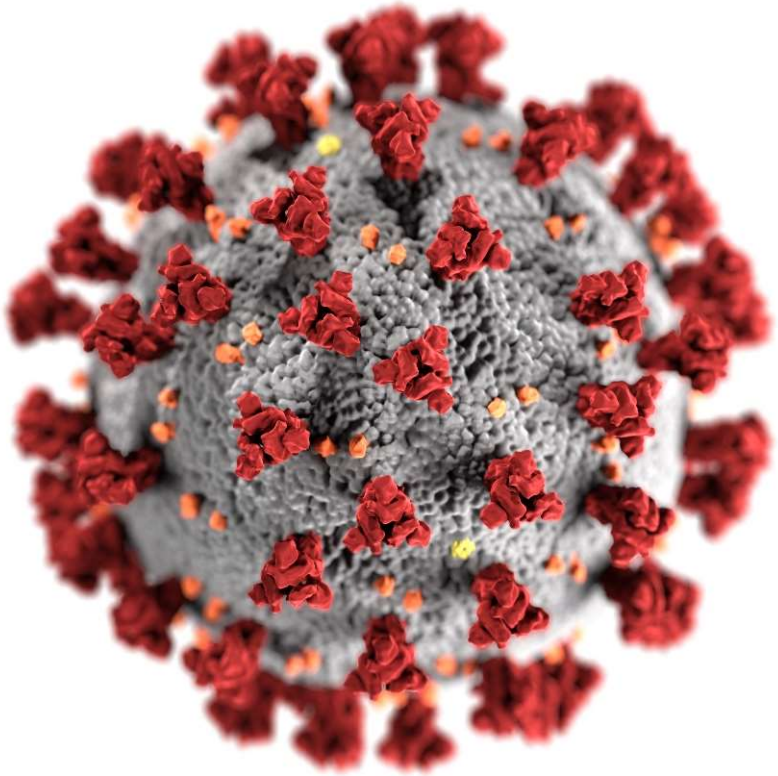
# Representativeness of Non-probability Serology Samples from Multiple Commercial Laboratories in the United States

## 2022 FCSM Research & Policy Conference

*October 26, 2022*


Yun Kim, ICF
Ronaldo Iachan, ICF
Lee Harding, ICF
Adam Lee, ICF
Davia Moyse, ICF
Kristie Clarke, CDC
Kevin Barney, CDC
Ruchi Pancholy, CDC

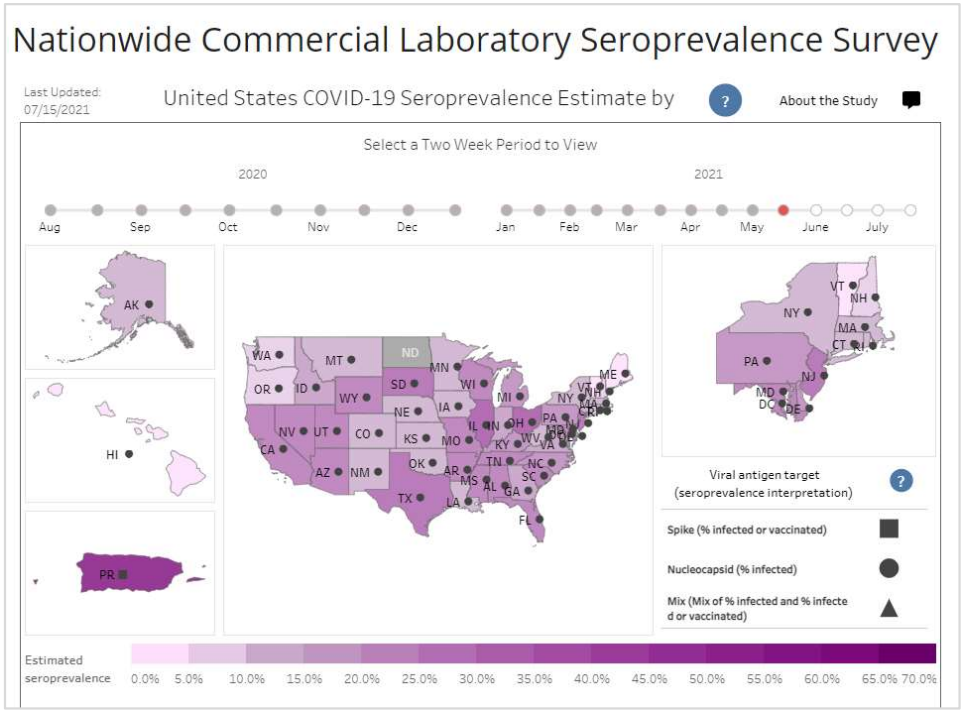



cdc.gov/coronavirus

# Background

- **What is the Nationwide Commercial Laboratory Seroprevalence (NCLS)?**

- CDC's NCLS study is designed to produce weighted estimates for the percentage of people with antibodies to SARS-CoV-2, the virus that causes COVID-19,




**Source:** https://covid.cdc.gov/covid-data-tracker/#national-lab

# Background

- **Objectives of Nationwide Commercial Laboratory Seroprevalence (NCLS) Study**
  - To produce jurisdiction-level estimates weighted to the population totals of each jurisdiction based on available demographics (e.g., age, sex, metro/non-metro status) using residual sera from commercial laboratories across the United States.
  - To produce national estimates by demographic groups and to assess changes over time

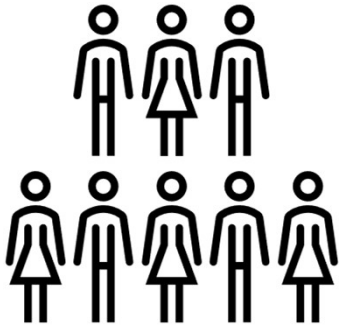
# Background

❑ **Design, Setting, and Participants**

- Included all 50 states, the District of Columbia, and Puerto Rico
- Repeated monthly cross-sectional study (Rounds)
- Used **non-probability samples** of people who visited commercial laboratories for non-COVID-19 related reasons and had blood drawn
- Collected specimens provided by people of all ages originally for routine screening or clinical management from the commercial laboratories
- Excluded people with COVID-19 related conditions

