medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

ESTIMATING PREVALENCE AND TIME COURSE OF SARS-CoV-2 BASED ON NEW HOSPITAL ADMISSIONS AND PCR TESTS

Jose E. Gonzalez*

*Aletheia Analytics LLC

ABSTRACT

Data posted in the COVID 19 tracking website for RT-PCR (PCR) results and hospital admissions are used to estimate the prevalence of the SARS-CoV-2 pandemic in the United States (1). Hospital admissions mitigate positive sampling bias in PCR tests due to their initially limited test numbers and application as a diagnostic, instead of a surveying tool.

As of July 31, the United States' cumulative recovered population is estimated at 47% or 155 million. The remaining susceptible population is 53%, or 47% excepting the 6% infectious population. The estimated mortality rate of the cumulative recovered population is 0.09% death per case.

New York and Massachusetts show SARS-CoV-2 prevalence of 87% and 55%, respectively. Likewise, each state exhibits relatively low current positive PCR results at 1 % and 1.7%. Also, these states show about twice the mortality rate of the nation. Florida, California, and Texas showed recovered population percent around 40%, higher current PCR positive test results ranging from 7% to 13%. A higher recovered population mitigates the current positive value attainable by limiting the infectivity rate Re.

This approach provides an alternate source of information on the pandemic's full time course since the serological testing only views a narrow time slice of its history due to the transient nature of the antibody response and its graduated expression dependency on the severity of the disease. The deficiency of serological testing to estimate the recovered population is made even more acute due to the large proportion of asymptomatic and sub-clinical cases in the COVID-19 pandemic.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

INTRODUCTION

The expectation to use straightforward statistics based on serological testing to measure the historical prevalence of SARS-CoV-2 was dashed with the discovery of the transient and disease-severity indexed antibody response in COVID-19 patients (2,3). This paper relies on the integration of % PCR positive test results over time, cycle-corrected for the length of disease, and coupled with hospital admissions to control for PCR testing sample bias, to estimate the kinetics and prevalence over the time course of the pandemic in the United States.

Early on the SARS-CoV-2 pandemic, studies relied on mortality rate to estimate R0 (4), the influenza-like-illness (ILI) surge to explore prevalence of 2020 (5), and serology to measure the extent of infection in New York state. (6,7). By comparison, this paper bases its approach on PCR test records and COVID-19 specific hospital admissions posted on the COVID-19 tracking website (1). New hospital admissions correlate with the ILI surge reported in 2020. As with ILI, new hospital admissions are exempt from positive sampling bias.

The correlation with new hospital admissions enables the mitigation of positive sampling bias found in PCR test results due to its use as a diagnostic tool instead of as a population survey measure. Likewise, the early limited numbers of daily PCR tests performed, and the high positive results detected contributed to positive sampling bias. New hospital admissions (NHA) for COVID-19 are free of such bias. When PCR tests are plentiful, such that % positives fall to roughly 5% of the number of tests, and the ratio to new hospital admissions becomes stable, the PCR sampling bias is largely mitigated. The ratio obtained by dividing NHA by %PCR new case in this stable region in turn enables the estimation of % positive cases in the population by dividing the observed NHA by this constant ratio on any date over the earlier time course of the pandemic.

RESULTS

Because PCR tests were rolled out with an emphasis as a diagnostic rather than as a survey tool, and only sick people were encouraged to test, positive results are likely to be enhanced by sampling bias. As sampling numbers increase well beyond the number of sick people tested, the positive bias diminishes. Figure 1 shows that beyond 500,000 daily tests, the % PCR positive remains steady at around 7% instead of decreasing with increasing daily test numbers, as is expected with the dilution of the positive bias. Thus, positive sampling bias is mitigated above this daily test number.

More importantly, it is possible to correlate % PCR New cases to a positive bias-independent parameter, New Hospital Admissions (NHA), as shown in Figure 2A. This figure illustrates the positive bias correction obtained from the NHA curve, which shows lower values at the beginning of the pandemic (associated with lower daily test numbers) than % PCR tests.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

In Figure 2B, the relationship with the CDC weekly P&I prevalence is included. It peaks over the same period as %PCR cases and New Hospital Admissions (NHA). This agreement with P&I intensity reinforces NHA's use as a marker for the prevalence of SARS-CoV-19 in the United States population. Normalizing the data for all three variables on week 20 shows a stable quantitative relationship between %PCR and NHA after that (R-square = 0.98), making it useful in estimating % positive cases of COVID-19 over the entire time based on NHA cases. NHA cases are free from the %PCR testing constraints that create sampling bias in that database.

A similar analysis, but independent of calendar time, is presented in Figure 3, which shows that the Ratio of New Hospital Admission (NHA) to % PCR New Cases remains within a narrow range (about 450 admissions/%New PCR cases) when above 500,000 daily PCR tests. This constant ratio enables the estimation of % New Cases throughout the time course of the data unencumbered by the positive sampling bias attributed to the PCR test.

Figure 4A shows the time course comparison of the SARS-CoV-2 United States infected population obtained from PCR tests and NHA. The figure shows the % of the United States population infected on any given date. All curves are 7-day moving averages of the data reported by the COVID-19 Tracking Project website (1). The NHA curve peaks at 14% while the %PCR curve draws a broad peak reaching 22% about April 11. The curves then trace converging paths to congruence on June 25, after which daily PCR tests consistently exceed 500,000.

