Prevalence of SARS-CoV-2 Antibodies in Liberty University Student Population

Emily Bonus

A Senior Thesis submitted in partial fullfillment
of the requirements for graduation
in the Honors Program
Liberty University
Spring 2023

Acceptance of Senior Honors Thesis

This Senior Honors Thesis is accepted in partial
fullfillment of the requirements for graduation from the
Honors Program of Liberty University.

_____

David E. Schweitzer, Ph.D.
Thesis Chair

_____

David M. Rockabrand, Ph.D.
Committee Member

_____

Christopher M. Nelson, M.F.A.
Assistant Honors Director

_____

Date

Abstract

In 2020, the virus SARS-CoV-2 gained attention as it spread around the world. Its antibodies are poorly understood, and little research focuses on those with few COVID-19 complications yet large numbers of close contacts: university students. This longitudinal study recorded SARS-CoV-2 antibody presence in 107 undergraduate Liberty University students twice during early 2021. After extensive data cleaning and the application of various statistical tests and ANOVAs, the data seems to show that in the case of COVID-19 infections, SARS-CoV-2 IgM antibodies are immediately produced, and then IgG antibodies follow later. However, the COVID-19 vaccine causes the production of both IgM and IgG antibodies right away, which then disappear more rapidly than those from a natural infection.

Prevalence of SARS-CoV-2 Antibodies in Liberty University Student Population

Introduction

The virus SARS-CoV-2, which causes the disease COVID-19, has swept the globe since early 2020. It has been feared for its high transmissibility and unsuspected complications. Recent attempts at creating an effective vaccine have had to settle for mitigating the symptoms of the illness but not entirely preventing infection or transmission [1]. Beyond that, little is known about antibody efficacy, immunity, and reinfection [2]. These factors have caused many universities to choose to conduct operations online. However, university-aged students who are regularly in contact with each other and potentially being exposed to SARS-CoV-2 provide a unique opportunity to study the virus, as they belong to the adult age group with the lowest rate of hospitalizations and deaths, although they contract the disease with the same frequency of the rest of the adult population [3]. This study seeks to utilize data collected from residential undergraduate students at Liberty University in early 2021. Statistical analysis will be performed to analyze the prevalence of SARS-CoV-2 antibodies with the intention of correlating this with vaccination rates, positivity test ratios, and exposures to infected persons. The data was collected under the supervision of Dr. David Rockabrand and Dr. David Dewitt under “IRB-FY20-21-140 SARS-CoV-2 Serologic Response from a Sample of Liberty University Students”.

Literature Review

Antibodies and Vaccines

An antibody is a protein produced by the B cells of the immune system in response to a foreign substance, or antigen [4]. Antibodies latch onto the specific antigen for which they were

created in order to facilitate its removal from the body [5]. In the case of a pathogen, after infection, the antibody-producing cells will multiply quickly and produce an “immunological memory.” This means that there are sufficient antibodies that recognize the pathogen constantly circulating in the blood so that the body can quickly fight off any subsequent re-infections [5]. There are five types of antibodies, but only two, IgG and IgM, will be covered here. IgG antibodies are responsible for long-term immunity and help immune cells such as leukocytes and macrophages find and destroy the antigen [5]. IgM, on the other hand, make up a lower proportion of the body’s total antibodies and primarily act in the early phases of infection, binding with a higher avidity than IgG antibodies and then signaling other immune cell responses [5].

In the case of SARS-CoV-2, IgM antibodies peak around day 12 after infection and last about 35 days, whereas IgG peak at 17 days and last 49 days [4]. This is much shorter than the amount of time that IgG antibodies last for other well-known viruses [4]. There are a variety of uncertainties in this knowledge because of gaps in understanding of SARS-CoV-2 antibodies as well poor antibody tests with high false-negative rates [4]. These points of confusion make it all the more critical to explore how SARS-CoV-2 antibodies behave in contrast with the prior antibody understanding just explained.

When studying antibodies and immunity, vaccines play a large role in the research. First discovered more than 200 years ago, vaccines were initially developed by empirical evidence alone but are now understood to function in various immunological ways [6]. Although mechanisms differ, most vaccines function by introducing a weakened pathogen or a subunit of the pathogen, which activates the immune system and results in the production of neutralizing

antibodies and antigen-specific memory T cells [6]. The best vaccines confer immunity for more than 50 years, while most can at least improve upon the protection of natural immunity [6].

During the COVID-19 pandemic, a newer type of vaccine rose to the forefront of the medical community: the mRNA vaccine. These vaccines are modified strands of RNA that encode a portion of the antigen so that when taken up by certain cells, those cells produce the same proteins as the antigen [7]. This allows the body to begin an immune response against those protein markers and therefore obtain some level of immunity against the true pathogen. Because the SARS-CoV-2 vaccine, an mRNA vaccine, is so new, its efficacy and action, especially against novel variants of SARS-CoV-2, are still being explored [1]. It is thought to lower viral load and therefore transmissibility, as well as possibly increase the rate of live virus clearance [1]. Because it does not confer complete immunity, infections despite vaccination are common [1]. The vaccine is recommended to be given at least twice, but even with both doses, its reduction of transmission of the virus, as well as the titer of antibodies it creates, decreases over time [1]. It is also more effective against the earlier, less infectious alpha variant of SARS-CoV-2 than the later delta variant of the virus [1]. In essence, SARS-CoV-2 vaccination does not guarantee a positive antibody test, immunity to infection, or a lack of transmissibility. Therefore, its impact on antibody status will be investigated alongside the other factors in this study.