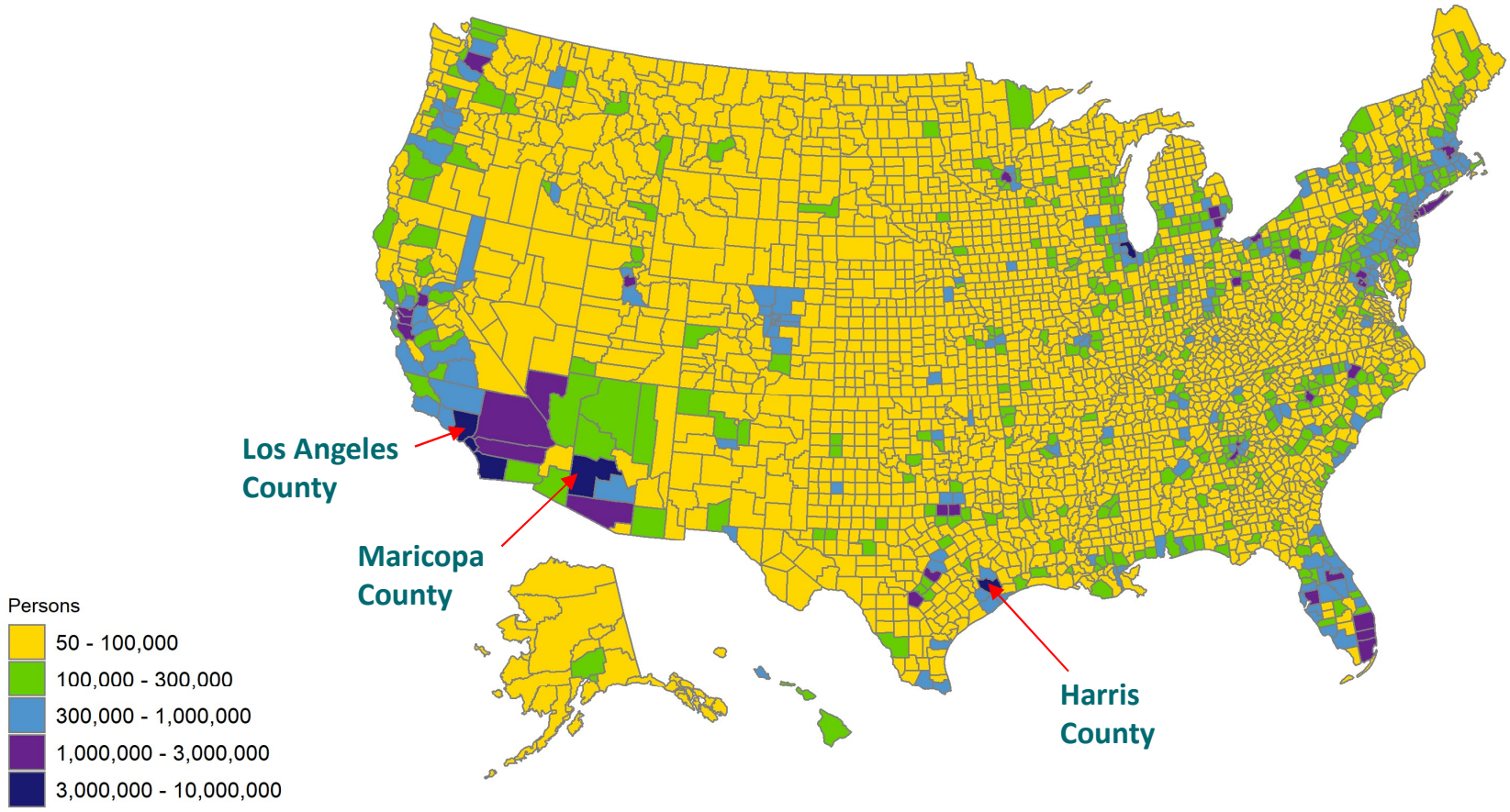
**Sample Representativeness?**

# Research Questions



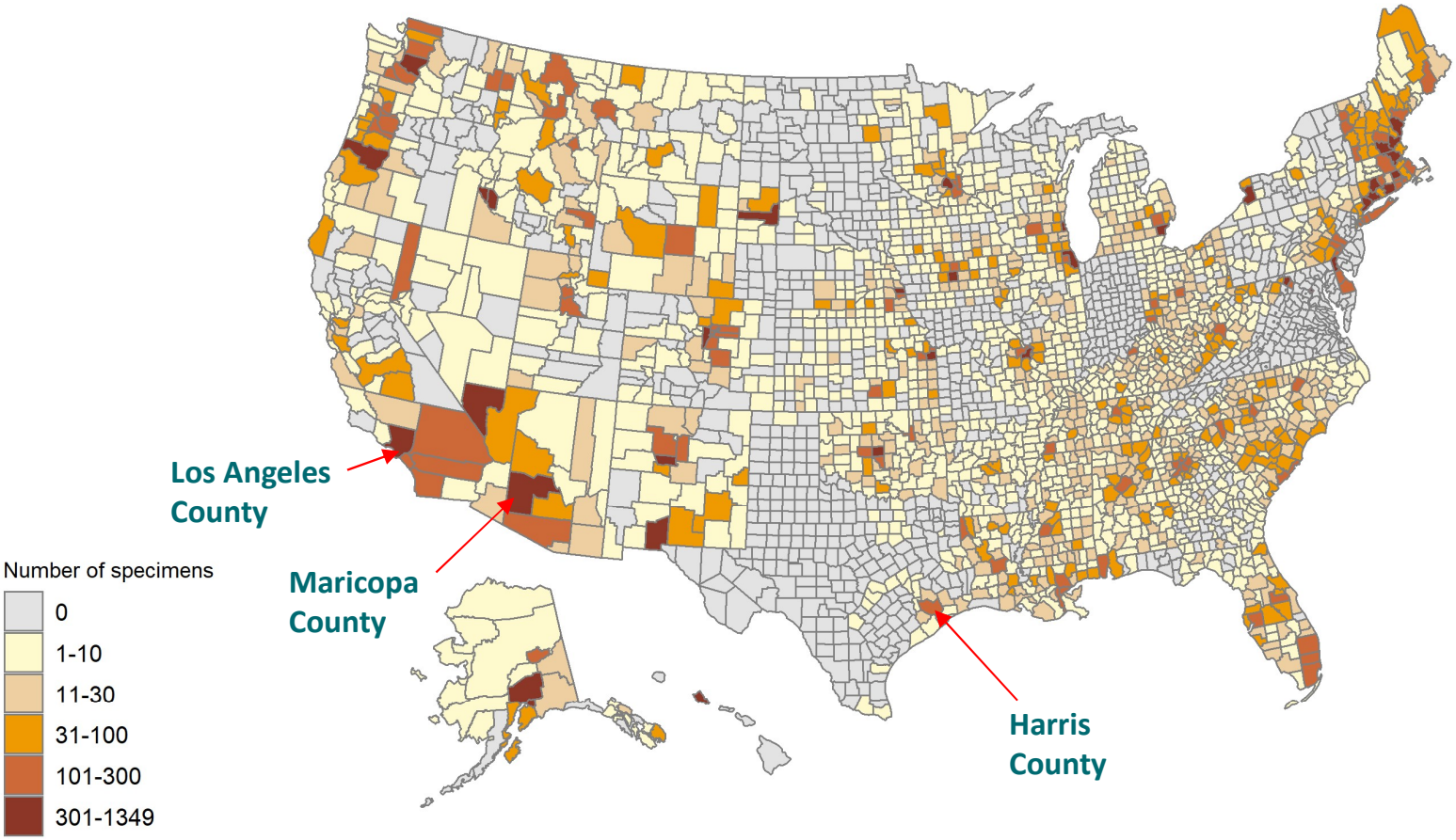
- Do the serology samples provide good geographic representation at the jurisdiction level?
- How do the counties in the serology sample