Similarly, Figure 4B shows the % of the United States population infected on any given date. The NHA curve shows a sharp peak at 51 million people infected in the United States on April 11, and a broad peak at 29 million people on July 14. Since March 26, the number of infected people never goes below 12 million. By comparison, the %PCR curve peaks broadly at 70 million new cases before gradually converging with the NHA curve on June 25. The higher population percent and infected numbers obtained in the %PCR curve occur at relatively low testing numbers, and consequently, higher positive bias remedied in the NHA curves.

The NHA derived infection time course curves allow the estimation of the number of people recovered nationwide. The daily curve values divided by the average detectable duration of the disease (20 days) gives the total number or percent of the recovered population in the United States (8, 9, 10). This is depicted in Figure 5 in both millions of people and population percent. Accordingly, by July 31, 47% of the United States population, or expressed in numbers, 155 million people had recovered from COVID-19, except for 145,425 deaths (1). Because on July 31, new cases in the United States registered at 6.5% (21 million people), the population remaining at risk is 46.5%, or 153 million. By comparison, the % PCR curve, with its positive early bias, results in an estimated 72% recovered population.

Table 1 summarizes the critical parameters of infection prevalence in the United States and select states. The column labeled NHA/%PCR shows that this ratio is roughly proportional to the state's population compared to the entire nation. Such agreement shows an internal consistency of new hospitalization events across states in response to new PCR-identified cases when PCR is conducted in a low-biased environment (higher testing numbers, lower percent positives). New York and Massachusetts have higher recovered populations at 87% and 55%,

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

respectively. Together with the densely populated Eastern seaboard states, Massachusetts to Delaware experienced a sudden and extensive onslaught of the epidemic in the Spring of 2020. Consequently, they are also experiencing much lower % new cases as reflected in the lower infectivity rate, Re, calculated from the initial R0, the reduced susceptible population, and the original mitigation practices (4).

The mortality column shows a fatality rate of 0.1% for the population of cases in the United States, including asymptomatic and subclinical cases (6). The Eastern seaboard states reported about twice the fatality rate of others in the table.

DISCUSSION

Rosenberg et al. conducted an antibody testing survey which showed a cumulative incidence of COVID-19 of 14% in New York state by March 29 (7). Silverman et al., using the CDC influenza surveillance networks to estimate the prevalence of SARS-CoV-2, found that over 8.3% of New York state residents were infected by March 28, while F.P. Havers et al. estimated the prevalence of SARS-CoV-2 at 6.9% over the period March 23 to April 1 (6,11). This paper estimates the cumulative incidence of 16% on March 29. The infection rate was moving with a doubling time of about four days in New York state in March. Thus, slight differences in timing would lead to substantial differences in prevalence estimates. These results lie within 40% of their common mean. F.P. Havers et al. also performed a serological convenience survey in South Florida, showing a prevalence of 1.85% between March 23 and April 1, and 4.9% between April 6 and April 10. The present paper has estimates for the entire state at 0.9% and 2.7%, respectively.

Recently, the discovery of transient antibody response to SARS-CoV-2 has cast doubt on the straightforward value of subsequent serological studies aimed at determining the cumulative prevalence of the virus. Additionally, the response is graduated to the severity of the disease such that milder infections will lead to an earlier drop of detectable SARS-CoV-2 antibody below the baseline. Fan Wu et al. reported that 30% of cases tested had extremely low neutralizing antibodies specific for SARS-CoV-2, and another 17% had low levels (3). J Seow showed that patients with early modest antibody responses led to undetectable levels 50 days post-infection. Higher antibody responses in other individuals remain stable at 60 days, possibly longer (2). These findings suggest that current serological studies are limited to time-sliced views of the whole infected population and are more effective in detecting severe disease. Since severe disease is present only in a small subset of the afflicted population, results will tend to underestimate the population of recovered COVID-19 patients (1, 2).

Starting on week 20 of 2020, % positive PCR and new hospital admissions display close linear agreement with a coefficient of determination of 0.98 (see Figure 2B). This observation, together with the lower positive % PCR results, and non-decline of detected % positives with increasing test numbers, justify the association of % PCR positive cases with NHA. This close

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

agreement forms the basis for estimating new cases between calendar weeks 10 and 20, when %PCR results seem to show a positive bias.

When representative, performing over 500,000 daily tests provides a high level of confidence and precision in survey results.

Even in New York state where, according to this study, the recovered population reaches over 87%, new cases continue to test daily at about 1% by PCR. The estimated actual daily increase of infected people is closer to 0.1% (19,500 people) after correcting for the daily rate of recovered patients. This observation supports the notion that SARS-CoV-2 infection will continue to advance through the population, even when the extent of the recovered population is remarkably high. Although under this circumstance, the advance should happen at a substantially lower daily rate and, therefore, pose a lower risk to the remaining susceptible population of contracting the infection, or overwhelming healthcare facilities and human resources.

Postponing exposure makes sense because medical treatments are becoming more effective over time. The risk is highest for the over 65 age group, who account for 80% of all COVID-19 deaths while comprising 16% of the United States population. Within this age group, COVID-19 mortality accounts for 9.5% of all deaths. By contrast, the age group below 24 has a 0.2% COVID-19 mortality rate, accounting for 0.9% of all deaths within the group, while comprising 32% of the United States (12).

