UNIVERSIDAD EAFIT

Vigilada Mineducación

# Documentos de trabajo

## Economía y Finanzas

**Nº 20-15**

2020

# An Estimate of Unidentified and Total U.S. Coronavirus Cases by State on April 21, 2020

Theodore R. Breton

cief
Centro de Investigaciones Económicas y Financieras
Universidad EAFIT

# An Estimate of Unidentified and Total U.S. Coronavirus Cases by State on April 21, 2020[+]

Theodore R. Breton[*]

April 23, 2020

**Abstract**
I use data on cumulative tests, positive tests, and deaths for the coronavirus in South Korea and the U.S. lower-48 states during April 2020 to estimate the extent of infection and the unidentified share of the infected population in each state and in the U.S. as a whole on April 21, 2020. I find that 3.8 million people, or 1.2% of the population, have been infected in the U.S., with rates of infection that range from 0.1% in more rural states to 7.0% in New York state. I estimate that only 20% of all U.S. cases have been identified through testing. The unidentified share of total cases ranges from 61% to 83% across the states. I estimate that 38% of all cases are asymptomatic, which is consistent with the high shares of unidentified cases.

**Key Words**
Coronavirus, U.S. states, rate of infection, asymptomatic share, community spread

**JEL classification**
I12, I18

SSRN 3583941

[+] Los conceptos expresados en este documento de trabajo son responsabilidad exclusiva de los autores y en nada comprometen a la Universidad EAFIT ni al Centro de Investigaciones Económicas y Financieras (Cief). Se autoriza la reproducción total o parcial del contenido citando siempre la fuente.

[*] Grupo de Estudios en Economía y Empresa, Centro de Investigaciones Económicas y Financieras (Cief), Universidad EAFIT, Carrera 49 Número 7 Sur 50, Medellín, Colombia, ted.breton@gmail.com.

## **I.** Introduction

Extremely limited testing for the coronavirus until recently and the uncertain share of asymptomatic cases has led to considerable speculation about the extent of coronavirus infection in the U.S. population. Confirmed tests indicate that hundreds of thousands have been infected, but some researchers have argued that a huge share of the population is already infected, citing preliminary results from antibody testing [Silverman, Hupert, and Washburn, 2020].

Researchers have been waiting for antibody tests to provide definitive data on the extent of the infection, but sufficient valid tests for mass testing may not be available for some time. Researchers have warned recently that many of the tests now coming to market are not reliable [Eder, Twohey, and Mandavilli, 2020]. These tests may be unable to distinguish between COVID-19 and other coronaviruses associated with less serious respiratory illnesses [Cohen, 2020].

In the absence of reliable results from antibody testing, in this paper I use state level data on positive coronavirus test rates and associated death rates, combined with an estimate of the death rate from the virus in South Korea, to estimate the number of coronavirus cases in each state and in the overall U.S. on April 21, 2020. This analysis is based on the assumption that, unlike the situation in other countries, South Korea's documented experience with the coronavirus is accurate and complete, so its death rate provides a benchmark for determining the true levels of coronavirus infection in the U.S.

The hypothesis behind the estimation methodology for the U.S. is that first, the share of positive tests in a state is a valid indicator of the share of the infected population that has been identified through testing. Second, the lower the share of positive tests, the higher the share of the infected population that has been identified. Third, the relationship between the shares of positive tests and the actual number of infected people in each state can be determined using the death rate estimated in South Korea and the relationship between the positive test rates and death rates in the 48 U.S. states.

The results from the analysis show a logarithmic relationship between the death rate and the positive share of tests across states. Applying this relationship to the data reveals that the rates of infection in the U.S. are highly skewed, with New York experiencing a 7% rate of infection, while Wyoming and a number of other more rural states have an infection rate of only 0.1%. Overall the analysis here indicates that the coronavirus has infected 3.8 million people, or 1.2% of the U.S. population, a large number, but nowhere near enough to create herd immunity.

The analysis also shows that only 20% of the infected individuals in the U.S. have been identified through testing through April 21, 2020. The other 80%, numbering over three million people, are too numerous to identify through contact tracing. South Korea managed to control the spread of the virus through a contact tracing process, but they only had 10,600 total cases.

The remainder of this paper is organized as follows: Section II presents the methodology used to create the estimates of the infected population. Section III presents the estimate of the death rate (deaths/total infections) associated with COVID-19 in South Korea. Section IV presents an estimate of the asymptomatic share of coronavirus cases. Section V presents the U.S. state data on the share of positive tests and the estimated death rates and the model to estimate the infected

share of the population in each state. Section VI presents the state-by-state and national estimates of the share of the U.S. population infected with the virus and the share of unidentified cases on April 21, 2020. Section VII concludes.

## II. Methodology

No states in the U.S. have managed to test enough people to determine the share that has the disease. The availability of tests has been so limited that no states have even managed to test all of the individuals with symptoms indicating that they might have the disease. No states have attempted to identify the asymptomatic cases.

In the approach taken here I estimate the infected share of the population in each state using the estimated death rates for the lower-48 states. If there is a known expected rate of death from the disease, then the death rates for those confirmed to have the disease can be used to estimate the missing share of infected individuals in the population. For example, if the true death rate (deaths/infected population) is known to be equal to 2% and the death rate among those tested is 4%, then the population tested is a sample containing only half of the infected individuals.

I do not use the death rate in each state to estimate the total cases in the state, because these rates vary based on the particular population (e.g., nursing homes) that happened to become infected in a state. Instead I use the death rates in all of the states to estimate a relationship between the share of positive tests in a state and the death rate. I do this because the share of positive tests relates more closely to the share of the total cases that have been identified in the state than its particular death rate.