differ from the counties not in the serology sample in terms of demographic and socioeconomic (SES) characteristics?

# U.S. Population by County in 2020



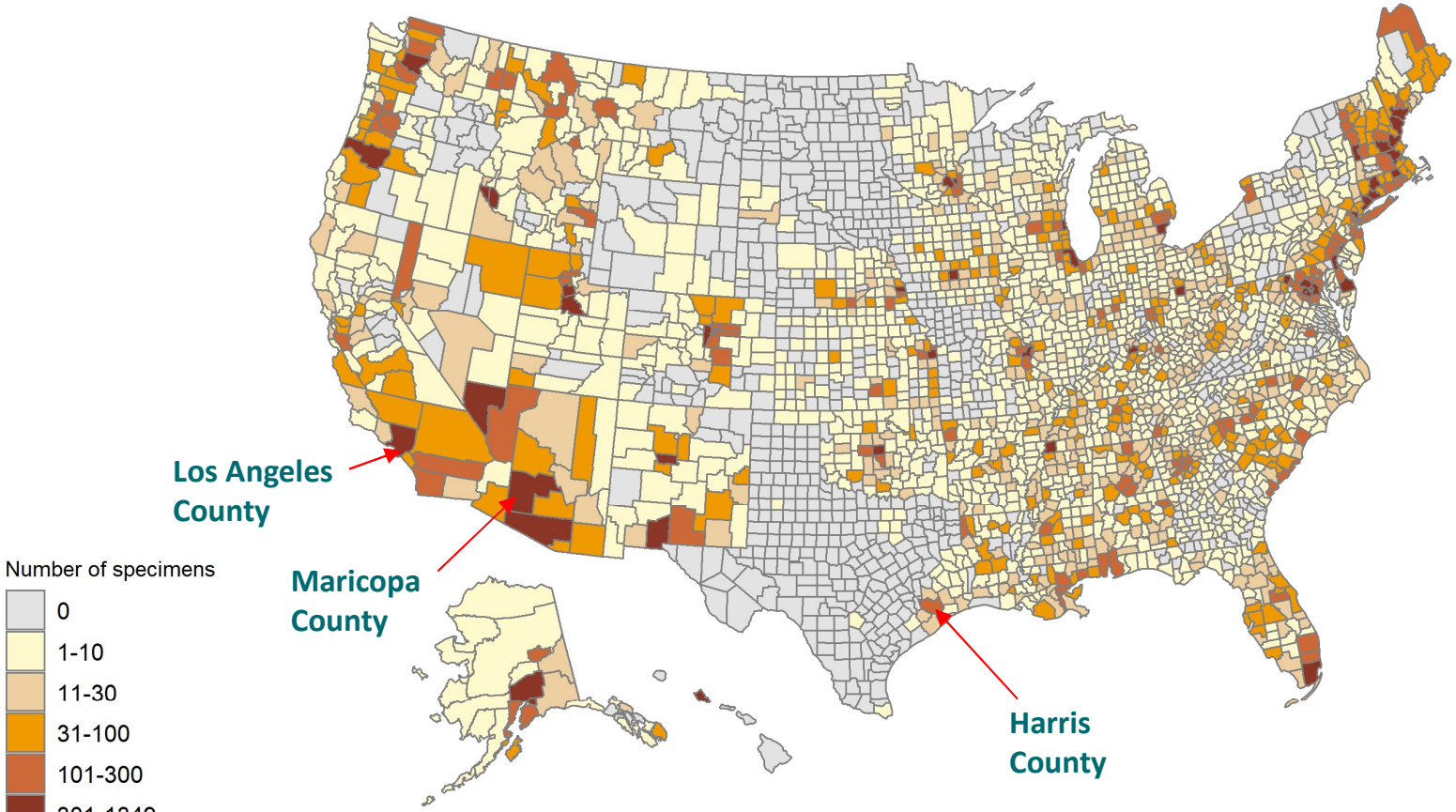

Source: American Community Survey (ACS) 2020 Public Use Microdata Sample (PUMS)