More effective mitigation measures should focus on safeguarding the population above 65 years of age to reduce mortality substantially and efficiently.

METHODS

Estimation of New Hospital Admissions

The New Hospital Admissions daily value is computed from the reported 'hospitalized currently' data column by assuming a 14-day average hospital stay which is applied to create a daily discharge value (1, 9, 13) described by the expressions:

NHA = New net hospitalizations + Hospital discharge

New net hospitalizations = hospitalizations current (today)- hospitalizations current (yesterday)

Hospital discharge = hospitalizations current (today) / 14

Florida and Kansas did not report current hospitalizations over a sufficient period to compute NHA for this analysis. Instead, both reported cumulative hospitalizations. In both cases, NHA was calculated from this data by subtracting today's cumulative hospitalizations from the

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

previous day. Other entities included in the COVID-19 database (HI, GU, MP, and VI) were excluded from the analysis because of data deficiencies.

Cumulative hospitalizations are also reported for the United States in the COVID-19 tracking database from The Atlantic (1). For comparison, the time course of the infection was also performed calculating new hospital admissions on this data and obtained similar results.

Case Summation

The calculation of the recovered population's extent is sensitive to the average duration of the illness as detectable by the PCR test. In this paper, the duration of detectability by PCR is set at a conservative average length of 20 days. For this estimation, the conservative choice estimates favor longer detection of the virus in infected patients. Among hospitalized patients, viral shedding persists over a range of 12 to 20 days from start of illness (9, 10, 13, 14, 15).

Average incubation time is reported at 4 to 5 days, and hospital stays come to 12 days. Summing these give a PCR positive duration of about 17 days. Asymptomatic and subclinical infections last less than clinical and infections requiring hospitalization. Although among COVID-19 patients of known disposition hospitalized patients account for about 20% of all infected persons, this is a much smaller number when considering asymptomatic and subclinical cases (8, 9, 10, 14, 15).

The daily percent positive PCR results corrected for sampling bias by the number of hospital admissions and divided by 20 days of viral detectability yields the daily recovered population % of the pandemic. Summing this daily estimate gives the cumulative recovered population.

Limitations:

The ratio of New Hospital Admissions to %PCR positives is used to calculate the percentage of the population infected when the total daily PCR test numbers are below 500,000. This practice encompasses the time from March 18 to June 20. The ratio was calculated from the data reported from June 21 to July 31. The average value for the ratio over this time when daily tests exceeded 500,000 is 450 units, and its standard deviation is 20 units. The 95% confidence interval of the standard error of the mean is seven units. At the higher limit value, 457 units, the recovered % on July 31 is 46.4%; and at the lower limit value 443 units, the recovered % would come to 47.4%. As the value of the ratio decreases, the % recovered increases approaching the value obtained from % PCR positives without mitigating for testing number bias. An increase of the average ratio value by two standard deviations, 40 units, would bring down the percent recovered to 44.3% from 47%. Conversely, a 40 unit decrease in the ratio value would increase the recovered population to 49%.

The duration of PCR viral detection in infected people affects the estimation of the nationwide prevalence of the infection proportionately. Thus, if the average duration of detection is longer

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

by 5% (21 days instead of 20 days), summation would be reduced by 5% (recovered population would be reduced to 44% from 47%), and vice versa.

REFERENCES

1) The COVID Tracking Project at the Atlantic https://covidtracking.com/data/api
2) Jeffrey Seow et al., Longitudinal evaluation and decline of antibody responses in SARS-CoV-2 infection : Medrxiv https://doi.org/10.1101/2020.07.09.20148429
3) Fan Wu et al., Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered 2 patient cohort and their implications https://doi.org/10.1101/2020.03.30.20047365. April 20, 2020)
4) A. R Ives et al., State-by-State estimates of R0 at the start of COVID-19 outbreaks in the USA https://doi.org/10.1101/2020.05.17.20104653
5) CDC Flu Weekly Incidence Report https://www.cdc.gov/flu/weekly/index.htm
6) J. D. Silverman et al., Using influenza surveillance networks to estimate state specific prevalence of SARS-CoV-2 in the United States Sci. Transl. Med. 10.1126/scitranslmed.abc1126 (2020).
7) E. S. Rosenberg et al., Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York https://doi.org/10.1016/j.annepidem.2020.06.004
8) S. A. Lauer, K. H. Grantz et al., The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application Ann Intern Med. doi:10.7326/M20-0504 Annals.org
9) J. A. Lewnard et al., incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in California and Washington: prospective cohort study BMJ 2020;369:m1923 http://dx.doi.org/10.1136/bmj.m1923
10) CDC interim Guidance for management of patients June 30, 2020 https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html
11) F.P. Havers et al., Seroprevalence of Antibodies to SARS-CoV-2 in Six Sites in the United States, March 23-May 3, 2020 https://doi.org/10.1101/2020.06.25.20140384
12) Weekly Updates by Select Demographic and Geographic Characteristics https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm
13) E. M. Rees et al., COVID-19 length of hospital stay: a systematic review and data synthesis medRxiv 2020.04.30.20084780; doi: https://doi.org/10.1101/2020.04.30.2008478
14) Lee YH, et al., Clinical Course of Asymptomatic and Mildly Symptomatic Patients with Coronavirus Disease Admitted to Community Treatment Centers, South Korea [published online ahead of print, 2020 June 22]. Emerg Infect Dis. 2020;26(10):10.3201/eid2610.201620. doi:10.3201/eid2610.201620
15) Young BE, et al. Epidemiologic Features and Clinical Course of Patients Infected With SARS-CoV-2 in Singapore. JAMA. 2020;323(15):1488–1494. doi:10.1001/jama.2020.3204