Coding Schemes for Categorical Variables

The antibody test used in this study provides four levels of the factor, which are “Neg”, “IgG”, “IgM”, and “IgM and IgG”. This results in categorical data (without inherent numerical value) for both the regressors and the response variable [8]. Because of these non-numeric and

non-ordered variables (nominal data), creating indicator variables, otherwise known as coding, is necessary. Each level of the categorical variable must be assigned a numerical value to make regression possible. However, one cannot simply assign, say, the integers to the levels of the variable, as this would imply a hierarchical structure with consistent step sizes between the effects [8]. Instead, there are many ways to code variables, each with different goals and its own unique application, with some popular methods being dummy coding, effect coding, and contrast coding. However, due to limited time, only simple dummy coding will be reviewed here.

Simple coding is useful when a variable can take on only two values, such as the survey questions in this study, and is therefore known as a dichotomous variable [9]. This is easily incorporated into the regression model as a binary variable, which has a value of 0 or 1, indicating whether a factor is present or absent [8]. This means that, conceptually, the regression model could take the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$

with x_1 being the numeric regressor and x_2 being the categorical regressor. Then the predicted value of y is affected by the $\beta_2 x_2$ term only if x_2 takes on a value of 1 (indicating the presence of the given factor). When x_2 is zero, the $\beta_2 x_2$ term would have no effect on the predicted value of y. This results in two separate regression lines being formed in parallel [8].

This can also be extended to a categorical variable that can take on more than two values. That involves creating multiple “dummy” variables. There will be one less dummy variable than the number of levels the categorical variable can take on, meaning all but one of the levels of that categorical variable will be assigned its own dummy variable. When the given level of the

original variable is present, the dummy variable associated with that level will take on a value of 1, and all the other dummy variables will take a value of 0 [10]. The level that does not have its own dummy variable will be signified by all the dummy variables assuming a value of 0. This level is now called the reference group. All other levels are compared to this level of the factor [9]. This type of coding may be necessary for the upcoming data analysis.

Generalized Linear Models

Many options are available to analyze this study's data after a coding scheme has been selected, including regression and hypothesis tests. However, the response variable is categorical, meaning that it inherently does not meet the normality assumption of simple linear regression. Instead, a form of a generalized linear model (GLM) must be employed. GLMs are nonlinear regression models that function as a mixture of linear and nonlinear models and allow response variables that are non-normal [8]. They are used in cases with a univariate response variable that may be binary, discrete, highly skewed, or have some other reason to grossly violate the normality assumption beyond the help of a simple data transformation [11]. As well, the response variable distribution must come from the exponential family. This family has the form

$$f(y_i, \theta_i, \phi) = e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + h(y_i + \phi)} \tag{2}$$

where ϕ is the dispersion (or scale) parameter and θ_i is the "natural location parameter" [8]. Members of this family include the normal, Poisson, binomial, inverse normal, exponential, and gamma functions, among others [11].

Link Functions

A GLM uses a link function g to transform otherwise binomial data into a continuous distribution so that correlations in the data can be better explored [12]. It connects the mean of a distribution to a linear predictor [11]. The linear predictor and subsequent expected value of the response variable are defined by equations 3 and 4 [8].

$$n_i = g[E(y_i)] = g(\mu_i) = \boldsymbol{x'\beta} \tag{3}$$

$$E(y_i) = g^{-1}(n_i) = g^{-1}(\boldsymbol{x'\beta}) \tag{4}$$

Each member of the exponential family of distributions has a natural link, called the canonical link, which arises from the relationship $n_i = \theta_i = \boldsymbol{x'\beta}$ [11]. The most common distributions and canonical links are shown in Table VI in the appendix [8]. Often GLMs function very similarly to variance-stabilization transformations. However, the link function uses the natural distribution of the response and is a transformation on the population mean, not the data.

Logistic Regression

Logistic regression (also known as logit) has a link function with the form

$$n_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \tag{5}$$

where π_i is the probability of success in the binomial distribution [8]. This form has many applications in biology, the social sciences, and medicine [11]. It maps π_i, which ranges from 0 to

1, to the real number line, and is therefore popular in modeling binomial data or when there is strong empirical evidence for an S-shaped monotonous function [8]. These properties make logistic regression the most likely candidate for this SARS-COV-2 antibody study.

In logistic regression, the y_i are assumed to be independent of one another and conforming to the equation

$$E(y_i) = \pi_i = \frac{e^{\boldsymbol{x'\beta}}}{1 + e^{\boldsymbol{x'\beta}}} \tag{6}$$

[11]. This model seeks to view the mean as a function of the regressors because the mean of the response itself is the parameter π_i in the distribution [8]. This is easily linearized to produce the logistic link, as shown above. It solves problems arising from Bernoulli response variables, specifically by allowing residuals to take on a range of values and confining the regression to the [0,1] range [8]. Therefore, the formula for logistic regression would become

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_i x_i \tag{7}$$

which can be used with MLEs to find the parameters [8]. The anti-log of the difference in log-odds between x_i and x_{i+1} returns a ratio "interpreted as the estimated increase in the odds of success associated with a one-unit change in the value of the predictor variable" [11].