# Total Number of Specimens (September 2021)



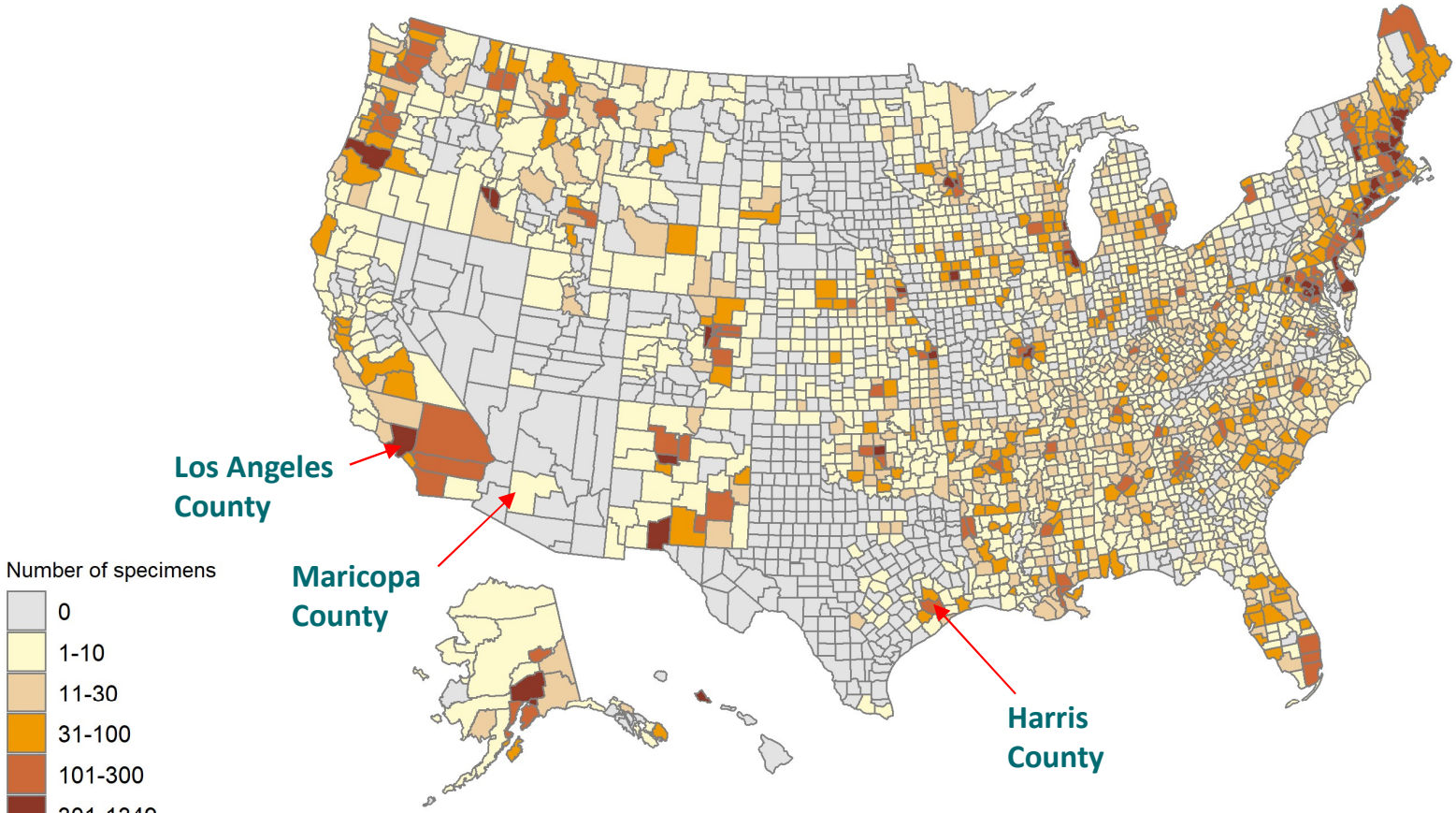

Source: CDC Nationwide Commercial Laboratory Seroprevalence Round 25 (September 2021)

# Total Number of Specimens (November 2021)




Source: CDC Nationwide Commercial Laboratory Seroprevalence Round 27 (November 2021)