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

FIGURES AND TABLES

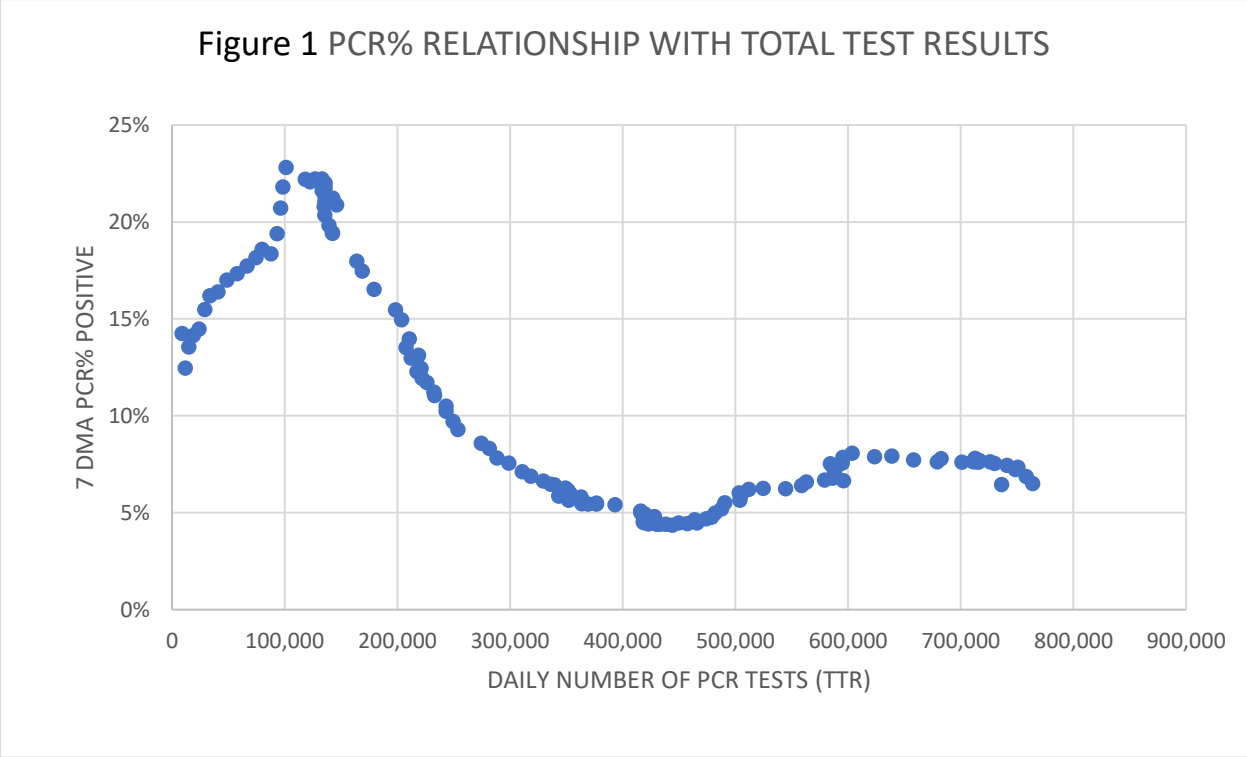

Figure 1 shows the relationship of the 7-day moving average of positive % PCR versus the daily number of tests with the 7-day moving average of the total number of daily PCR tests. The % positive PCR stabilizes beyond 500,000 daily tests. 7-day moving averages are used in these figures and estimations because of the strong weekly cyclicality of data reporting.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