Logistic regression can be broadened to a response variable with more than two nominal categorical outcomes. In this case, a baseline category is chosen, and a logit is produced for all other categories with respect to the baseline [16]. If the categorical outcomes are ordinal, cumulative probabilities are used, with only intercepts changing per response level [8].

Other Common Link Functions

Another important link for noncontinuous data is the Poisson link. It is very useful in the field of biostatistics because many clinical trials make use of count data (data made of positive integers representing the number of cases or outcomes), which follows the Poisson distribution [13]. There are also many useful link functions for continuous numerical data. For example, the identity link is chosen when the data follows a normal distribution. However, it was previously stated that GLMs are used when the data is non-normal. This leads to the fact that simple linear regression is just a special case of the GLM where the data does not need to be transformed by a link function because its 'link' is just the mean [11]. Another link, the reciprocal link, is often applied to data that follows a gamma distribution (which includes the exponential distribution as well). Data that is strictly positive and strongly right-skewed is often fit well by these gamma distributions. Therefore, the reciprocal link for gamma and exponential distributions is commonly used in survival analysis in clinical studies [14].

The link functions discussed up until this point are canonical, which simplify computations and are natural choices for many common distributions [11]. However, it is also appropriate to consider noncanonical link functions. The most common are the probit, power family, log-log, and complementary log-log links [8]. The probit link is used when "the outcomes are evenly distributed 0 and 1s, and the modeled probabilities are symmetric and expected to follow the Normal cumulative distribution function" [15]. The probit link is represented by

$$n_i = \phi^{-1}[E(y_i)] \tag{8}$$

where ϕ is the cumulative standard normal distribution function [11]. A useful application of probit regression is in toxicological studies with dose-response curves [15].

In binary cases where the outcomes are asymmetric (meaning either 0 or 1 is much more common), the log-log or complementary log-log functions may be more appropriate, depending on whether the probabilities approach 0 or 1 more quickly [15]. The links used are $-\ln[-\ln(\theta)]$ and $\ln[-\ln(1-\theta)]$ for log-log and complementary log-log, respectively [15]. Both of these links, as well as the probit link, are potentially quite useful in the study on SARS-COV-2 antibodies and will be tested in the data analysis stage of the study.

Finally, the power family link is a general formula for finding a link function that helps to reduce the time needed to search for the best link to use [11]. It follows the formula

$$\eta_i = \begin{cases} E(y_i)^{\lambda}, & \lambda \neq 0 \\ \ln[E(y_i)], & \lambda = 0 \end{cases} \tag{9}$$

[8]. λ can be found using an iterative approach to maximum likelihood estimators until convergence is reached. Both the identity link and the log link are special cases of the power link, along with the square root, negative square root, and reciprocal negative square links [11].

Data Cleaning

Data collected by hand on human participants is subject to transcription errors, incomplete data points, and non-responses. To mitigate harmful effects, there are many candidates for statistical techniques to clean the data. Most current research in this area focuses on cleaning big data, but the principles can easily be applied to a smaller dataset, such as the one driving this

study [17]. In fact, the young field of data cleaning has many applications beyond the simple errors mentioned above but can be useful in all sorts of data collection and usage scenarios [18].

Anomalies in the data may be syntactic (concerning the format or variables), semantic (violated constraints, duplicated, or inconsistencies), or coverage anomalies (missing values or tuples) [17]. To address any of these, a combination of automatic processes and domain knowledge must be employed, and each type of error must be handled differently [17]. For example, parsing a value (checking if it conforms to the schema for its tuple) and replacing it with the minimal edit distance is effective at remedying syntax errors [17]. As well, data transformation and duplicate detection are applicable to certain syntactic and semantic anomalies [17]. Other more complex methods include integrity constraint enforcement and statistical methods such as outlier detection [17]. In addition, if a certain part of the data seems incorrect or has too many missing values, there are cases where it is permissible to create a prediction model to predict the missing data [18]. All of these methods fall under one of the following types of methods of error detection: statistical, clustering, pattern-based, and association rules [18]. Currently, software tools are being developed to accurately clean data without requiring much of an expert's time [18].

Methods

Study Design

In the spring semester of 2021, a two-month observational study was conducted by Dr. David Rockabrand and Dr. David Dewitt to investigate the prevalence of SARS-CoV-2 antibodies in Liberty University undergraduate students. The study consisted of a short questionnaire about

vaccination status, symptoms of illness, demographics, and more, as well as a SARS-CoV-2 antibody test. The data was collected from each student twice. The first collection was from February 8th, 2021 - February 11th, 2021, and the second began on April 14th, 2021, and went until April 22nd, 2021. The IRB approval for this study can be found under "IRB-FY20-21-140 SARS-CoV-2 Serologic Response from a Sample of Liberty University Students."

Participants

The participants in this study were 107 undergraduate Liberty University students enrolled in select lower-level biology lab courses. They were invited to participate during a lab period. Participation was entirely voluntary, and every student was eligible regardless of class status, past COVID-19 infection, or other factors. Students could choose not to answer any question even after opting into the study.

Antibody Testing

The method of measuring a student's antibodies was a fingerstick test cassette for SARS-CoV-2 antibodies. The Healgen Covid-19 IgG/IgM Rapid Test Cassette was used. This cassette differentially detects the presence of IgG and IgM-type antibodies against SARS-CoV-2 through the SARS-CoV-2 Spike S1 antigen within it. This antigen can recognize SARS-CoV-2 antibodies and produce a positive test result.