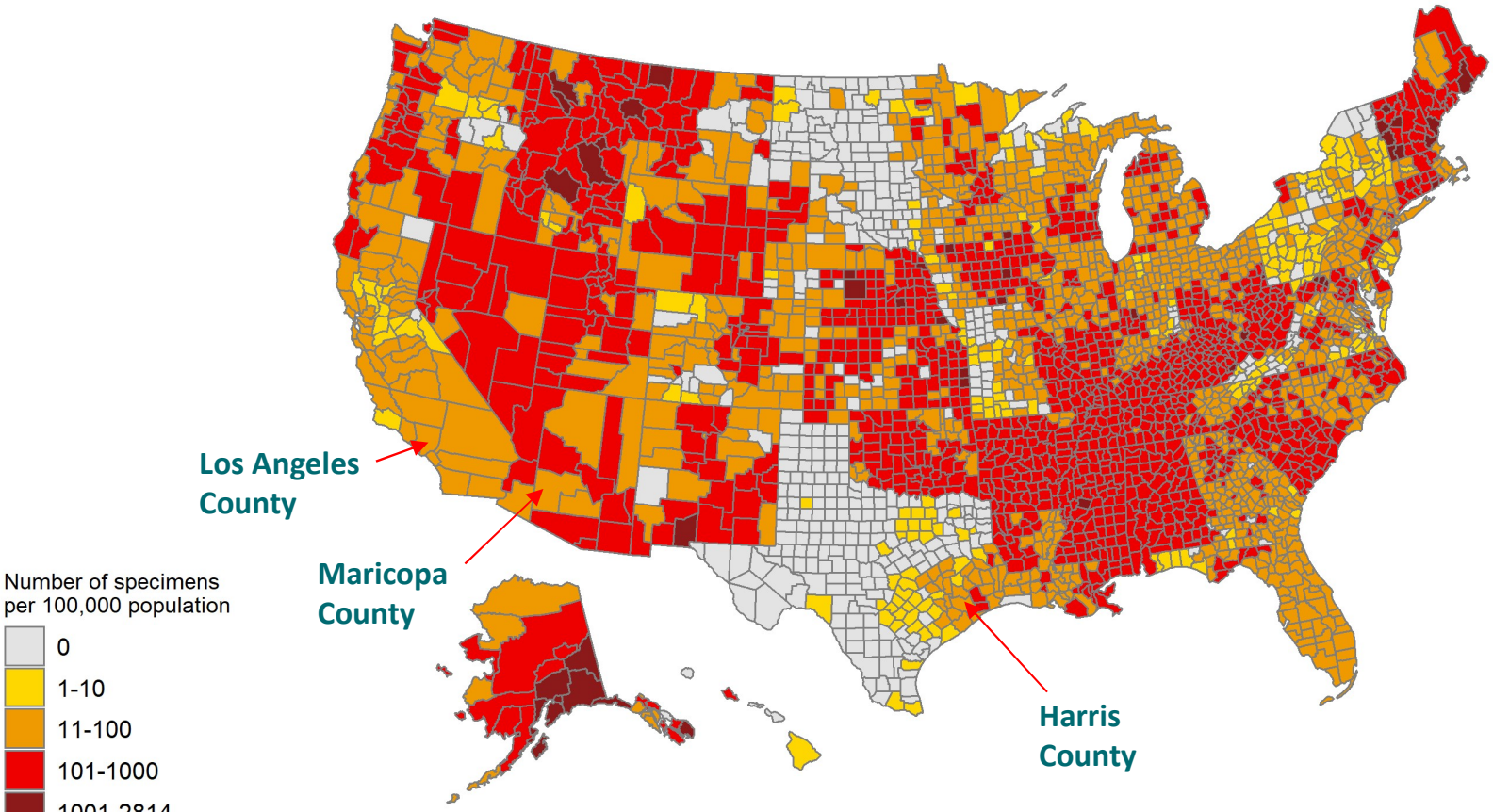
# Total Number of Specimens (January 2022)




Source: CDC Nationwide Commercial Laboratory Seroprevalence Round 29 (January 2022)

# Specimens per 100,000 Population in Rounds 25-30

## (September 2021 – February 2022)




Source: American Community Survey (ACS) 2020 Public Use Microdata Sample (PUMS)
CDC Nationwide Commercial Laboratory Seroprevalence Round 25-30 (September 2021 – February 2022)

# Methods: Overview

| | Research Questions | Methods Summary |
|---|---|---|
| 1 | Does the serology sample provide good geographic representation at the state level? | Dissimilarity Index |
| 2 | How do the counties in the serology sample differ from the counties not in the serology sample in terms of demographic and SES characteristics? | County-level multivariable logistic regression model |

# Method 1

## Dissimilarity Index

Measure of the evenness with which two groups are distributed across component geographic areas that make up a larger area.

$$= \frac{1}{2}\sum_{i=1}^{N}\left|\frac{Specimen\ counts\ in\ county(i)\ in\ the\ state\ sample}{Total\ sample\ in\ the\ state} - \frac{County(i)\ population\ in\ the\ state}{Total\ population\ in\ the\ state}\right|$$

where N = the number of counties in the state.

*Reference*: Agresti, A. Categorical Data Analysis, Second Edition. 2002.

# Method 2

## ❑ Statistical Analysis

- Chi-square Test

  Compared demographic and SES characteristics between counties in the sample and counties not in the sample

- County-level Multivariate Logistic Regression Model

  Examined the association of county average response in the serology sample with county-level demographic and SES characteristics

# Method 2

## ❑ Dependent Variable

- 1: counties having at least one sample in the serology data (Sep. 2021- Feb. 2022)
- 0: counties having no sample in the serology data (Sep. 2021- Feb. 2022)

## ❑ Independent Variables – from American Community Survey (ACS)

- Categorical variables based on the terciles within the state of the following county-level characteristics :

  % Black, Hispanic, or Asian/Pacific Islander people

  % Age 65 and older

  % Under poverty line

  % With college education

  % Employed

  Population size

- Census Regions (Midwest, Northeast, South, West)

# Geographic Representativeness

| Jurisdiction | Number of County | Sample Size | Dissimilarity Index |
|---|---|---|---|
| South Dakota | 66 | 751 | 74.26 |
| North Dakota | 53 | 75 | 63.89 |
| Texas | 254 | 3942 | 63.01 |
| Missouri | 115 | 8837 | 51.96 |
| Montana | 56 | 4851 | 40.92 |
| Pennsylvania | 67 | 9036 | 38.59 |
| Indiana | 92 | 4446 | 38.35 |
| Virginia | 133 | 7450 | 37.47 |
| Iowa | 99 | 8899 | 37.46 |
| Vermont | 14 | 7289 | 37.20 |
| West Virginia | 55 | 9107 | 35.09 |
| Massachusetts | 14 | 8847 | 34.34 |
| New Jersey | 21 | 7645 | 33.85 |
| Louisiana | 64 | 9332 | 33.27 |
| Hawaii | 5 | 9054 | 30.74 |
| Puerto Rico | 78 | 9222 | 30.72 |
| Wyoming | 23 | 878 | 30.49 |
| Maryland | 24 | 7777 | 29.96 |
| Washington | 39 | 8754 | 29.18 |
| Oklahoma | 77 | 9172 | 29.12 |
| New Mexico | 33 | 9100 | 29.01 |
| Wisconsin | 72 | 7887 | 28.52 |
| Oregon | 36 | 8681 | 27.21 |
| Kansas | 105 | 9122 | 27.13 |
| Minnesota | 87 | 8987 | 26.23 |
| New York | 62 | 9126 | 25.95 |