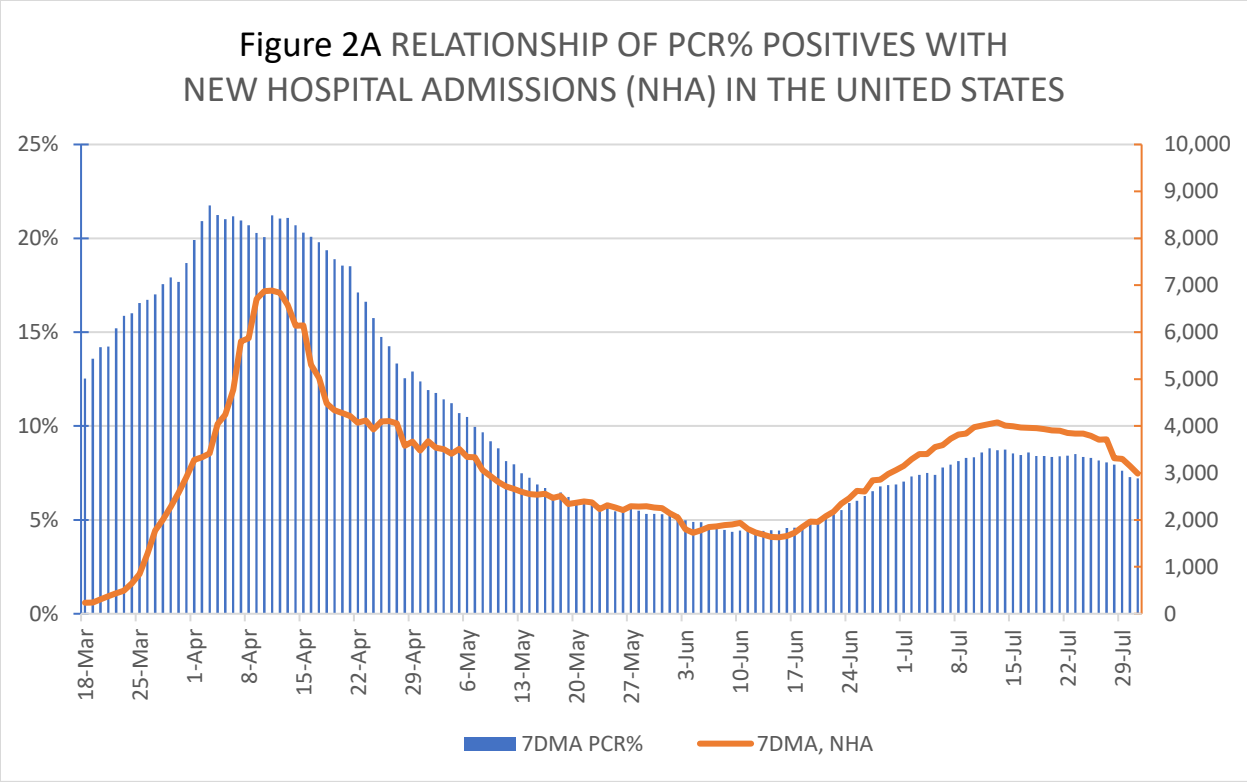

Figure 2A exhibits the relationship between the 7-day moving average of positive %PCR and NHA. NHA is represented on the right ordinate. After June 21, the daily number of PCR tests exceed 500,000. The ratio of these two variables is computed for the daily values after this date, obtaining an average of 450 NHA/PCR % positive. Including the data back to May 21, the start of calendar week 20, when daily tests are over 380,000, leads to a lower ratio, 425 NHA/PCR % positive. Since the ratio drives the new daily case %, this reduction in ratio value increases the % cumulative recovered population from 47% to 50% by July 31.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

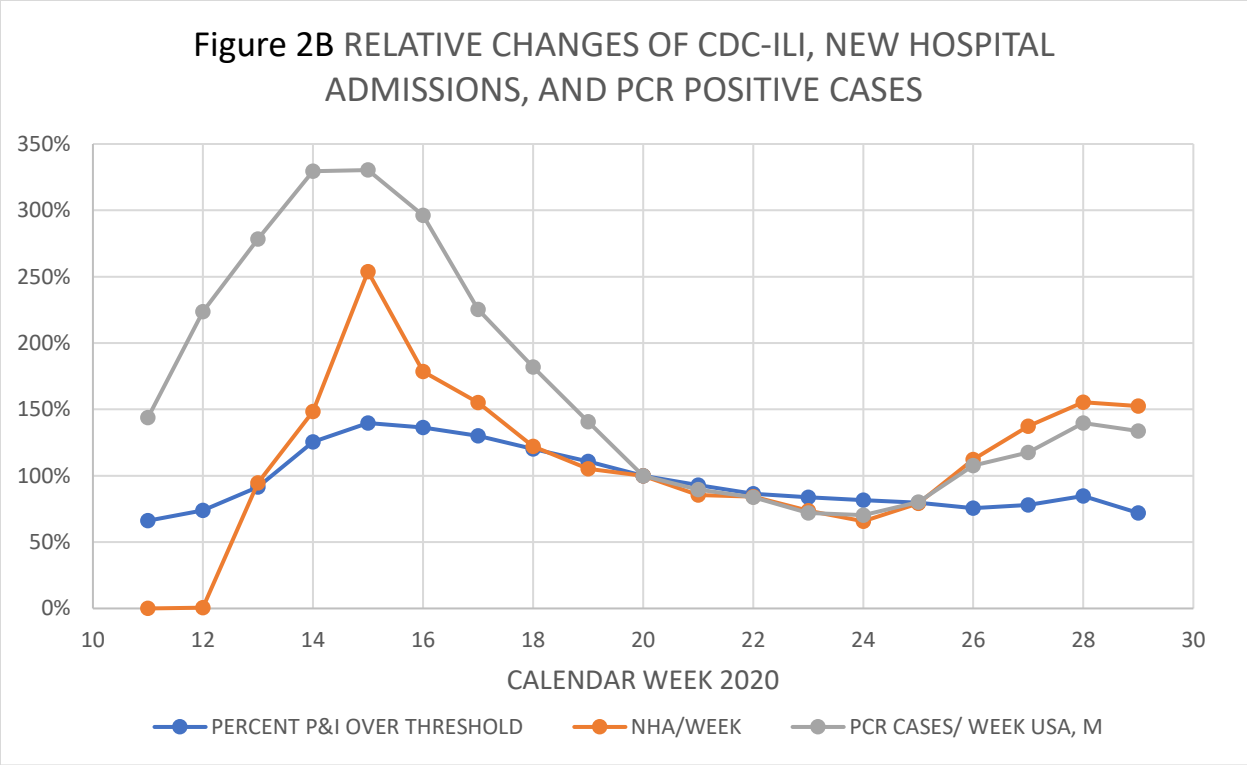

Figure 2B shows the relationship of normalized weekly ILI surge, NHA, and New PCR cases to a 100% on calendar week 20. All three variables tend to move together. NHA and PCR cases track tightly starting on calendar week 20. Regression analysis of these two variables after that date yields a coefficient of determination of 0.98.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