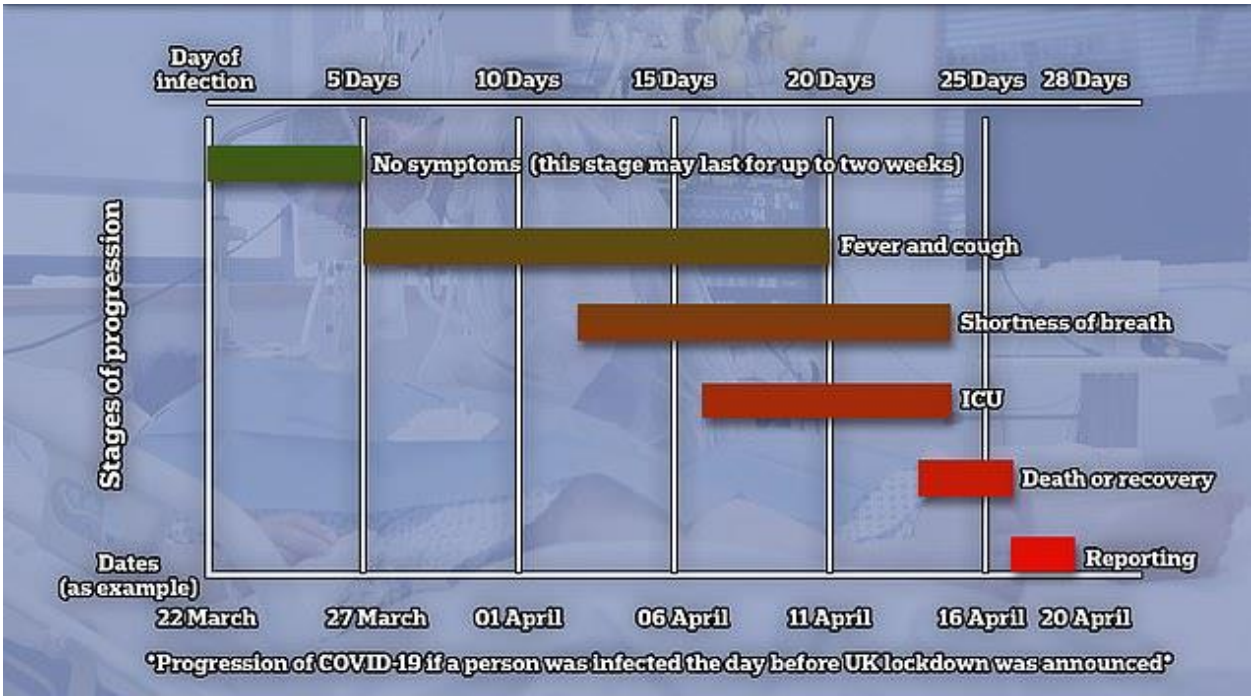
I first use the South Korean data to estimate the death rate for coronavirus in a population where everyone with the virus appears to have been tested. I use the estimated average lag between the onset of the disease and death from the UK and China and estimates of the lag between testing and reporting in South Korea to calculate the deaths in South Korea that correspond to the reported cases. I use this same approach to estimate the death rates in the U.S. by state. After comparing the distribution of cases by age in South Korea and the U.S., I conclude that they are sufficiently similar to use the South Korean death rate as a reasonable estimate of the actual death rate in the U.S.

I then plot the death rate in each state vs. the share of positive tests on April 7$^{th}$ and use a logarithmic trendline to determine the relationship between these two variables in the state data. The logarithmic relationship provides the best fit to the data.

I then calculate the ratio between the estimated death rate and the South Korean death rate to determine the relationship between the total cases and the reported cases in each state. This relationship provides the basis to estimate the identified and unidentified shares of the total cases in each state. It also provides the basis to estimate the number of infected individuals in each state, which in total provides a national estimate of the infected population in the U.S.

*The Daily Mail* has published a time line to show the progression of the disease in the UK [Boyd, 2020]. This time line is shown in Figure 1. It shows five days from contagion to the onset of symptoms and 18-21 days from the onset of symptoms until death. This estimate is similar to that of Verity et al. [2020], who used data for 24 patients in China to estimate that the average delay from the onset of symptoms until death is 17.8 days.

**Figure 1**
**Timeline for Progression of the Coronavirus**




I assume that the lag between the onset of symptoms and death is 19 days in both South Korea and the U.S. I use information about the test reporting lags in South Korea and the U.S. during April to estimate the death rates associated with the reported cases of the virus in each country.

I validate the methodology by examining whether the estimates of the unidentified share of cases in each state are consistent with its reported testing strategy and the limitations on the availability of testing in March and early April. In addition, since the states did not test asymptomatic individuals, I also examine whether the estimated shares of unidentified cases are consistent with my estimate from existing studies of the share of all cases that have been found to be asymptomatic. I examine three studies and end up using the data from a screening study in Iceland and a study of the distribution of case severity in China to determine the asymptomatic share.

## III. The Death Rate in South Korea

The death rate is the share of positive cases that end in death. Since countries generally only count a fraction of the symptomatic cases, the number of deaths divided by the number of reported cases on the same date substantially overestimates the true death rate. In addition, deaths occur later than reported cases, so when cases are rising, the death rate calculated from cases and deaths on the same date is underestimated. Death rates may also be underestimated if individuals die of coronavirus without any testing for the disease and their cause of death is attributed to something else, such as pneumonia.