**Table 1**. Dissimilarity index for geographic representativeness by states (Sep. 2021 – Feb. 2022)

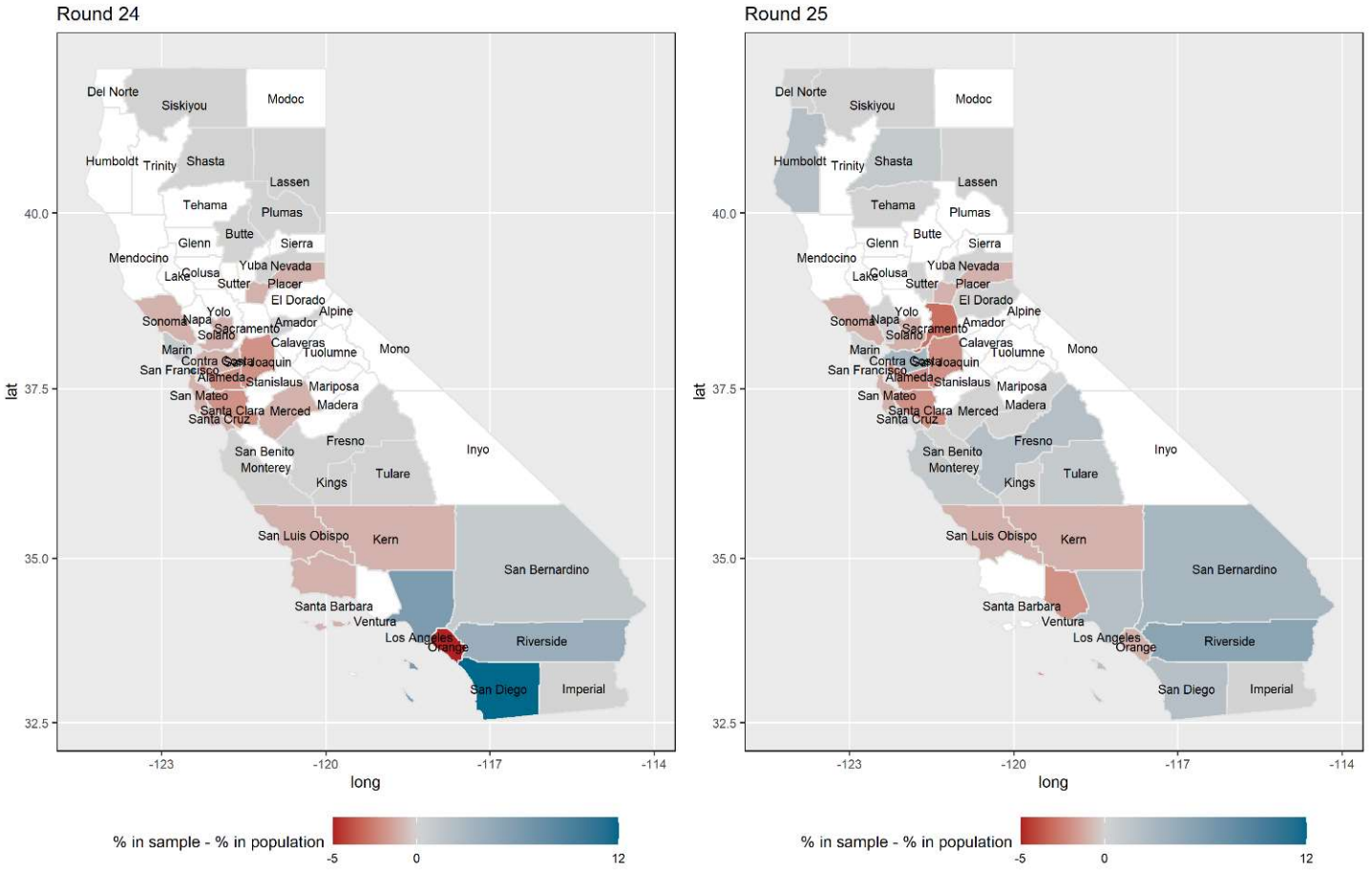
# Geographic Representativeness

| Jurisdiction | Number of County | Sample Size | Dissimilarity Index |
|---|---|---|---|
| Maine | 16 | 8464 | 25.16 |
| Tennessee | 95 | 9796 | 24.95 |
| Idaho | 44 | 8646 | 23.44 |
| Alaska | 29 | 8656 | 22.16 |
| Colorado | 64 | 9073 | 21.88 |
| Ohio | 88 | 9477 | 21.82 |
| Georgia | 159 | 9164 | 20.81 |
| Michigan | 83 | 7791 | 20.18 |
| Nebraska | 93 | 8823 | 20.03 |
| New Hampshire | 10 | 8896 | 18.76 |
| Kentucky | 120 | 9914 | 18.18 |
| Mississippi | 82 | 8931 | 15.11 |
| California | 58 | 9852 | 14.91 |
| Arkansas | 75 | 6016 | 13.65 |
| Utah | 29 | 5510 | 13.21 |
| North Carolina | 100 | 9867 | 11.93 |
| Illinois | 102 | 10519 | 10.86 |
| Florida | 67 | 9998 | 10.54 |
| Connecticut | 8 | 9041 | 9.92 |
| South Carolina | 46 | 9971 | 9.76 |
| Alabama | 67 | 9917 | 9.55 |
| Rhode Island | 5 | 8333 | 7.28 |
| Arizona | 15 | 7305 | 6.84 |
| Nevada | 17 | 6756 | 4.28 |
| Delaware | 3 | 7739 | 0.83 |
| District of Columbia | 1 | 8969 | 0.00 |