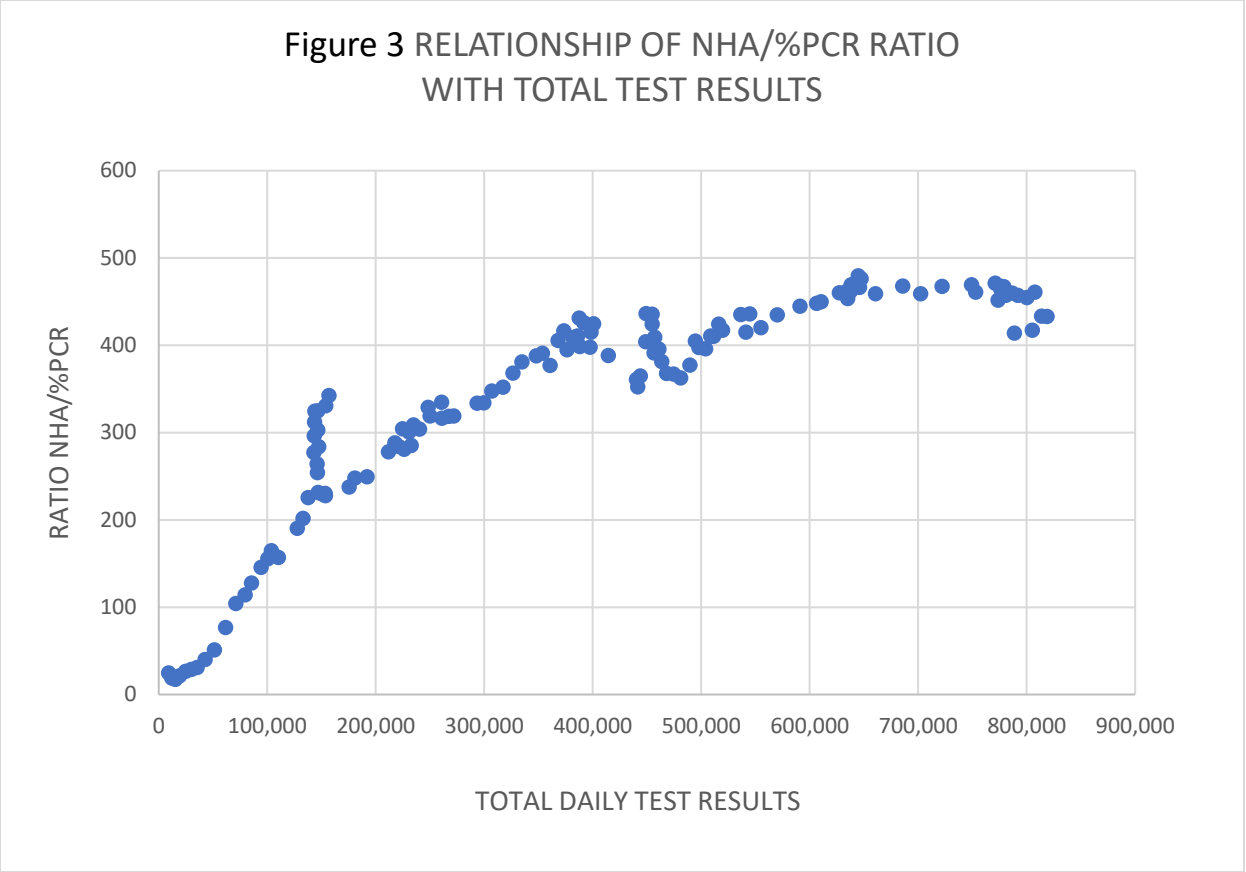

Figure 3 depicts the relationship between New Hospital Admissions (NHA) to %PCR positive ratio over increasing total PCR tests performed daily. Past 500,000 daily tests the ratio stabilizes at a value of 450. The graph is based on 7-day moving averages for NHA, %PCR positive, and test results.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