Clearly the best estimate of the death rate from all cases of coronavirus would come from a large representative sample of the population used to determine the share infected and a follow-up study of the infected group to identify the share of this group that died. No country has carried out such a study, although Iceland has done part of such a study.

South Korea has pursued an alternative approach, in which they attempted to identify all the cases of infection in the country, including those with no symptoms, and then tracked these individuals to determine the associated deaths. The government mobilized the entire population to identify all cases of the coronavirus and carried out the highest rate of testing per capita in the world. It set up drive-by testing, so anyone with symptoms or suspected of being infected could quickly be tested. It then followed up every positive test with thorough contact tracing and tested the contacts, whether they had symptoms or not [DW, 2020].

The evidence indicates that the government managed to find and test almost every symptomatic and asymptomatic case in the country. Since the number of new cases declined to almost zero in April, there must not have been many asymptomatic individuals who were not tested, identified, and quarantined. This means that South Korea is unique among countries in including all of its asymptomatic cases in the national estimate of total cases.

The delay between the onset of symptoms and a confirmed test result is short in South Korea due to the prevalence of testing facilities and its high testing capacity relative to the number of individuals suspected of being infected. The Government of the Republic of Korea [2020] estimated that in early April this delay was 2-3 days. I assume 19 days between the onset of symptoms and death, which leaves 16 days as the average lag between confirmed test results and death at that time.

As shown in Table 1, the cumulative confirmed cases of the virus in South Korea on April 4th was 10,156 and the deaths sixteen days later on April 20th were 236. This yields a death rate for all cases of the virus of 2.3%.

**Table 1**
**South Korea Coronavirus Statistics**

| | Tests Completed | Total Cases | Total Deaths | Population | Positive Test Rate | Death Rate | Infected Rate |
|---|---|---|---|---|---|---|---|
| April 4th | | 10,156 | | | | | |
| April 18th | 541,284 | 10,653 | 232 | 51,600,000 | 2.0% | | 0.02% |
| April 20th | 551,054[1] | 10,674 | 236 | 51,600,000 | 2.0% | 2.3% | 0.02% |

Source: Roser, Ritchie, Ortiz-Ospina and Hasell (2020)
[1]11,981 tests outstanding (2-3 days)

This rate could be low if some deaths are yet to occur or are not included in the statistics, but this is unlikely due to the small number of recent coronavirus cases and the government's coronavirus reporting practices. The data in the table show that in the two days prior to April 20th there were only 21 new cases and two new deaths. These numbers are so small that slight increases

would have no noticeable effect on the ratio of cumulative deaths/cumulative cases in mid-April. In addition, deaths are unlikely to be underestimated because South Korea attributes all deaths of persons who tested positive for the coronavirus to the virus, even if they had other underlying health conditions [AFP-JiJi, 2020].

The average death rate is very sensitive to the age distribution of infected individuals in a country because death rates are much higher for older patients. So the South Korean death rate is only applicable for the U.S. if the age distributions of coronavirus patients are similar in both countries.

Table 2 compares the age distribution of reported Coronavirus cases in South Korea and the U.S. in April, 2020. Since the available data for the two countries were reported for slightly different age categories, I created a distribution for the U.S. by decades for comparability with the decadal data available for South Korea. A comparison of the two distributions shows that the share of cases in the U.S. distribution is more heavily weighted toward the older age categories, with 40% of the cases in the over-60 group compared to only 25% for this group in South Korea.

Due to the testing limitations, the U.S. data on positive tests in April exclude a considerable share of the mild and moderate cases and all of the asymptomatic cases. Since these cases are more frequent in the young than in the old, their exclusion in the U.S. reduces the share of reported cases in the younger age categories. If the missing mild and moderate cases in the U.S. were added to the total cases, the U.S. distribution would become more like the South Korean distribution. Lacking data on a large share of coronavirus cases in the U.S., I assume for the analysis that the age distributions are similar in the two countries and that the South Korean death rate of 2.3% is applicable for the actual death rate in the U.S. when all cases of the coronavirus are included.

**Table 2**
**Age Distribution of Cases in South Korea and the U.S.**
**(Percent)**

| Age Categories | South Korea* | U.S.** |
|---|---|---|
| 0-9 | 1 | 1 |
| 10-19 | 5 | 4 |
| 20-29 | 27 | 11 |
| 30-39 | 11 | 11 |
| 40-49 | 13 | 15 |
| 50-59 | 18 | 18 |
| 60-69 | 13 | 17 |
| 70-79 | 7 | 13 |
| 80+ | 5 | 10 |

*Statista, 2020 for April 16th
**Sonnemaker and Kiersz, 2020

## IV. The Asymptomatic Share of Total Cases

My methodology does not require an explicit estimate of the asymptomatic share of all coronavirus cases. But an estimate for this portion of the cases is useful because it provides a

minimum estimate of the missing share of cases in the U.S. Few states have made any effort to perform contact tracing, so reported coronavirus cases in the U.S. rarely include asymptomatic cases.

Few countries manage to identify the asymptomatic cases. South Korea has included these cases in their estimates, but it has not provided data on the asymptomatic share, perhaps because this would require that investigators monitor asymptomatic cases to determine whether they become symptomatic later.

Iceland has carried out the only large screening study designed to determine the asymptomatic share of coronavirus cases in the population [Gudbjartsson, et al., 2020]. Two groups were tested that included 13,080 participants in total, of which one was self-selected and the other was picked at random. Only 17% of the participants were selected at random, but the share testing positive in this group was similar to the share in the entire sample. Out of the total participants in both groups, 100 individuals (0.6%) tested positive and were quarantined for 14 days, and 43% reported that they were asymptomatic during this period. If any of these individuals developed symptoms later, the asymptotic share would be lower, but this seems relatively unlikely after a 14-day quarantine.