**Table 1**. Dissimilarity index for geographic representativeness by states (Sep. 2021 – Feb. 2022) (continued)

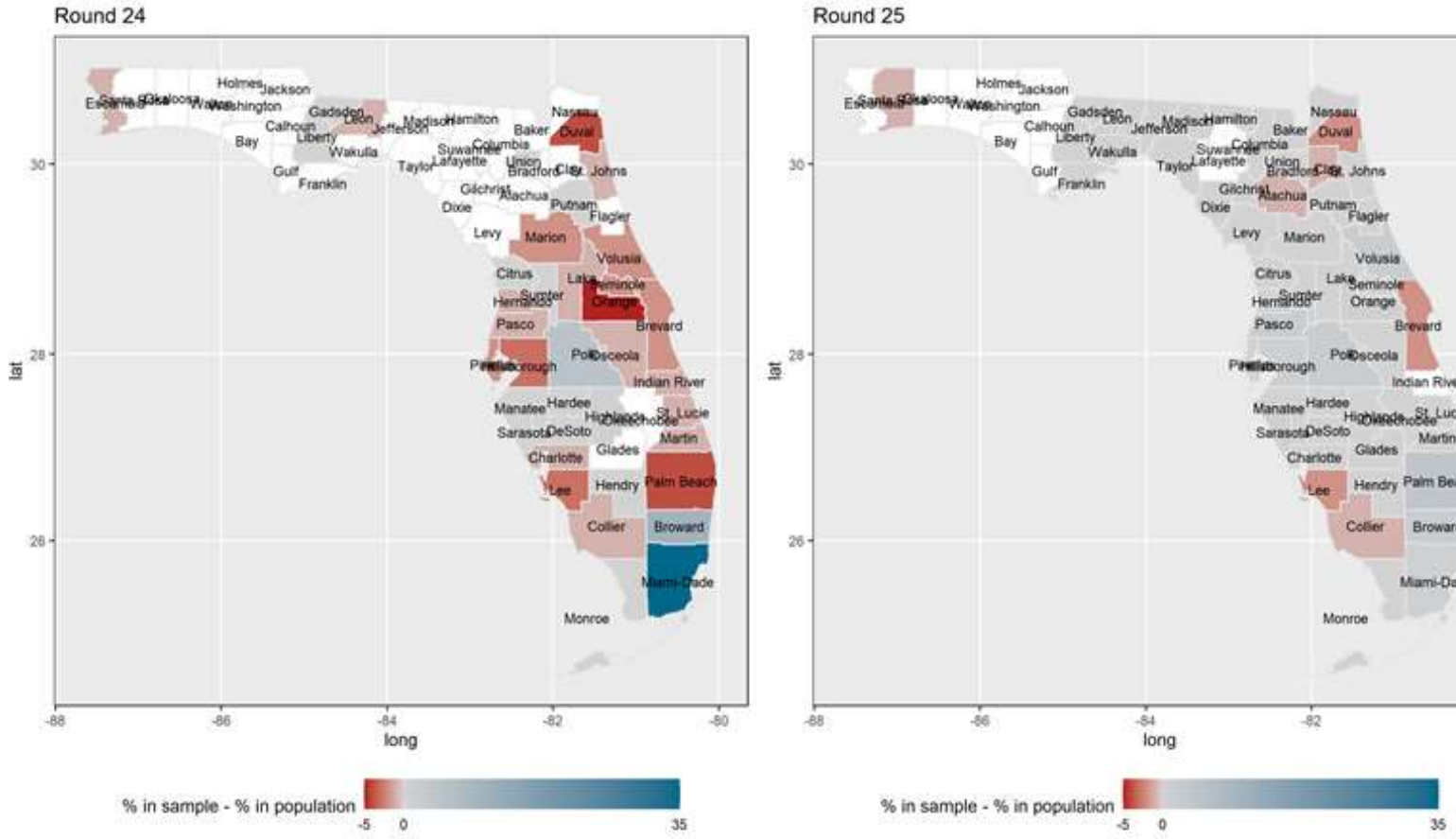
# Geographic Representativeness

- CA



# Geographic Representativeness

- FL



# Comparisons by County Characteristics

**Table 2**. Demographic and socioeconomic characteristics of counties nationally (Sep. 2021 – Feb. 2022)

| County Characteristics | Counties with specimens in sample (n=2522) n(%) | Counties without specimens in sample (n=698) n(%) | P-value |
|---|---|---|---|
| **% 65 and older** | | | |
| Lowest tercile | 973 (34.7) | 91 (21.8) | |
| Middle tercile | 909 (32.4) | 135 (32.4) | <0.001 |
| Highest tercile | 921 (32.9) | 191 (45.8) | |
| **% Black people** | | | |
| Lowest tercile | 826 (29.5) | 175 (42.0) | |
| Middle tercile | 938 (33.5) | 152 (36.5) | <0.001 |
| Highest tercile | 1039 (37.1) | 90 (21.6) | |
| **% Hispanic people** | | | |
| Lowest tercile | 787 (28.1) | 152 (36.5) | |
| Middle tercile | 920 (32.8) | 153 (36.7) | 0.002 |
| Highest tercile | 1096 (39.1) | 112 (26.9) | |
| **% Asian/Pacific Islander people** | | | |
| Lowest tercile | 874 (31.2) | 166 (39.8) | |
| Middle tercile | 948 (33.8) | 123 (29.5) | <0.001 |
| Highest tercile | 981 (35.0) | 128 (30.7) | |

* Significance of the difference between two county groups are tested using Chi-square tests.
* County characteristics were defined based on the tercile within the state.

# Comparisons by County Characteristics

**Table 2**. Demographic and socioeconomic characteristics of counties nationally (Sep. 2021 – Feb. 2022) (continued)

| County Characteristics | Counties with specimens in sample (n=2522) n(%) | Counties without specimens in sample (n=698) n(%) | P-value |
|---|---|---|---|
| **% Higher education** | | | |
| Lowest tercile | 898 (32.0) | 159 (38.1) | |
| Middle tercile | 911 (32.5) | 153 (36.7) | <0.001 |
| Highest tercile | 994 (35.5) | 105 (25.2) | |
| **% Poverty** | | | |
| Lowest tercile | 944 (33.7) | 121 (29.0) | |
| Middle tercile | 897 (32.0) | 151 (36.2) | 0.112 |
| Highest tercile | 962 (33.3) | 145 (34.8) | |
| **% Employed** | | | |
| Lowest tercile | 900 (32.1) | 160 (38.4) | |
| Middle tercile | 909 (32.4) | 151 (36.2) | <0.001 |
| Highest tercile | 994 (35.5) | 106 (25.4) | |
| **Population size** | | | |
| Lowest tercile | 847 (30.2) | 227 (54.4) | |
| Middle tercile | 924 (33.0) | 132 (31.7) | <0.001 |
| Highest tercile | 1032 (36.8) | 58 (13.9) | |

* Significance of the difference between two county groups are tested using Chi-square tests.
* County characteristics were defined based on the tercile within the state.

