MEMORANDUM

To: David J. Wright and Tunde Akinseye

From: Ben Laufer, Sam Ricafrente, Lucas Fonda, and Wyatt De Mers

Date: April 24, 2024

Re: SARS-CoV-2 Seroprevalence in US Blood Donors Statistical Analysis Recommendation

---

The purpose of this memo is to describe the statistical methods and findings from an analysis of your SARS-CoV-2 Seroprevalence in US Blood Donors Data. We hope that this information helps you address your research question:

> *"What variations exist in the detection rates of observed SARS-CoV-2 seroconversion through nucleocapsid antibody assays across demographic and health-related predictor variables?"*

This memo is organized into four sections.

I. **Abstract of Key Findings** – an overview of key results from the analysis.
II. **Background and Data** – a summary of our understanding of your research questions and basic descriptive statistics to get an overview of your data including variables measured and how the data was collected.
III. **Statistical Methods** – a description of the models and methods.
IV. **Results and Discussion** – numerical and graphical summaries, interpretation of results, and limitations
V. **Technical Output** – varied computer output for reference

If you have any additional questions about this work following our consulting meeting today, please feel free to contact us at sricafre@calpoly.edu so that we may set up another meeting to discuss your questions.

I. Abstract

We present an analysis of SARS-CoV-2 seroconversion detection among a cohort of blood donors across the United States, using data collected from December 2021 through 2022. Our analysis aimed to understand how factors such as age, gender, ethnicity, geographic location, and vaccination status affect the probability of donors self-reporting a swab within the seroconversion interval. Utilizing logistic regression without interactions, we found that age and geographic location significantly influenced self-swabbing rates, with older donors and donors from the Midwest demonstrating higher odds of reporting. In addition, males were less likely to report seroconversion compared to females, and vaccinated individuals had lower self-reporting rates than their unvaccinated counterparts. Black donors reported at higher rates than other racial groups. Urban donors were less likely to report seroconversion compared to rural donors. These patterns suggest notable behavioral differences in the cohort's response to potential SARS-CoV-2 infection. However, the study's limitations—including potential volunteer bias, self-reported vaccination data, and exclusion of booster dose effects—must be considered, as they could affect the interpretation and generalizability of these findings. Despite these constraints, the results offer valuable insights into seroconversion detection behaviors in blood donors.

II. Background and Data

Our understanding is that you seek assistance in selecting the appropriate statistical method, analyzing data, and interpreting the results to address your main research question: *"What variations exist in the detection rates of observed SARS-CoV-2 seroconversion through nucleocapsid antibody assays across demographic and health-related predictor variables?"* To investigate this question, we are utilizing the data you provided from the study conducted by the CDC in the National Blood Donor Cohort. We understand that the data was collected across all 50 states from December 2021 until the end of 2022 by the National Blood Donor Cohort. Additionally, there was retrospective SARS-CoV-2 Ab data that was collected from June 1, 2020, until June 30, 2021, from universal screenings of donations. Donors were required to fill out a survey quarterly that was aimed at collecting information about their known infection, infection outcome, and vaccination history. Throughout the study, a total of 142,612 donors were followed, organized into various cohorts based on their infection and vaccination status, including Not Infected, Vaccinated (NIV); Not Infected, Not Vaccinated (NINV); Infected, Vaccinated (IV); and Infected, Not Vaccinated (INV).

We recognize that there are two datasets used, that being the primary infection (PI) and reinfection (RI) datasets. Each dataset includes the same variables, with a detailed listing provided later in the memorandum. The primary infection dataset, with approximately 33,000 individuals, includes donors with known primary infection seroconversion donation intervals. The reinfection dataset, with approximately 10,000 individuals, includes only those with known reinfection seroconversion donation intervals. We understand that the response variable is the probability of self-swabbing within the seroconversion interval based on the multitude of predictor variables outlined above.