The study excluded individuals with severe symptoms, such as those that would be hospitalized, so the 43% asymptotic share in the study is an overestimate of the actual share in the population. A WHO-China joint mission analyzed Chinese data for a large group of coronavirus patients and determined that about 20% of all symptomatic cases required hospitalization [Verity, 2020]. Assuming this ratio is applicable to Iceland, adding the severe cases raises the 57 symptomatic cases in the study to 71, which reduces the asymptomatic share to 38% (43/71+43). This seems to be the best currently available estimate of the asymptotic share of coronavirus cases.

In another study Nishiura et al. [2020] evaluated the asymptomatic share of cases in the 565 Japanese citizens evacuated from Wuhan and placed in quarantine for 14 days. Thirteen tested positive for coronavirus, of which nine had symptoms or developed them during the quarantine and four did not. The asymptomatic share of coronavirus patients is 31%, but the sample is too small to be very precise.

Mizumoto, et al.[2020] estimated the asymptomatic share of those passengers on the Diamond Princess cruise ship who tested positive. Since the initial testing for coronavirus was not random and the asymptomatic passengers could not be monitored after they left the ship, their estimate is heavily based on modeling assumptions. They estimate that the asymptomatic share of 53% at the time they left the ship probably declined to 18% later.

The estimate from the Icelandic study that 38% of all coronavirus cases are asymptotic is the most robust of the three estimates. Given the greater uncertainty associated with the Nishiura et al. and Mizumoto et al. estimates, the 38% figure appears to be consistent with all three studies.

## V. U.S. Data on the Share of Positive Tests and the Associated Death Rates

Several groups track the numbers of U.S. tests, confirmed cases, and deaths by state and publish their results several times a day. These data can be used to calculate the share of individuals that test positive and the death rates associated with the confirmed cases.

As discussed earlier, the number of reported cases of the coronavirus are a subset of the actual number of cases because the U.S. did not have sufficient testing capacity to test all of the individuals with coronavirus symptoms. The method of rationing the limited available tests was similar across states because the CDC issued guidelines for testing priorities [Connor, 2020]:

- Priority one: Hospitalized patients and symptomatic healthcare workers
- Priority two: High-risk patients with coronavirus symptoms
- Priority three: Symptomatic individuals in the community, if resources allow

None of these priorities include asymptomatic individuals.

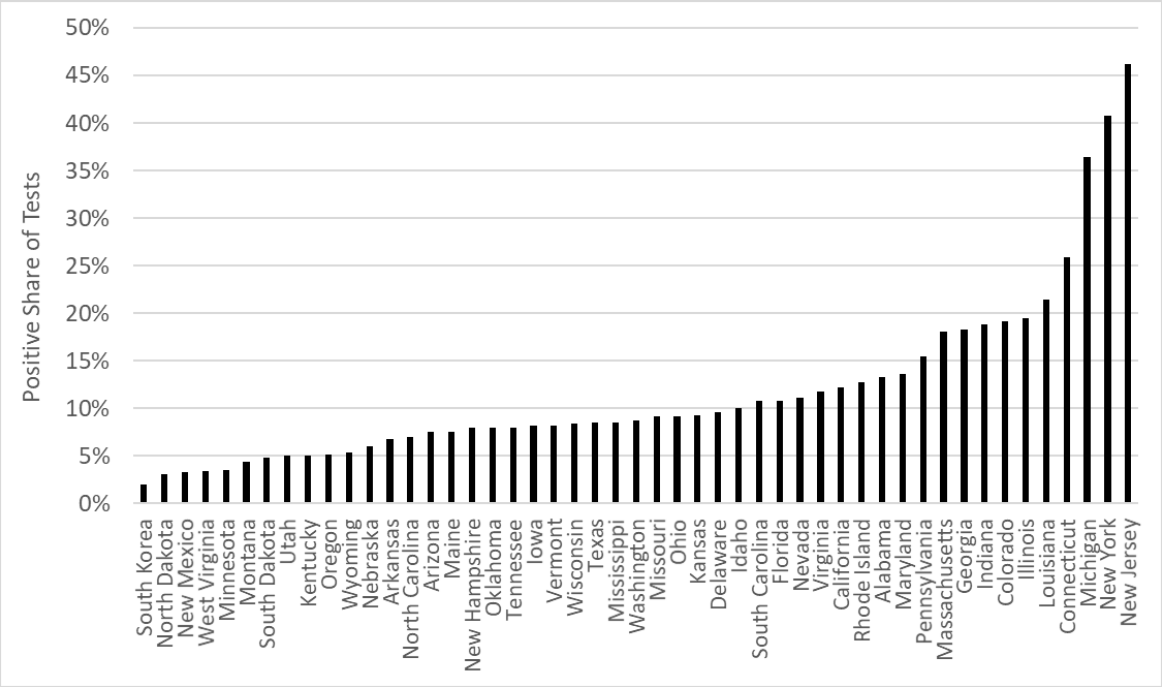
The likelihood that tests would be positive was highest for individuals in Priority one and declines in Priority two and three. As a result, as the number of tests increases, testing is extended to groups less and less likely to test positive, and the positive share of cumulative tests declines. States overwhelmed with coronavirus cases tested only a small fraction of all the individuals they thought were infected and obtained very high shares of positive tests.