Statistical Analysis

The data fell into three categories: binary answers to infection-related questions, demographic data, and the presence or absence of one or both types of antibodies. First, it was necessary to clean the data, which involved removing subjects who did not have data for both time

points and then standardizing the responses so that the inconsistencies in data recording would not hinder analysis. After this, the data was analyzed using the R statistical programming language.

The analysis began with exploration. This consisted of forming hypotheses about the significance of various factors and then sorting the data into two parts: those participants with the factor present and those with it absent. Then the other meaningful variables could be compared between the two disjoint subsets that together account for all the data. This was accomplished using two-way *t*-tests to compare if the means of the significant variables were the same or different between the set of data with the factor and the set without it. Then, after extensive observations, ANOVAs were run to compare the significance of various factors on different combinations of response variables, with the goal of then creating generalized linear models from the significant factors.

Challenges of Using Imperfect Data

The challenge for this study is to gain meaningful insights from data that was collected without knowledge of its future use. In this case, the problem can be broken into two distinct categories: data cleaning and study design. A variety of anomalies requiring data cleaning were common in this dataset, especially domain format errors, irregularities, and missing values [17]. For example, when IgG and IgM antibodies were present together, research assistants recorded the data point as “both”, “IgG + IgM”, or “IgG, IgM”. An error in the surveys caused COVID test results to be recorded as “positive” or “negative” in February but “Y” or “N” in April. A final example is that some students either chose not to respond to certain questions or were absent for one of the survey dates, resulting in missing data. These anomalies required significant amounts

of time to correct. In the end, 19 data points were removed, but higher-fidelity data resulted.

The second challenge of using imperfect data is that due to time and resource constraints, the study design included some poor sampling techniques, which resulted in unmet assumptions. In this case, all participants were recruited from labs at Liberty University. This means the results will not be as generalizable as they would have been if the samples included all types of undergraduate students, other universities, or simply more participants. Secondly, more frequent sampling and surveying would have allowed a more precise estimation of the time of COVID infection, vaccination, or appearance of antibodies instead of the binary variables that were produced instead. As well, some inconsistencies in the surveys created uncertainty about the dates from which a student should begin reporting COVID tests, contacts, and sickness. Finally, because each participant shared a class with one or more other participants, they not only share certain characteristics but also could have infected one another with COVID-19. This questions the strength of the assumption of independence and therefore affects future tests and models.

Unlike data cleaning, study design cannot be modified retrospectively. Instead, one must stress that the results of the study show correlations and possible connections but alone do not prove causation. Secondly, the results cannot be extended beyond undergraduate Liberty University students. Finally, the unique ways in which questions were asked resulted in a lack of specific pieces of important data, in this case, the dates of COVID infections or vaccination, which would make the results of generalized linear models much more interpretable. This demonstrates both the necessity of informed statistical input even in the pre-design phase of a study, as well as the fact that every study has its limitations.

Results

Seropositivity and Vaccination Rates Among Liberty University Students

In Table I below, seropositivity prevalence for any type of SARS-CoV-2 antibody is broken down by various infection and demographic factors.

TABLE I
CHARACTERISTICS AND SEROPOSITIVITY OF PARTICIPANTS

	Survey 1: Administered 2/2021			Survey 2: Administered 4/2021		
	Total	Seropositive	Percent	Total	Seropositive	Percent
Overall	105	27	26%	105	38	36%
Male	16	7	44%	16	9	56%
Female	89	20	22%	89	28	31%
< Age 20	72	21	29%	72	26	36%
≥ Age 20	33	6	18%	33	11	33%
In a dorm	81	24	30%	81	30	37%
Off Campus	24	3	13%	24	7	29%
Freshman	67	18	27%	67	24	36%
Sophomore	17	7	41%	17	8	47%
Junior	12	1	8%	12	2	17%
Senior	9	1	11%	9	3	33%
Positive COVID Test*	20	9	45%	8	4	50%
Negative COVID Test*	49	10	20%	26	8	31%
No COVID Test*	33	8	24%	71	25	35%
Sick at Least Once**†	19	6	32%	18	6	33%
Never Sick**†	84	21	25%	85	30	35%
Close Contact*††	70	21	30%	49	18	37%
Not a Close Contact*††	26	5	19%	55	19	35%
Vaccinated	4	3	75%	22	15	68%
Unvaccinated	98	24	24%	83	22	27%

Non-responses are not counted
*Survey 1: results from 3/2020 - 2/2021, Survey 2: results from 1/2021 - 4/2021
**Survey 2: results from 11/2020 - 2/2021, Survey 2: results from 1/2021 - 4/2021
†Includes COVID-19 and flu-like symptoms
††Includes contact with individuals positive for COVID-19 or with COVID-19 symptoms

Antibody Fluctuations

Given the longitudinal nature of this study, below in Table II the fluctuations in antibody results over the course of the studies are recorded by participant.