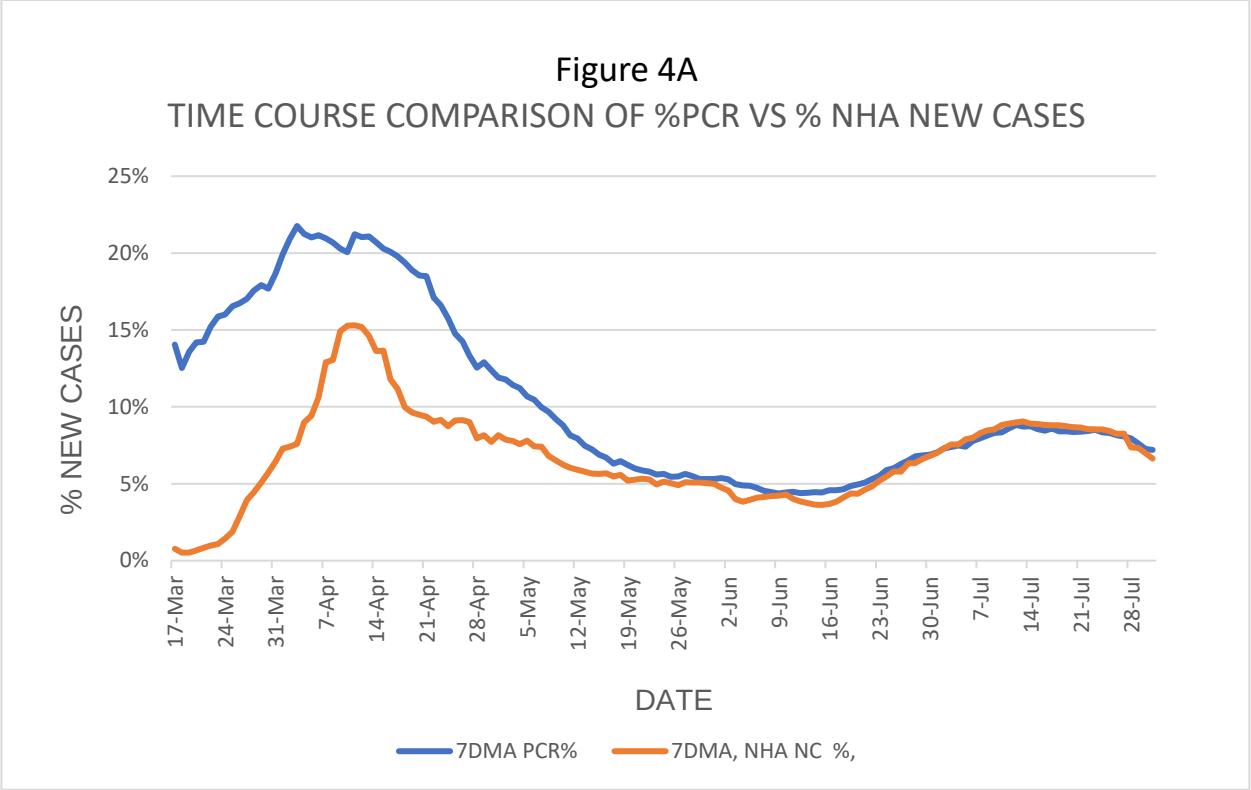

Figure 4A illustrates the time course %PCR positives, and the % new cases calculated by dividing new hospital admissions by 450 (the ratio of NHA to PCR% after total test results exceed 500,000 daily tests). After June 21, total daily test results exceed 500,000. The lower curve obtained from NHA mitigates the positive sampling bias before this date.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

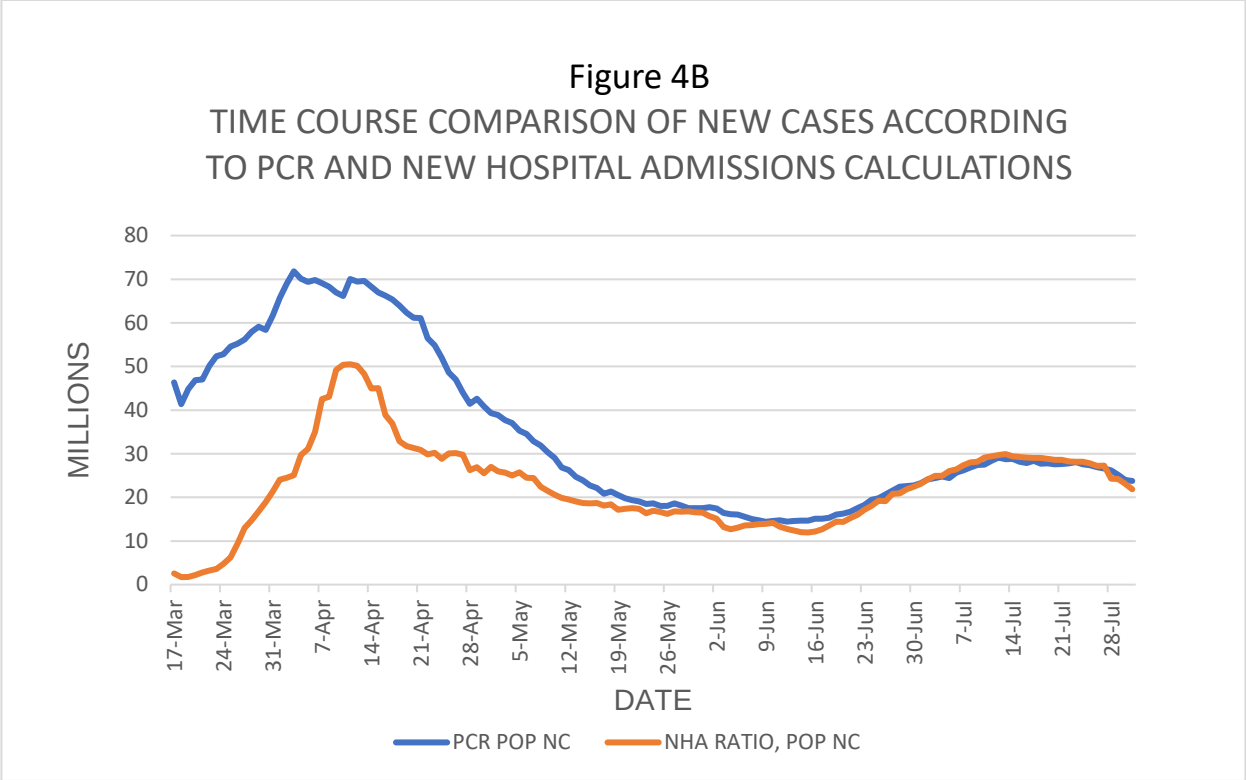

Figure 4B shows the information presented in 4A converted to millions of people infected from a perspective that the NHA calculations render the obtained information as a nationwide population survey. All the data plotted is based on 7-day moving averages.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