In order to be able to effectively analyze the difference between the PI and RI dataset, we decided to merge the two datasets. This was done by using the left join function in SAS studio. With this function we combined the two datasets into one in which we matched each observation with its corresponding patient id (pat_id). Additionally, we created a new variable (called data_source) that classifies which dataset the observation came from (labeled either PI or RI).

From this, we were able to see if any key differences with blood donors between the two datasets.

Below we have listed the variables that have been provided to us and that are to be included in the analyses.

- pat_id: Donor identifier
- age:
    - 1 = 16-29
    - 2 = 30-49
    - 3 = 50-64
    - 4 = 65+
- gender: Male, Female
- race2: Race/ethnicity in 5 categories
- dhq_vaccinated:
    - 0 = Unvaccinated
    - 1 = Vaccinated prior to seroconversion interval
    - 2 = Vaccinated during seroconversion interval
- Census Region: 4 census regions, plus unknown/missing
- urban_rural:
    - 1 = Urban
    - 2 = Rural
    - 3 = Missing
- true_sero/ri_true_sero:
    - 0 = No self-reported swab within seroconversion interval
    - 1 = Yes, self-reported swab within seroconversion interval

To help better understand the data set, we analyzed descriptive statistics to summarize the characteristics of the blood donor data. The majority of the donors are middle-aged, with the largest age group of 40-59 years comprising 41.68% of the cohort. The gender distribution is almost equal but slightly favors males, making up 54.48% of participants. In terms of race and ethnicity, the majority of donors are Non-Hispanic White, accounting for 91.89%. We also found

that a significant portion of the cohort was vaccinated prior to the seroconversion interval, approximately 46.94%. Additionally, the urban-rural classification shows a predominant urban residency, with 77.57% of participants living in urban areas. These percentages provide an overview of the demographics and characteristics of our study population.

You were also interested in receiving several deliverables for your manuscript. This includes creating a journal-quality table of results. We also understand that odds ratios for missing levels and vaccinated individuals within the seroconversion interval should be excluded from the table but included in the model. Additionally, you require a well-articulated paragraph for the methods section of the manuscript, providing a comprehensive description of the model used for analysis. It is also our understanding that we will perform model checking and assess whether the logistic regression assumptions are met by the dataset. Lastly, we aim to identify any limitations of the model that should be acknowledged and discussed for the manuscript's discussion section.

III. Statistical Methods

As per your request, we have run a **logistic regression model** using SAS. This model allows us to estimate the odds of self-swabbing within the seroconversion interval among blood donors.

The **final model** can be expressed as:

$$log_e[\frac{\pi(x)}{1-\pi(x)}] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14} + \beta_{15} x_{15} + \beta_{16} x_{16} + \beta_{17} x_{17} + \varepsilon$$

where $\pi(x)$ = probability of self-swabbing within the seroconversion interval
α = log odds of self-swabbing within the seroconversion interval when all predictors are at their reference level
x_1 = gender (binary: male=0, female=1)
x_2 = age category 1 (16-29=1, otherwise=0)
x_3 = age category 2 (30-49=1, otherwise=0)
x_4 = age category 3 (50-64=1, otherwise=0)
x_5 = ethnicity Hispanic (Hispanic=1, otherwise=0)
x_6 = race non-Hispanic Asian (Asian=1, otherwise=0)
x_7 = race non-Hispanic Black (Black=1, otherwise=0)
x_8 = race non-Hispanic Other (Other=1, otherwise=0)
x_9 = race non-Hispanic White (White=1, otherwise=0)
x_{10} = census region 1 (Northeast=1, otherwise=0)
x_{11} = census region 2 (Midwest=1, otherwise=0)
x_{12} = census region 3 (South=1, otherwise=0)
x_{13} = census region 4 (West=1, otherwise=0)
x_{14} = DHQ vaccinated status 0 (unvaccinated=1, otherwise=0)
x_{15} = DHQ vaccinated status 1 (vaccinated during=1, otherwise=0)
x_{16} = urban rural classification 1 (urban=1, otherwise=0)
x_{17} = urban rural classification 2 (rural=1, otherwise=0)
ε = error term

For our analysis, all 33,604 participants from the data set were incorporated. The key results from this analysis are that all of the variables in the model are **statistically different** from zero, with all corresponding p-values below the significance level of .05. Therefore, our final model

included **all** of the variables in the dataset. The estimated coefficients and p-values can be found below in Table 1. The coefficients in the logistic regression model are interpreted as the change in **log odds** of seroconversion for a **one-unit change** in the predictor variable, holding other predictor variables constant.