Figure 2 shows the share of cumulative positive tests by state for the lower-48 states and South Korea on April 7th, using U.S. data published in *Politico* [Jin, 2020]. These data show that about half the states have shares of positive tests under 10%, with the lowest share at 3%. All of the states have a higher share of cumulative positive tests than South Korea (at 2%). Five states have shares of cumulative positive tests over 20%, and three states have shares over 30%. These three states, New Jersey, New York, and Michigan, are the ones who have reported tremendous limitations on their capacity to test individuals with coronavirus symptoms.

Maag [2020] reports that in late March very sick New Jersey residents spent many nights in their cars waiting in line at testing stations after being turned away in emergency rooms. In late March New York City restricted testing to hospitalized patients to prevent the many infected individuals with less serious symptoms from leaving home [Cuzey, 2020]. Michigan restricted testing to those with the most serious symptoms until testing was extended to those with mild symptoms in mid-April [Clarke, 2020].

At the beginning of April, Georgia, another state a with high shares of cumulative positive tests (18%), restricted tests to very sick patients, including those in group situations, such as nursing homes, and emergency and health care workers [Trubey, 2020]. Testing continued to be restricted to high-priority patients until mid-April due to a shortage of testing capacity [Bynum, 2020].

**Figure 2**
**Positive Share of Cumulative Coronavirus Tests on April 7, 2020**




Until mid-March even states with lower shares of cumulative positive tests, such as California and Rhode Island, reported that they could not test everyone who needed a test [Becker, 2020 and Mooney, 2020]. Even in mid-April 2020, after testing capacity in the U.S. improved relative to the number of suspected coronavirus cases, tests in most locations were still rationed to the higher priority applications [Connor, 2020].

Patients with severe symptoms and elderly and vulnerable patients with less severe symptoms were tested, and some states tested a large share of individuals with mild or moderate symptoms, but there was no contact tracing system that might lead to testing of asymptomatic individuals. Some patients died at home or in nursing homes without ever being tested. As a consequence, the confirmed cases reported by the states do not include asymptomatic cases and do not include an unknown share of patients with mild, moderate, and even severe symptoms.

Death rates (deaths/infected population) for coronavirus calculated from the confirmed cases for most states are much higher than the 2.3% rate in South Korea. Although death rates for the virus vary as a function of the patient's age and the quality and availability of medical care, there is no reason to believe that on average a higher share of Americans than South Koreans are likely to die from the virus. This means that the higher average death rates in the U.S. are likely to be due largely to the missing cases of infected individuals in the data.

If this is the case, then we should also expect to see a strong correlation between the share of cumulative positive tests and the death rates across states. Death rates could vary due to many

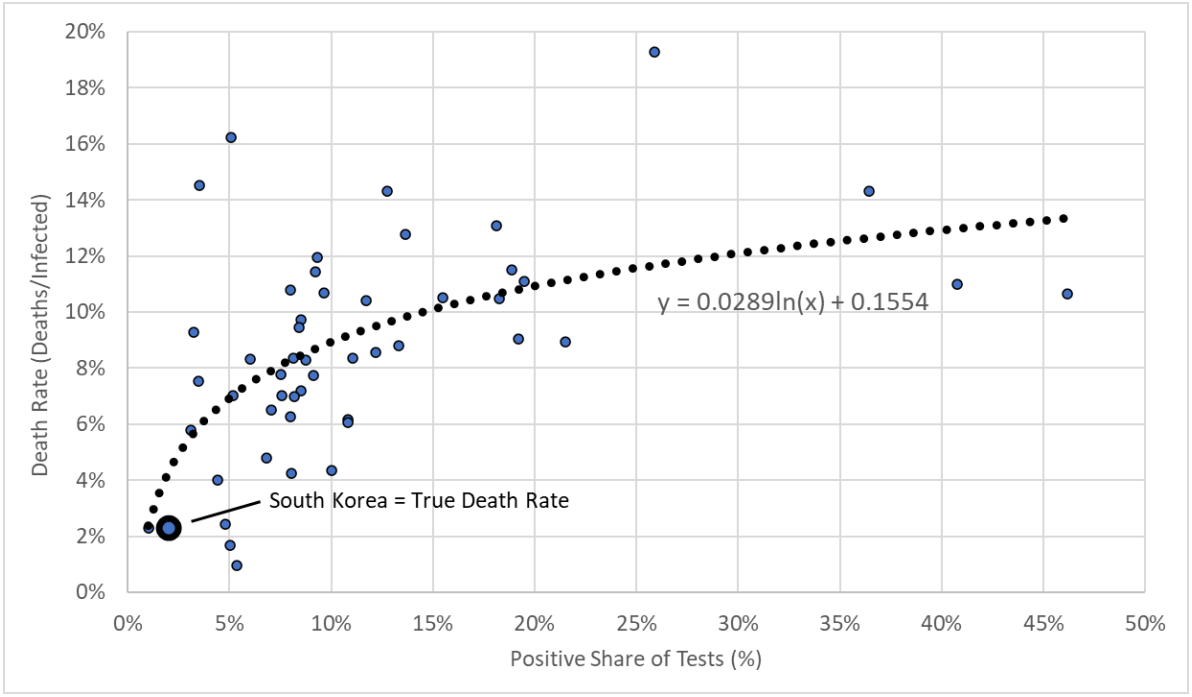
factors, but a higher share of individuals testing positive is almost certain to indicate that the more difficult coronavirus cases to identify are missing from the data.

Figure 3 shows a plot of the coronavirus death rate from cumulative reported cases and the share of cumulative positive tests for reported cases in the lower-48 states through April 7, 2020. The death rates are calculated using total cumulative deaths on April 21st divided by total cumulative positive cases on April 7th.

