration: degree = 1, nprune = 9.
- The final model included hinge functions for BMI, time, and age, as well as main effects for gender and smoking.
- MARS achieved the best performance:
  - Training RMSE: ~0.53
  - Test RMSE: ~0.53

4. Results

4.1 Final Model Interpretation (MARS)

- Higher BMI and current smoking were associated with lower antibody levels.
- Older age showed a modest negative effect on response.
- The effect of time since vaccination was non-linear, indicating potential antibody waning patterns.
- The model selected only a subset of predictors, improving interpretability.

4.2 Model Performance

- Cross-validated RMSE: ~0.53 (MARS), ~0.53–0.54 (GAM)
- Test RMSE on new dataset (dat2): MARS: 0.53, GAM: 0.57

The MARS model maintained stable performance on the independent dataset, indicating good generalizability. GAM showed slightly worse test RMSE, suggesting possible overfitting or sensitivity to distributional changes.

5. Discussion

The MARS model revealed that demographic (age, gender), behavioral (smoking), and clinical (BMI) variables are meaningful predictors of post-vaccination antibody levels. The non-linear association with time since vaccination is particularly important for monitoring waning immunity.

The model performed similarly on an independent dataset, supporting its generalizability. Slight degradation in GAM performance may reflect overfitting or mismatches in distributional characteristics across datasets.

6. Conclusion

We built a robust predictive model of antibody responses following vaccination using demographic and clinical data. Our findings underscore the importance of body composition, lifestyle, and time in shaping immune responses. Future work may expand this framework with immunogenomic or behavioral covariates and explore longer-term antibody trajectories.