Table 1: Logistic Regression Analysis

		Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
	Intercept	0.2719	0.096	8.087	0.005
Gender	Female	-0.08	0.011	51.022	<0.001
Donor's Age at Baseline	16-29Y	-0.160	0.046	12.317	<0.001
	30-49Y	-0.080	0.024	11.744	<0.001
	50-64Y	0.034	0.021	2.554	0.110
Donor Race/ Ethnicity	Hispanic	-0.086	0.061	2.002	0.157
	Asian	-0.045	0.083	0.298	0.585
	Black	0.511	0.100	26.212	<0.001
	Other	0.023	0.097	0.056	0.811
	White	-0.101	0.042	5.780	0.016
Location (Census Regions)	Northeast	-0.008	0.088	0.009	0.926
	Midwest	0.246	0.086	8.218	0.004
	South	0.143	0.087	2.683	0.101
	West	0.015	0.087	0.029	0.864
Self-Reported Vaccine Status	Not Vaccinated	0.060	0.180	11.145	0.001
	Vaccinated	-0.140	0.180	61.133	<0.001
2013 NCHS Urban-Rural Classification Scheme of Counties	Urban	-0.244	0.028	77.099	<0.001
	Rural	0			
Data Source	PI	-0.073	0.014	26.678	<0.001

In our logistic regression analysis, we opted against using an interaction model to maintain clarity in interpreting the results. We recognized that while interaction terms can sometimes reveal relationships between variables, they also complicate the model structure and interpretation of odds ratios. This added complexity can obscure the clarity of the results without necessarily providing insights of practical relevance to our research questions. The model we chose captures essential trends and patterns as requested, without overburdening the analysis with excessive detail that may not enhance our understanding of the predictors' effects.

In our logistic regression analysis, we operated under several assumptions. We assumed that each predictor variable has a linear relationship with the log odds of the outcome, meaning that as the value of a predictor increases, the log odds of the outcome either consistently increase or decrease, depending on the nature of the relationship. The errors, or the differences between observed and predicted values, were expected to follow a normal distribution, which supports the validity of our statistical tests. Consistent variance of errors across the range of predictor variables was also assumed, known as homoscedasticity. Additionally, we presumed that the observations were independent of each other, meaning the outcome for one individual did not influence that of another. Lastly, we assumed minimal multicollinearity, ensuring that our independent variables were not overly interrelated. These assumptions are important for the interpretability of our logistic regression results.

IV. Results and Discussion

Per your request, below in Table 2 I have provided a journal quality table of unadjusted and adjusted odds ratios (w/ 95% CI) from the PI dataset, as well as the number of each observation found in the dataset, and that rate.

Table 2: Odds Ratios for Seroconversion by Donor Characteristics

Variables	Values		Overall	Unadjusted		Adjusted	
Total		N	Rate/100	OR	95% CI	OR	95% CI
Donor's Gender	Male	15295	45.52%	0.846	0.810, 0.883	0.851	0.814, 0.889
	Female	18309	54.48%				
Donor's Age at Baseline	30-49Y	8114	3.45%	1.129	0.998, 1.277	1.091	0.963, 1.235
	50-64Y	14006	24.15%	1.275	1.075, 1.131	1.228	1.087, 1.386
	65+Y	10323	41.68%	1.458	1.291, 1.647	1.458	1.288, 1.651
Donor Race/ Ethnicity	Hispanic	1316	3.92%	0.848	0.670, 1.073	0.894	0.705, 1.134
	Asian	520	1.55%	0.842	0.642, 1.104	0.938	0.714, 1.232
	Black	357	1.06%	1.672	1.238, 2.260	1.628	1.201, 2.206
	White	30879	91.89%	1.005	0.815, 1.241	0.885	0.715, 1.095
Location (Census Regions)	Midwest	11965	35.61%	1.401	1.312, 1.497	1.296	1.211, 1.386
	South	6549	19.49%	1.204	1.119, 1.296	1.162	1.078, 1.251
	West	10050	29.91%	1.010	0.944, 1.081	1.015	0.947, 1.088
Self-Reported Vaccine Status	Vaccinated Prior	14626	43.52%	0.774	0.740, 0.809	0.787	0.751, 0.825
2013 NCHS Urban-Rural Classification Scheme of Counties	Urban	26067	77.57%	0.718	0.682, 0.756	0.784	0.742, 0.827