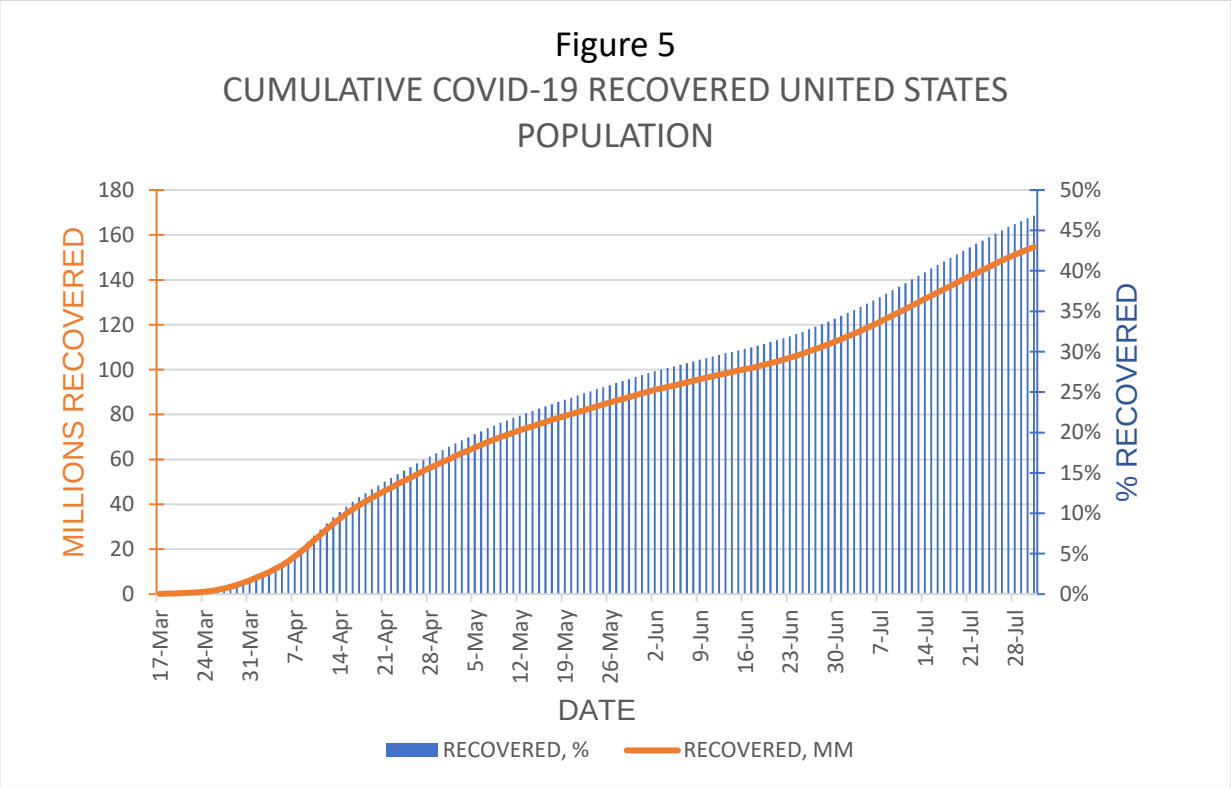

Figure 5 expresses the time course for the COVID-19 recovered population in terms of percent and millions of patients (MM). This information is derived by summing the daily incidence obtained from the NHA new case curve divided by the duration of detectable disease (20 days) (8, 9, 10).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.15.20175653; this version posted August 17, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Table 1.

COMPARISON OF HOSPITALIZATION AND MORTALITY RATES, RECOVERED AND SUSCEPTIBLE POPULATIONS, INCIDENCE AND INFECTIVITY CONSTANT (Re) ON JULY31 ACROSS THE UNITED STATES AND SELECT STATES

LOCATION	MILLIONS	NHA PER % NEW CASE	NHA PER RECOVERED CASE	DEATHS	DEATHS/CASE, %	RECOVERED, %	SUSCEPTIBLE, %	NEW CASE % ON JULY 31	Re
UNITED STATES	330	450	0.26%	145,425	0.09%	47%	44%	9.40%	2.14
CALIFORNIA	40	72	0.34%	9,508	0.07%	35%	58%	6.70%	2.03
TEXAS	29	55	0.30%	7,455	0.06%	40%	48%	11.90%	1.82
FLORIDA	21	36	0.35%	7,402	0.09%	37%	49%	13.30%	1.95
NY STATE	20	42	0.34%	32,798	0.19%	87%	12%	0.94%	0.83
MASSACHUSETTS	7	11	0.32%	8,609	0.23%	55%	44%	1.70%	1.76

Table 1 summarizes the information computed for these states following the approach used for the nation. The susceptible column excludes the new cases found on July 31 from its totals. Deaths and hospitalizations are obtained from the COVID Tracking Project (1). Re is calculated from the initial R0 published by Ives et al. (4).