# Results

**Table 3**. Multivariate logistic regression on sample representation (Sep. 2021 – Feb. 2022, national data)

| County characteristics * | Having at least one specimen in the sample | |
|---|---|---|
| | OR | 95% CI |
| **% 65 and older** (ref: Lowest tercile) | | |
| Middle tercile | 0.963 | (0.700, 1.320) |
| Highest tercile | 0.921 | (0.652, 1.295) |
| **% Black people** (ref: Lowest tercile) | | |
| Middle tercile | 1.058 | (0.825, 1.358) |
| Highest tercile | 1.711*** | (1.276, 2.304) |
| **% Hispanic people** (ref: Lowest tercile) | | |
| Middle tercile | 1.086 | (0.833, 1.419) |
| Highest tercile | 0.872 | (0.656, 1.160) |
| **% Asian/Pacific Islander people** (ref: Lowest tercile) | | |
| Middle tercile | 0.799 | (0.614, 1.040) |
| Highest tercile | 0.748 | (0.547, 1.026) |
| **% College degree or higher** (ref: Lowest tercile) | | |
| Middle tercile | 0.905 | (0.695, 1.178) |
| Highest tercile | 0.898 | (0.646, 1.251) |
| **% Poverty** (ref: Lowest tercile) | | |
| Middle tercile | 0.831 | (0.624, 1.106) |
| Highest tercile | 0.893 | (0.648, 1.230) |
| **% Employed** (ref: Lowest tercile) | | |
| Middle tercile | 0.950 | (0.723, 1.247) |
| Highest tercile | 1.056 | (0.740, 1.511) |
| **Population size** (ref: Lowest tercile) | | |
| Middle tercile | 2.002*** | (1.549, 2.595) |
| Highest tercile | 4.969*** | (3.410, 7.331) |

OR: Odds Ratio; CI: Confidential Interval
County characteristics were defined based on the tercile within the state.
Significance level: *p<0.05 **p<0.01 ***p<0.001

# Conclusions

- The people in the serology sample are more likely to come from the counties with highest tercile of percentage of people who are Black, and highest tercile of size.
  - The odds of being represented in the sample is 1.7 times higher for the counties with the highest tercile of percentage of people who are Black, compared to the lowest tercile counties.
  - The odds of being represented in the sample is 4.9 times higher for the counties with the highest tercile of population size, compared to the lowest tercile counties.

- **The sample population overrepresents people living in counties with a higher percentage of people who are Black and higher population counties**.

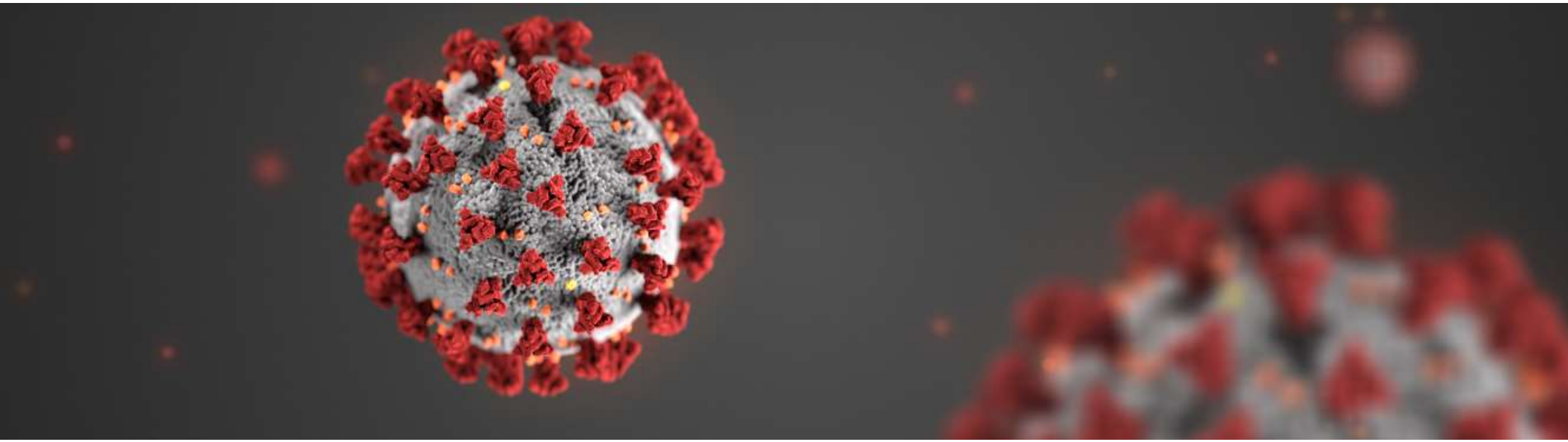
# Limitations

- Analyses were conducted using the county characteristics due to the absence of individual demographic and SES information of people. Therefore, while counties with a higher percentage of people who are Black are more represented in the survey, the true racial and ethnic representativeness of survey data cannot be calculated.

- Analyses used only the serology sample collected Sep. 2021 – Feb. 2022 (Rounds 25-30). The representativeness can fluctuate over rounds and the sample population characteristics may differ from round to round.

# Implications

- The geographic analysis and dissimilarity index help commercial labs improve data collection and geographic representativeness.
  - For example, with very high dissimilarity indices, it is recommended to revisit data collection practice in Texas and Missouri to secure more geographic representation and validity of seroprevalence estimates.
- The potential bias of the sample can provide information about sample characteristics that can better inform data interpretation for any research using this serology sample.
  - Due to the sample characteristics, the seroprevalence estimates from the serology sample is likely to be related to the behavioral or healthcare-related characteristics of people living in large counties or counties with high concentrations of Black population.



For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.