We decided to use females, ages 16-29, "other" ethnicities, people living in the Northeast, those that were unvaccinated, and people who live in rural areas as the reference groups for the odds ratios. We chose these as a result of them being the least represented groups for each variable. This was the case for all variables except for females. However, because there were only two

groups for the donor's gender, being male and female, our preference would have made no difference.

As shown in Table 1, there is a significant difference in the odds of self-reporting a swab within the seroconversion interval between donors with primary infection seroconversion donation intervals and donors with known reinfection seroconversion donation intervals. Specifically, the odds of correctly self-reporting a swab within the seroconversion interval for donors with a primary infection is estimated to be 7.04% lower than for donors with a known reinfection. This makes logical sense, as donors who have not already gotten SARS-CoV-2 will not have seen the symptoms firsthand. Ultimately, those donors may not be as weary about getting tested when compared with those who have already contracted SARS-CoV-2. The ones with reinfection have experienced the effects of the disease prior to them self-reporting a positive swab.

The odds of self-reporting a swab within the seroconversion interval is 0.149 times lower for donors who are male compared to donors who are female. The odds of self-reporting a swab within the seroconversion interval is 1.628 times higher for donors who are black compared to a donor with an ethnicity different than the ones listed. The odds of self-reporting a swab within the seroconversion interval is 1.296 higher for donors who are from the Midwest compared to a donor living in the Northeast. An intriguing feature of the data shows that the odds of self-reporting a swab within the seroconversion interval for older donors tends to be higher when compared with younger donors. The odds ratios confidence intervals for both 50 to 64 year olds, as well as 65 plus, are completely above 0, (1.087, 1.386) and (1.288, 1.6510), respectively. Meaning, the odds of self-reporting a swab within the seroconversion interval for those previously mentioned ages is higher than the odds of self-reporting a swab within the seroconversion interval when compared with 16 to 29 year olds. At the beginning of the analyses, we believed that younger donors most likely were not the ones being proactive in their pursuit of figuring out if they had SARS-CoV-2 or not when compared with older donors. Older donors, maybe because of health concerns, appear to be more proactive about self-reporting a swab when they are actually infected. As clearly laid out in the above odds ratios, there are patterns that suggest notable behavioral differences in the cohort's response to potential SARS-CoV-2 infection.

The findings from our analysis may not fully generalize to the broader U.S. population due to differences in health status and demographic representation, particularly among minority groups. The logistic regression analysis used assumes each response is independent and follows a binomial distribution, which may be compromised in our non-randomly selected, volunteer-based sample. Since the data from the study relies on volunteers and self-reported data, we can only suggest associations rather than establish cause and effect. Factors such as the absence of data on booster doses and reinfections, as well as the exclusion of behavioral differences and the order of vaccination relative to infection, introduce additional uncertainties. Additionally, the irregular timing of donations could have introduced inconsistencies in the data collection, limiting our analysis. These limitations highlight the need for cautious interpretation of the patterns observed.