TABLE II
CHANGES IN ANTIBODY RESULTS PER PARTICIPANT

February Result	April Result	Number of Participants	Proportion of Total
Negative	Negative	62	59.6%
Negative	IgG	9	8.7%
Negative	IgM	0	0%
Negative	IgG and IgM	7	6.7%
IgG	Negative	0	0%
IgG	IgG	9	8.7%
IgG	IgM	0	0%
IgG	IgG and IgM	3	2.9%
IgM	Negative	4	3.8%
IgM	IgG	0	0%
IgM	IgM	0	0%
IgM	IgG and IgM	1	1.0%
IgG and IgM	Negative	1	1.0%
IgG and IgM	IgG	2	1.9%
IgG and IgM	IgM	0	0%
IgG and IgM	IgG and IgM	6	5.8%

There is another interesting category that informed the statistical tests to follow. That is those participants who were negative for antibodies in February and then received a COVID test (which may have been positive or negative) before the April survey. This allows further investigation into the correlation between antibodies and a proven recent COVID infection. The results of this investigation can be found in Table VII in the Appendix.

Vaccination Effects

There were 17 participants vaccinated during the course of the study, and below in Table III, their antibody results are given for the beginning and end of the study.

TABLE III
PARTICIPANTS VACCINATED DURING THE STUDY

	Antibody Result in April	
Antibody Result in February	Positive	Negative
Positive	4	1
Negative	7	5

Factors Affecting Seropositivity

Although antibodies are known to fluctuate over time, *t*-tests were conducted to see whether having a history of a confirmed COVID infection or vaccine would cause students to have demonstrable SARS-CoV-2 antibodies at present, and the results are shown in Table VIII in the Appendix. However, as a more robust test, ANOVA was conducted on the data. The data was coded using a binary scheme, showing the presence or absence of each factor. Because of previous analysis, it became clear that time was a significant factor and that the two surveys should be two different variables. Also, because there were more than two results possible for an antibody test, and they were not related numerically, it was decided to treat them separately. ANOVAs were run to find the effects of the variables on the presence of antibodies (treating IgM, IgG, both types, or either type as different ANOVAs).

The ANOVAs analyzed the effect of 10 different factors on the April antibody results, including February antibody results, COVID test results, symptoms of illness, and contact with infected persons. Then, each model was decreased to simply the most important factors (*p*-value

$\leq .05$). Again, each ANOVA only dealt with a specific subset of the antibodies (both for the April dependent variable and the February factor) so that the data could be restricted to binary variables. The sample size in each of the following ANOVA tests is 103 after missing data was removed.

It is important to note that a factor on the April survey occurred after February and a factor on the February survey occurred before February. Unfortunately, exact dating of factors was not available, given the survey design. However, it can be reasonably inferred that the April survey pertains to the prior two months, and the February survey pertains to greater than two months and less than 11 months prior. Hence, for simplicity, April questions will be referred to as "recent" and February as "old." A positive COVID test will also be referred to as a "COVID infection" because although it is understood that Type I and Type II errors abound, there is not a more accurate way of diagnosing a COVID case in this study.

On the following page, in Table IV and Table V, are the ANOVA tables for the final combinations of the most significant factors that affect IgG and IgM antibody status. These tables demonstrate the most remarkable results of the study and draw from the previous results to make informed decisions about what is worth studying. As seen, the models were reduced until only factors with *p*-values below the threshold were included. The ANOVA tables for the presence of both types of antibodies or either type of antibody are less important to the results but can be found in the Appendix, in Table IX and Table X.

Unfortunately, at this point in the analysis, it became clear that the challenges of the dataset made it unfit to explore in the context of generalized linear models. While the various GLMs explored in the literature review could be fit to the data, the results would be

TABLE IV
ANOVA TABLE FOR IGG ANTIBODIES IN APRIL

Factor	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Prior IgG Antibodies	1	8.103	8.103	66.08	1.13e-12
Older COVID-19 Infection	1	1.401	1.401	11.42	0.00103
Recent Vaccination	1	2.359	2.359	19.23	2.84e-05
Residuals	101	12.385	0.123		

TABLE V
ANOVA TABLE FOR IGM ANTIBODIES IN APRIL

Factor	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Prior IgM Antibodies	1	1.847	1.8465	17.151	7.17e-05
Recent COVID-19 Infection	1	0.616	0.6165	5.726	0.01856
Recent Vaccination	1	0.911	0.9109	8.461	0.00446
Residuals	101	10.874	0.1077		

uninterpretable. This is because, with all binary variables, every point is on the extreme of the S-shaped curve, and with no points near the inflection point, there is no way to make predictions as to the presence or absence of antibodies. Therefore, the ANOVA results, showing the most significant factors affecting a student's antibody status, became the chief discovery of this study.

Discussion

Seropositivity and Vaccination Rates Among Liberty University Students

Liberty University undergraduate students were represented in this study to evaluate the prevalence of SARS-CoV-2 antibodies among them. Although at the time of the study Liberty University was operating with in-person classes and on-campus residences, and the COVID-19 outbreak had occurred 11 months prior, the antibody rate among students was relatively low, at just over a quarter of the students having either type of antibody. It rose, as expected during the

course of the semester, to over a third who showed antibodies to SARS-CoV-2. It is tempting to assume that these numbers are the total number of students ever infected with COVID-19 or that they are the progression of the student body to herd immunity. However, as was shown in the literature review and will be demonstrated in the upcoming analyses, antibody levels are less connected to proven infection than previously assumed and frequently decrease, allowing infections to occur again. The factors that influence antibody levels, and therefore what a positive antibody test may mean, will compose the rest of this discussion of results.