As discussed earlier, the average lag between the onset of symptoms and death is estimated to be about 19 days. Initially test results were only reported in the U.S. with very long lags, but in early April the lag became much shorter, as the public's awareness of the virus became widespread and private laboratories began to perform large numbers of tests on a more expedited basis. The reported lag for test results from LabCorp and Quest Diagnostics on April 8th was 3-4 days [Strickler and Kaplan, 2020]. With this lag for results, I assume that the average time lag between the onset of symptoms and test results on April 7th was 5 days. This leaves a 14-day average lag between reports of cases of coronavirus and the deaths associated with these cases.

I use Jin's [2020] data on cumulative tests and positive test results by state on April 7th and the New York Times' data on cumulative deaths by state on April 21st to calculate the share of positive tests and the death rates for the lower-48 states. Both sources report the positive test results, but neither includes all three measures. A comparison of their positive test results for the same days shows that the two data sets are similar, but not identical, which is not surprising given that the data are updated at different times over the course of each day.

**Figure 3**
**Apparent U.S. Death Rates vs. Positive Shares of Tests on April 7th**



The data for the death rate and the positive share of tests in the figure show a definite positive logarithmic relationship ($R^2 = 0.30$). As the positive share of cumulative tests rises across states for the period through April 7th, the calculated death rate rises, but at a decreasing rate. The share of cumulative positive tests associated with the assumed 2.3% true death rate in the trendline is 1%, which is close to the 2% share of positive tests at that death rate in South Korea. The estimated relationship of the death rate (deaths/infected population) to the share of cumulative positive tests is:

1) Death rate = 0.0289 ln(share of positive tests) + 0.1554

This relationship can be used to estimate the ratio of the actual infected population to the identified infected cases as a function of the positive share of cumulative tests in each state, assuming that there are no missing cases of infection when the death rate equals 2.3% (as in South Korea):

2) TotaI Cases/Identified Cases = 1.257 ln(positive share of tests) + 6.757

Equation (2) can be used to estimate the total number of coronavirus cases in each state. The sum of all the cases can be used to estimate the infected share of the U.S. population. These results are shown in the next section.

Equation (2) is based on the assumption that the reported deaths due to the coronavirus in the U.S. are as complete and comparable to the reported deaths in South Korea. This is almost certainly not the case because many deaths in the U.S. involved elderly patients who died at home or in nursing homes without ever being tested for the virus. Until recently, these deaths were often attributed to pneumonia, to some underlying condition, or to some fatal physiological response to the virus, such as a stroke.

As discussed earlier, South Korea attributed any deaths of patients who tested positive to the coronavirus to the virus. Medical providers in the U.S. could not and did not do this because they tested only a small fraction of the patients who had the disease.

Coroners in many countries are now reviewing deaths that occurred over the last few months attributed to other causes to determine whether COVID-19 was the actual cause of death, and they are raising the number of deaths attributed to the virus. As coroners in the U.S. raise their estimates of the deaths that were due to the coronavirus, states may or may not include these revisions in the statistics reported to the CDC. Even if they do, these deaths show up later, so the cumulative deaths included in the *New York Times* data for a particular date may underestimate the total number of deaths [Wu, McCann, Katz, and Peltier, 2020].

The deaths reported for New York by the *New York Times* include those associated with patients who tested positive for the coronavirus in New York City, but apparently do not include deaths that the City is now saying were probably caused by the virus. There appear to be about 5,000 of these probable deaths through April 21st [Bui, Katz, Parlapiano, and Sanger-Katz, 2020]. Including these deaths to calculate the cumulative New York death rate raises the rate from 11.0% to 14.8%, which would slightly raise the slope of the trendline shown in Figure 3. This change for this one state does not have much effect on the estimated relationship in Equations (1) and (2), as long as the rates in other states are not changed.

No statistics are available to estimate the magnitude of the missing deaths in the U.S. data for all the states, but some researchers are comparing the recent surge in mortality to the deaths attributed to the virus to see whether the underreporting of deaths due to the virus could be substantial. The *New York Times* has estimated that the surge in mortality in New York City in recent months was 30% larger (17200/13240) than the reported deaths attributed to the coronavirus [Wu, McCann, Katz, and Peltier, 2020]. New York City's recent addition of probable deaths from the coronavirus implicitly attributes all of the recent surge in mortality to the virus.

Of course, even if these additional deaths in NYC are due to the virus, this does not imply that the magnitude of this increase is relevant for the death statistics reported throughout the U.S. What is clear is that the deaths reported by the states to estimate the U.S. death rates in the analysis are underestimated by some amount, which means that the ratio of total cases/identified cases in Equation (2) is underestimated. Still, the estimates of total cases/identified cases produced by the Equation are sufficiently accurate to be indicative of the actual total cases in the country.

Another way to evaluate whether Equation (2) is reasonable is to apply it to the states with the lowest and highest shares of positive tests to see if the estimated sizes of the unidentified share of the infected population seem reasonable, given the known testing conditions in those states.

Applying equation (2) to North Dakota, one of the states with the lowest share of positive tests (3.1%) on April 7th, the equation indicates that the state identified 42% of the total cases of infection and failed to identify 58% of these cases. Since 38% of the cases are estimated to be asymptomatic, if the state did not identify any of these cases, then it identified 68% of the symptomatic cases (42/62). Given the extensive testing relative to the number of symptomatic cases in this state, this estimate seems reasonable.