V. Technical Output

Model fit and goodness of fit test using PI DATASET

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	57.4	Somers' D	0.149
Percent Discordant	42.5	Gamma	0.149
Percent Tied	0.1	Tau-a	0.075
Pairs	281871604	c	0.575

Partition for the Hosmer and Lemeshow Test					
		true_sero = 0		true_sero = 1	
Group	Total	Observed	Expected	Observed	Expected
1	3380	1380	1384.26	2000	1995.74
2	3373	1514	1518.81	1859	1854.19
3	3364	1571	1587.91	1793	1776.09
4	3360	1698	1655.08	1662	1704.92
5	3367	1691	1714.80	1676	1652.20
6	3362	1811	1769.37	1551	1592.63
7	3364	1837	1831.35	1527	1532.65
8	3366	1855	1901.67	1511	1464.33
9	3358	1993	1984.36	1365	1373.64
10	3310	2112	2114.36	1198	1195.64

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
8.0955	8	0.4242

Linearity using PI DATASET

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	46535.171	45982.752
SC	46543.594	46134.355
-2 Log L	46533.171	45946.752

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	586.4198	17	<.0001
Score	581.3127	17	<.0001
Wald	572.0149	17	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
gender	1	47.9730	<.0001
race2	5	33.7524	<.0001
Census_Region	4	96.8402	<.0001
urban_rural	1	76.4274	<.0001
age_cat	3	6.9364	0.0740
dhq_vaccinated	2	113.7289	<.0001
age_yrs2	1	22.1078	<.0001

MAIN EFFECTS PLOT WITH vaccinated to unvaccinated people using PI DATASET

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	46535.171	46405.712
SC	46543.594	46430.980
-2 Log L	46533.171	46399.712

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	133.4590	2	<.0001
Score	133.3938	2	<.0001
Wald	133.2124	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
dhq_vaccinated	2	133.2124	<.0001

Analysis of Maximum Likelihood Estimates

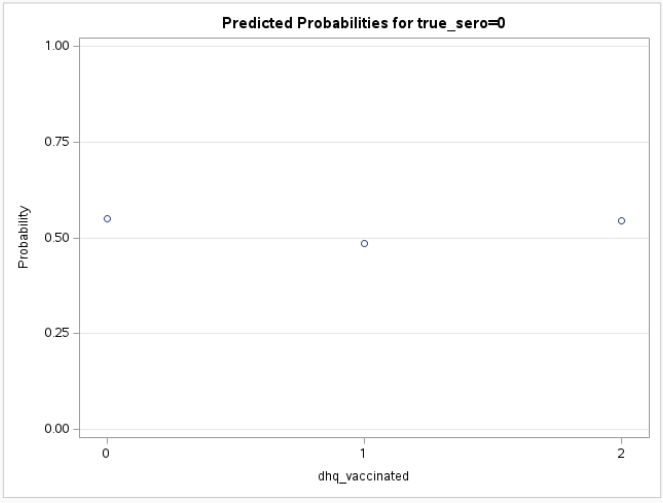
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.1086	0.0141	59.3667	<.0001
dhq_vaccinated	0	1	0.0928	0.0171	29.6226	<.0001
dhq_vaccinated	1	1	-0.1638	0.0168	94.6869	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dhq_vaccinated 0 vs 2	1.022	0.947	1.104
dhq_vaccinated 1 vs 2	0.791	0.733	0.853

Association of Predicted Probabilities and Observed Responses

Percent Concordant	32.3	Somers' D	0.064
Percent Discordant	25.9	Gamma	0.109
Percent Tied	41.7	Tau-a	0.032
Pairs	281871604	c	0.532