When studying vaccination rates, Liberty University undergraduates' rate increased more rapidly in this two-month window than the general US population while staying at a lower proportion overall. The average proportion of people vaccinated (with at least one dose) in the United States from February 8th, 2021, to February 11th, 2021, was 10% [19], compared to this study's participants who had a proportion of only 3.7% in the same time period. However, by the April survey, this study's participants had risen to a vaccination rate of 21.5%. The national average for the same April time interval was 39% [19]. It is understandable that young, healthy university students would have a lower vaccination rate than the national average, which includes many old or immunocompromised citizens. However, this study's vaccination rate increased 1.5 times faster than the increase in the national average, which may be due to the availability of SARS-CoV-2 vaccinations on Liberty University's campus.

Antibody Fluctuations

After analysis of antibody fluctuations, it seems clear that antibody permanency and life-long immunity are not applicable to SARS-CoV-2. It appeared that a participant was slightly

more likely to have IgG antibodies than both types if they began with no antibodies. Yet they never had only IgM at the end. Because, generally, IgM alone is likely only present during an acute COVID infection in which a student would not be in class, these results make sense. Before a student returns to class, their body has already started to create IgG antibodies.

The results, however, become more complex. Most frequently, those with only IgM antibodies returned to a negative result, instead of moving to both types of antibodies as expected. Yet many of those with only IgG changed to having both types of antibodies, which is also unexpected because IgM is said to disappear over time. From these patterns and careful inspection of Table II, it seems that IgM antibodies rise and fall more frequently than IgG. They can disappear or reappear, which the literature did not predict, while IgG seems to persist a bit more steadily, as the literature said. It was also shown that in a surprising proportion of cases, antibodies can decrease to undetectable levels, again proving that SARS-CoV-2 antibodies do not behave similarly to other diseases, whose antibodies persist in the body for many years.

Vaccination Effects

Given the selected vaccination results in Table II, over a third of those vaccinated during the study had no antibodies in April. Although the vaccination date was not recorded, it can be inferred that these participants had been vaccinated for less than or equal to two months. Therefore, the vaccine either caused antibodies that were very short-lived in these participants, or it did not cause antibodies at all. However, it was effective at causing antibodies in other participants. Proportions in this section are not precise because of those vaccinated after already having antibodies. In any case, however, it is clearly demonstrated that the frequency of

COVID-19 infections despite vaccination [1] is not shocking, for many vaccinated participants had no demonstrable levels of SARS-CoV-2 antibodies shortly after vaccination.

Factors Affecting Seropositivity

Raw proportions of students in various categories and when they changed classifications can be informative, as shown. However, the results often lend themselves to erroneous interpretations. Therefore, the methods of analysis in this study moved toward hypothesis tests and ANOVA. The results here are the main conclusions of the study, as they draw from the prior literature and are informed by the analysis that occurred up until this point.

Time-Independent T-Tests

Considering the results of the *t*-tests in Table VIII, it seems that both COVID-19 infections and vaccinations make a difference in a participant having demonstrable levels of SARS-CoV-2 antibodies. It also seems that vaccination status may be more significant in predicting antibody presence than COVID-19 infection history. However, the results were very similar, and vaccinations were a more recent development than COVID-19 infections. Because time is not included in these *t*-tests, it may be that antibodies fade over time, and the results are simply due to uncertainty about when infection or vaccination occurred. Therefore, the rest of the results and discussion focus on breaking down the data by antibody type and time.

IgG Analysis of Variance

Recall that in the ANOVA testing, only one antibody category was analyzed per test, of which IgG and IgM on their own clearly became the most important. Keeping in mind the discussion of the time periods involved, it can be seen that IgG antibodies are significantly

affected by old COVID infections, recent vaccinations, and recent IgG antibodies. First, it is reassuring to see that IgG antibodies seem to persist for greater than two months because this agrees with prior understanding of this type of antibody. Yet the fact that recent infections (within two months) are not a significant factor questions the literature's result that IgG should reach a peak at about one-month post-infection [4].

Secondly, vaccination's effect on antibody results is inconsistent with general understanding. IgG antibodies seemed to appear soon after vaccination but were not correlated with vaccination in the past. This suggests that a vaccine is more effective than a natural infection in providing IgG antibodies quickly but that they do not persist as long as the antibodies from a natural infection. This raises the question as to the methods of vaccine-initiated antibody development, as well as the relative protection from illness between vaccine-initiated and infection-initiated antibodies, but these questions are beyond the scope of this study. Instead, it suffices to say that the COVID-19 vaccine does not behave the same as a natural infection in time of production or duration of IgG antibodies.

IgM Analysis of Variance

Moving to the study of IgM antibodies alone, the ANOVA tests run in the same manner showed that the significant factors were the recent presence of IgM antibodies, recent vaccination, and recent COVID-19 infections. The recency of both vaccination and natural infection is expected in the case of IgM antibodies. It is slightly surprising, then, that this time IgM antibodies in February could predict IgM antibodies two months later because the average persistence of IgM antibodies is said to be only 35 days [4]. Therefore, IgM antibodies may persist longer than

previously postulated. This, combined with the previous antibody fluctuation prevalence results, helps to shed light on the slightly less predictable nature of IgM antibodies. However, they must not last much longer than this study's duration because old COVID infections and old vaccinations were not significant factors in these ANOVAs.