Applying equation (2) to New Jersey, the state with the highest share of positive tests (46.2%) on April 7th, the equation indicates that the state identified 17% of the total cases of infection and failed to identify 83% of these cases. Since 38% of the cases are estimated to be asymptomatic, if the state did not identify any of these cases, then it identified 27% of the symptomatic cases (17/62). Given that about 20% of all symptomatic cases require hospitalization, if the state identified all of these cases, then the implication is that it only managed to identify 8% of the existing cases with mild or moderate symptoms. Given the reports on the tremendous difficulty encountered by individuals seeking tests in New Jersey in late March, this estimate also seems reasonable.

The data used to estimate Equation (1) are included in the Appendix.

**VI. Estimated Total and Unidentified Coronavirus Cases**

Table 2 presents the results for an application of the model using the data on April 21st for the share of positive test results in each state. The state results are used to calculate results for the U.S. as a whole. For each state the table shows the ratio of total cases/reported cases, the share of unidentified cases, and the infected share of the population.

What is striking is that most cases in every state are unidentified. Overall, only 20% of all cases have been confirmed through testing, and 80% have not been identified. The unidentified share ranges from 61% to 83% across states. The total number of unidentified cases in the U.S. was

about 3 million. This finding demonstrates how difficult it would be to contact trace all the cases in the U.S. at the current stage of the epidemic.

The rate of infection in the U.S. varies dramatically across states, from about 0.1% in Wyoming, Montana, West Virginia, Minnesota, and Oregon to 7.0% in New York. Contact tracing would be feasible, at least in theory, in states with the lower rates of infection, as South Korea managed to do it with a 0.2% infected share of the population.

**Table 2**
**Unidentified Coronavirus Cases and Share of U.S. Population Infected on April 21,2020**

| States | Positive | Tests | Pos share | Tot/Ident | Tot Cases | Unidentified | Population | Infected |
|---|---|---|---|---|---|---|---|---|
| Alabama | 5,025 | 45,900 | 10.9% | 3.98 | 19982 | 74.9% | 4,858,979 | 0.4% |
| Arizona | 5,064 | 54,500 | 9.3% | 3.77 | 19093 | 73.5% | 6,828,065 | 0.3% |
| Arkansas | 1,923 | 26,553 | 7.2% | 3.46 | 6648 | 71.1% | 2,978,204 | 0.2% |
| California | 30,978 | 290,500 | 10.7% | 3.94 | 122159 | 74.6% | 39,144,818 | 0.3% |
| Colorado | 9,730 | 46,195 | 21.1% | 4.80 | 46695 | 79.2% | 5,456,574 | 0.9% |
| Connecticut | 19,815 | 62,806 | 31.5% | 5.31 | 105156 | 81.2% | 3,590,886 | 2.9% |
| Delaware | 2,745 | 16,470 | 16.7% | 4.50 | 12366 | 77.8% | 945,934 | 1.3% |
| Florida | 26,660 | 266,225 | 10.0% | 3.86 | 103025 | 74.1% | 20,271,272 | 0.5% |
| Georgia | 18,947 | 84,328 | 22.5% | 4.88 | 92465 | 79.5% | 10,214,860 | 0.9% |
| Idaho | 1,672 | 17,445 | 9.6% | 3.81 | 6369 | 73.7% | 1,654,930 | 0.4% |
| Illinois | 31,508 | 148,358 | 21.2% | 4.81 | 151535 | 79.2% | 12,859,995 | 1.2% |
| Indiana | 11,686 | 64,639 | 18.1% | 4.61 | 53837 | 78.3% | 6,619,680 | 0.8% |
| Iowa | 3,159 | 25,820 | 12.2% | 4.12 | 13003 | 75.7% | 3,123,899 | 0.4% |
| Kansas | 1,986 | 18,761 | 10.6% | 3.93 | 7813 | 74.6% | 2,911,641 | 0.3% |
| Kentucky | 2,960 | 32,572 | 9.1% | 3.74 | 11077 | 73.3% | 4,425,092 | 0.3% |
| Louisiana | 24,523 | 142,099 | 17.3% | 4.55 | 111544 | 78.0% | 4,670,724 | 2.4% |
| Maine | 875 | 14,951 | 5.9% | 3.19 | 2791 | 68.6% | 1,329,328 | 0.2% |
| Maryland | 13,684 | 71,397 | 19.2% | 4.68 | 64047 | 78.6% | 6,006,401 | 1.1% |
| Massachusetts | 38,077 | 162,241 | 23.5% | 4.94 | 187910 | 79.7% | 6,794,422 | 2.8% |
| Michigan | 32,000 | 113,798 | 28.1% | 5.16 | 165192 | 80.6% | 9,922,576 | 1.7% |
| Minnesota | 2,470 | 46,850 | 5.3% | 3.06 | 7553 | 67.3% | 5,489,594 | 0.1% |
| Mississippi | 4,512 | 51,434 | 8.8% | 3.70 | 16685 | 73.0% | 2,992,333 | 0.6% |
| Missouri | 5,807 | 56,013 | 10.4% | 3.91 | 22694 | 74.4% | 6,083,672 | 0.4% |
| Montana | 433 | 11,051 | 3.9% | 2.68 | 1163 | 62.8% | 1,032,949 | 0.1% |
| Nebraska | 1,474 | 15,680 | 9.4% | 3.78 | 5579 | 73.6% | 1,896,190 | 0.3% |
| Nevada | 3,830 | 32,347 | 11.8% | 4.07 | 15607 | 75.5% | 2,890,845 | 0.5% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| New Hampshire | 1,392 | 14,118 | 9.9% | 3.84 | 5352 | 74.0% | 1,330,608 | 0.4% |
| New Jersey | 88,806 | 178,057 | 49.9% | 5.88 | 522407 | 83.0% | 8,958,013 | 5.8% |
| New Mexico | 1,845 | 37,042 | 5.0% | 2.99 | 5510 | 66.5% | 2,085,109 | 0.3% |
| New York | 247,512 | 633,861 | 39.0% | 5.57 | 1379868 | 82.1% | 19,795,791 | 7.0% |
| North Carolina | 6,764 | 79,484 | 8.5% | 3.66 | 24755 | 72.7% | 10,042,802 | 0.2% |
| North Dakota | 627 | 14,747 | 4.3% | 2.79 | 1748 | 64.1% | 756,927 | 0.2% |
| Ohio | 12,516 | 90,436 | 13.8% | 4.27 | 53457 | 76.6% | 11,613,423 | 0.5% |
| Oklahoma | 2,680 | 35,646 | 7.5% | 3.50 | 9391 | 71.5% | 3,911,338 | 0.2% |
| Oregon | 1,956 | 40,045 | 4.9% | 2.96 | 5794 | 66.2% | 4,028,977 | 0.1% |
| Pennsylvania | 33,232 | 162,952 | 20.4% | 4.76 | 158133 | 79.0% | 12,802,503 | 1.2% |
| Rhode Island | 5,090 | 37,080 | 13.7% | 4.26 | 21688 | 76.5% | 1,056,298 | 2.1% |
| South Carolina | 4,377 | 40,480 | 10.8% | 3.96 | 17337 | 74.8% | 4,896,146 | 0.4% |
| South Dakota | 1,685 | 12,326 | 13.7% | 4.26 | 7171 | 76.5% | 858,469 | 0.8% |
| Tennessee | 7,238 | 100,689 | 7.2% | 3.45 | 24955 | 71.0% | 6,600,299 | 0.4% |
| Texas | 19,458 | 190,394 | 10.2% | 3.89 | 75691 | 74.3% | 27,469,114 | 0.3% |
| Utah | 3,213 | 68,311 | 4.7% | 2.91 | 9364 | 65.7% | 2,995,919 | 0.3% |
| Vermont | 816 | 12,981 | 6.3% | 3.28 | 2676 | 69.5% | 626,042 | 0.4% |
| Virginia | 8,990 | 56,735 | 15.8% | 4.44 | 39927 | 77.5% | 8,382,993 | 0.5% |
| Washington | 11,790 | 138,642 | 8.5% | 3.66 | 43139 | 72.7% | 7,170,351 | 0.6% |
| West Virginia | 902 | 22,155 | 4.1% | 2.73 | 2465 | 63.4% | 1,844,128 | 0.1% |
| Wisconsin | 4,499 | 51,102 | 8.8% | 3.70 | 16658 | 73.0% | 5,771,337 | 0.3% |
| Wyoming | 313 | 7,386 | 4.2% | 2.78 | 871 | 64.1% | 586,107 | 0.1% |
| **Total** | **767,244** | **3943602** | **19.5%** | **4.70** | **3800346** | **79.8%** | **318,576,557** | **1.2%** |