Logistic Regression for combined RI/PI Dataset

Analysis of Maximum Likelihood Estimates (Standard without interaction effects) (Combined RI/PI Dataset)						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.2719	0.0956	8.0871	0.0045
gender	female	1	-0.08	0.0112	51.0218	<.0001
age_cat	1	1	-0.1599	0.0455	12.3167	0.0004
age_cat	2	1	-0.0804	0.0235	11.7441	0.0006
age_cat	3	1	0.0337	0.0211	2.5544	0.11
race2	hispanic_ethnicity	1	-0.0859	0.0607	2.0018	0.1571
race2	non-hispanic asian	1	-0.0452	0.0827	0.2984	0.5849
race2	non-hispanic black	1	0.5108	0.0998	26.2122	<.0001
race2	non-hispanic other	1	0.0231	0.0965	0.0575	0.8106
race2	non-hispanic white	1	-0.1014	0.0422	5.7796	0.0162
Census_Region	Region 1 (Northeas	1	-0.00821	0.0879	0.0087	0.9256
Census_Region	Region 2 (Midwest)	1	0.2462	0.0859	8.2178	0.0041
Census_Region	Region 3 (South)	1	0.143	0.0873	2.6828	0.1014
Census_Region	Region 4 (West)	1	0.0148	0.0867	0.0292	0.8643
dhq_vaccinated	0	1	0.0595	0.0178	11.1446	0.0008
dhq_vaccinated	1	1	-0.1398	0.0179	61.1329	<.0001
urban_rural	1	1	-0.2441	0.0278	77.0994	<.0001
urban_rural	2	0	0	.	.	.
data_source	PI	1	-0.0732	0.0142	26.6779	<.0001

Adjusted Odds Ratios for combined RI/PI Dataset

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gender female vs male	0.851	0.814	0.889
age_cat 2 vs 1	1.091	0.963	1.235
age_cat 3 vs 1	1.228	1.087	1.386
age_cat 4 vs 1	1.458	1.288	1.651
race2 hispanic_ethnicity vs non-hispanic other	0.894	0.705	1.134
race2 non-hispanic asian vs non-hispanic other	0.938	0.714	1.232
race2 non-hispanic black vs non-hispanic other	1.628	1.201	2.206
race2 non-hispanic white vs non-hispanic other	0.885	0.715	1.095
race2 unavailable vs non-hispanic other	0.722	0.502	1.040
Census_Region Region 2 (Midwest) vs Region 1 (Northeast)	1.296	1.211	1.386
Census_Region Region 3 (South) vs Region 1 (Northeast)	1.162	1.078	1.251
Census_Region Region 4 (West) vs Region 1 (Northeast)	1.015	0.947	1.088
Census_Region Uncoded vs Region 1 (Northeast)	0.979	0.430	2.231
dhq_vaccinated 1 vs 0	0.787	0.751	0.825
dhq_vaccinated 2 vs 0	1.005	0.930	1.087
urban_rural 1 vs 2	0.784	0.742	0.827

Unadjusted Odds Ratio for AGE Variable

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age_cat 2 vs 1	1.129	0.998	1.277
age_cat 3 vs 1	1.275	1.131	1.438
age_cat 4 vs 1	1.458	1.291	1.647

Unadjusted Odds Ratio for LOCATION Variable

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Census_Region Region 2 (Midwest) vs Region 1 (Northeast)	1.401	1.312	1.497
Census_Region Region 3 (South) vs Region 1 (Northeast)	1.204	1.119	1.296
Census_Region Region 4 (West) vs Region 1 (Northeast)	1.010	0.944	1.081
Census_Region Uncoded vs Region 1 (Northeast)	0.993	0.437	2.255

Unadjusted Odds Ratio for GENDER Variable

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gender female vs male	0.846	0.810	0.883

Unadjusted Odds Ratio for RACE/ETHNICITY Variable

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
race2 hispanic_ethnicity vs non-hispanic other	0.848	0.670	1.073
race2 non-hispanic asian vs non-hispanic other	0.842	0.642	1.104
race2 non-hispanic black vs non-hispanic other	1.672	1.238	2.260
race2 non-hispanic white vs non-hispanic other	1.005	0.815	1.241
race2 unavailable vs non-hispanic other	0.690	0.481	0.991

Unadjusted Odds Ratio for URBAN/RURAL Variable

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
urban_rural 1 vs 2	0.718	0.682	0.756
urban_rural 3 vs 2	0.655	0.289	1.486

Unadjusted Odds Ratio for VACCINATION STATUS Variable

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
dhq_vaccinated 1 vs 0	0.774	0.740	0.809
dhq_vaccinated 2 vs 0	0.978	0.906	1.056