Final Comments

While the GLMs were, in the end, not suitable for this study, they can still be useful in this biostatistical application to SARS-CoV-2 antibodies. If the data had been more extensive in its inclusion of the dates of vaccination and confirmed COVID infections, then there would have been ample opportunity to use GLMs. The issue was the presence of binary independent variables. Even with binary dependent variables, having non-binary independent variables (whether discrete or continuous) can create an interpretable and useful logistic regression curve in a study similar to this one. The issue was, in this case, the lack of specific data.

Concluding this discussion, it must be noted that this study included only undergraduate students at Liberty University. Therefore, it cannot be generalized to the entire population of the country. Instead, it would be safer to conclude that the results are applicable to young, healthy individuals who are frequently exposed to one another and have low rates of complications from SARS-CoV-2. The results are interesting in and of themselves, regardless of their generalizability, and bring to light both agreements and discrepancies with current knowledge of the virus.

Conclusion

To review, a sample of Liberty University undergraduate students were tested for SARS-CoV-2 antibodies twice, two months apart, during the spring semester of 2021. They were

also surveyed at both times for a variety of infection, vaccination, and demographic information. This study undertook analyzing and interpreting the data, both to understand the prevalence of SARS-CoV-2 antibodies in a student population and with the intention of fitting generalized linear models to the data to predict antibody behavior in university students.

The challenge, however, came from the nature of the data. It was collected without knowledge of its future applications, which led to discrepancies in dates and notations, as well as a general lack of certainty about the timing of infections and vaccinations. It required extensive data cleaning, but even so, almost all of the variables were still binary. This meant that generalized linear models, while still technically usable, were now likely uninterpretable. Therefore, the analysis shifted focus to instead pinpointing significant relationships among variables to either support hypotheses that the literature had put forward about the behavior of SARS-CoV-2 antibodies or to formulate new possible correlations.

The results showed that antibody and vaccination proportions were rising among students but stayed below the national levels. However, it was surprising that large proportions of students with confirmed COVID infections or vaccinations did not have detectable antibodies. In fact, the fluctuations in antibody levels demonstrated variable onset times for both antibody types and a shorter antibody duration than other viruses.

IgG antibodies were the most commonly observed and behaved as expected in some aspects. They were likely to be observed two months after an original observation, owing to their status as longer-term antibodies. They were also significantly impacted by older confirmed COVID infections, again as expected due to the time they take to develop. However, a final

significant factor was recent vaccination. And because older vaccinations did not correlate with IgG antibodies, it seems that the COVID-19 vaccines provide shorter-lived antibodies than natural infections. Yet it also means that vaccination caused IgG antibody production much more rapidly than natural infection. In summary, these results suggest quick production and short duration of vaccine-initiated antibodies, a novel idea that deserves future research.

On the other hand, IgM antibodies were also significantly impacted by prior IgM antibody observation, hinting that they persisted over the course of the two months, contradicting the prior assumptions that they disappear rapidly. However, they were also significantly impacted by recent COVID infection and recent vaccination, as would be expected with their association with acute illness. Finally, the presence of both types of antibodies together aided in seeing IgM antibodies' transition to IgG antibodies as time went on.

In conclusion, the university setting is an insightful place to study a rapidly-developing disease such as SARS-CoV-2. The results demonstrate the trends in antibody production and disappearance that have been hypothesized. Not only this, but they helped illuminate the gaps in knowledge about the quality and duration of vaccine-initiated antibodies. It seems that complete immunity is unlikely in this case, but understanding antibody prevalences may aid in the continued aftermath of COVID-19, as well as in informing responses to future diseases.

Future Research

This study brought about many answers but also created many questions. A straightforward next step is to design a similar study that, instead of only having binary responses to questions about COVID infections and vaccines, would instead record the dates that these

events occurred. In this way, the problem of non-interpretable GLMs would be solved, and the analysis could show when the IgM and IgG antibodies rise and fall and how those timelines differ between vaccine-initiated antibodies and infection-initiated antibodies. As well, an increase in sample size, the use of more robust sampling techniques, and the inclusion of more than one US university would greatly improve the statistical power of the tests in use, thereby causing the results to hold more weight. These changes would allow even more insightful conclusions that could greatly further research in this rapidly-developing field.

References

[1] D. W. Eyre *et al.*, "Effect of Covid-19 Vaccination on Transmission of Alpha and Delta Variants," *New England Journal of Medicine*, vol. 286, no. 12, February, 2022. [Online Serial]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa2116597. [Accessed Dec. 19, 2022].

[2] R. L. Tillett *et al.*, "Genomic evidence for reinfection with SARS-CoV-2: a case study," *The Lancet Infectious Diseases*, vol. 21, no. 1, pp. 52-58, October, 2020. [Online Serial]. Available: https://doi.org/10.1016/S1473-3099(20)30764-7. [Accessed Jan. 9, 2023].

[3] Centers for Disease Control and Prevention, "Risk for COVID-19 infection, hospitalization, and death by age group," *cdc.gov*, Nov. 8, 2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html. [Accessed Dec. 19, 2022].

[4] J. Kopel, H. Goyal, and A. Perisetti, "Antibody tests for COVID-19," *Baylor University Medical Center Proceedings*, vol. 34, no. 1, pp. 63-72, October, 2020. [Online Serial]. Available: https://www.tandfonline.com/doi/full/10.1080/08998280.2020.1829261. [Accessed Jan. 5, 2023].

[5] D. Jacofsky, E. M. Jacofsky, and M. Jacofsky, "Understanding Antibody Testing for COVID-19," *The Journal of Arthroplasty*, vol. 35, no. 7, pp. S74-S81, April, 2020. [Online Serial]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7184973/. [Accessed Jan. 5, 2023].

[6] B. Pulendran and R. Ahmed, "Immunological mechanisms of vaccination," *Nature Immunology*, vol. 12, pp. 509-517, May, 2011. [Online Serial]. Available: https://www.nature.com/articles/ni.2039?report=reader. [Accessed Jan. 5, 2023].

[7] F. P. Polack *et al.*, "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine," *The New England Journal of Medicine*, vol. 383, pp. 2603-2615, December, 2020. [Online Serial]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa2034577. [Accessed Jan. 5, 2023].

[8] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. Hoboken, NJ: Wiley, 2021.

[9] W. W. Daniel and C. L. Cross, *Biostatistics: A Foundation for Analysis in the Health Sciences*. Hoboken, NJ: Wiley, 2018.

[10] X. Chen, P. Ender, M. Mitchell, and C. Wells, "Additional coding systems for categorical variables in regression analysis" in *Regression with SPSS*. UCLA, 2011. [Online] Available: https://stats.oarc.ucla.edu/sas/webbooks/reg/chapter5/regression-with-saschapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/.

[11] R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson, *Generalized Linear Models: With Applications in Engineering and the Sciences*, ed. 2. Hoboken, NJ: Wiley, 2010.

[12] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, ed. 4. London, UK: Routledge, 2019.

[13] M. J. Hayat and M. Higgins, "Understanding poisson regression," *Journal of Nursing Education*, vol. 53, no. 4, pp. 207-215, March, 2014. [Online Serial]. Available: https://pubmed.ncbi.nlm.nih.gov/24654593/. [Accessed Dec. 20, 2022].

[14] G. Grover, A. S. A. Sabharwal, and J. Mittal, "An application of gamma generalized linear model for estimation of survival function of diabetic nephropathy patients," *International Journal of Statistics in Medical Research*, vol. 2, no. 3, pp. 209-219, July, 2013. [Online Serial]. Available: http://dx.doi.org/10.6000/1929-6029.2013.02.03.6. [Accessed Dec. 20, 2022].

[15] J. D. Canary, L. Blizzard, R. P. Barry, D. W. Hosmer, and S. J. Quinn, "Summary goodness-of-fit statistics for binary generalized linear models with noncanonical link functions," *Biometrical Journal*, vol. 58, no. 3, pp. 674-690, May, 2016. [Online Serial]. Available: https://doi.org/10.1002/bimj.201400079. [Accessed Dec. 20, 2022].

[16] A. Agresti, *Foundations of linear and generalized linear models*, Hoboken, NJ: Wiley, 2015.

[17] H. Müller and J. C. Freytag, "Problems , Methods , and Challenges in Comprehensive Data Cleansing," *Professoren des Inst. Für Informatik*, 2005. [Online Serial]. Available: https://tarjomefa.com/wp-content/uploads/2015/06/3229-English.pdf. [Accessed Dec. 20, 2022].

[18] J. I. Maletic and A. Marcus, "Data cleansing: A prelude to knowledge discovery," in *Data mining and knowledge discovery handbook*. Boston, MA: Springer, 2009. [Online] Available: https://link.springer.com/chapter/10.1007/978-0-387-09823-4_2.

[19] Covid Act Now, "Data API," *Act Now Coalition*. [Online]. Available: https://covidactnow.org/data-api. [Accessed Nov. 7, 2021].

Additional References

R.V. Hogg, E. A. Tanis, and D. Zimmerman, *Probability and Statistical Inference*, ed. 10. Hoboken, NJ: Pearson, 2020.

Appendix

TABLE VI
CANONICAL LINKS FOR GENERALIZED LINEAR MODELS

Distribution	Canonical Link
Normal	$\eta_i = \mu_i$ (identity link)
Binomial	$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ (logistic link)
Poisson	$\eta_i = \ln(\lambda)$ (log link)
Exponential / Gamma	$\eta_i = \frac{1}{\lambda_i}$ (reciprocal link)

TABLE VII
PARTICIPANTS NEGATIVE IN FEBRUARY THAT RECEIVED A COVID TEST BEFORE APRIL

	Antibody Result in April	
Result of COVID Test	Positive	Negative
Positive	3	4
Negative	3	17

TABLE VIII
WELCH'S TWO-SAMPLE T-TESTS FOR DIFFERENCES IN ANTIBODY PREVALENCE

Factors	Result	p-value
Positive COVID-19 Test vs. Not	Significant Difference	0.015
Vaccinated vs. Not	Significant Difference	< 0.001

TABLE IX
ANOVA TABLE FOR BOTH ANTIBODY TYPES PRESENT IN APRIL

Factor	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Prior Both Antibodies Present	1	2.508	2.5080	24.919	2.5e-12
Recent COVID-19 Infection	1	0.592	0.5919	5.881	0.01709
Recent Vaccination	1	0.982	0.9824	9.761	0.00233
Residuals	101	10.165	0.1006		

TABLE X
ANOVA TABLE FOR EITHER ANTIBODY TYPE PRESENT IN APRIL

Factor	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Prior Antibodies Present	1	6.117	6.117	42.986	2.37e-09
Older COVID-19 Infection	1	1.401	1.401	9.843	0.00223
Recent Vaccination	1	2.359	2.359	16.575	9.30e-09
Residuals	101	14.372	0.142		