## VII. Conclusions

The U.S. is testing to identify cases of the coronavirus, but the testing has been completely inadequate to determine the extent of infection in the population. This study uses the death rate and the share of positive tests in each state to estimate the share of cases in each state that have been identified and the size of the infected population. The results indicate that no state has managed to

identify even half of the cases, so it is difficult to control the spread. But it is also clear that not all states have been severely affected by the virus, with rates of infection on April 21, 2020 ranging from 0.1% to 7.0% of the population.

Overall 3.8 million people have been infected, which is 1.2 percent of the population. This is nowhere near a high enough share to create herd immunity, but it is too many people to permit contact tracing for all of them in the states with the highest infection rates. The wide differences in rates of infection indicate that different strategies are appropriate to manage the virus in different states.

## References


AFP-JiJi, 2020, "In the Coronavirus Pandemic, Counting the Dead is a Difficult Process," *The Japan Times*, April 11, 2020

Becker, Rachel, 2020, "Where California Stands with Coronavirus Testing Right Now," *Cal Matters*, March 18, 2020

Boyd, Conner, 2020, "The coronavirus death lag explained: How it can take three weeks between catching the disease and being hospitalised (and three days for the NHS to record the fatality)" *Daily Mail.com*, April 6, 2020, https://www.dailymail.co.uk/news/article-8192993/The-coronavirus-death-lag-explained-weeks-fatality-recorded.html

Bui, Quoctrung, Katz, Josh, Parlapiano, Alicia, and Anger-Katz, Margot, 2020, "What 5 Coronavirus Models Say the Next Month Will Look Like," The Upshot, *New York Times*, April 22, 2020,

Bynum, Russ, 2020, "Georgia Expanding Coronavirus Testing as Deaths Pass 550," *The Union-Recorder*, April, 2020

Clarke, Kyla, 2020, "Michigan Officials Extend Coronavirus (COVID-19) Testing to Those with Mild Symptoms," *Click on Detroit*, April 14, 2020

Cohen, Elizabeth, 2020, "Prominent Scientists Have Bad News for the White House about Coronavirus Antibody Tests, CNN Health, April 15, 2020

Connor, Katie, 2020, "Can You Get Tested for Coronavirus Today? Here's Who Qualifies for COVID-19 Testing," C/NET, April 19, 2020

Cuzey, Bobby, 2020, "Why NYC Says Widespread Testing for Coronavirus May Be Counterproductive,"*Spectrum News*," April 6, 2020

DW, 2020, "Up to 30% of Coronavirus Cases Asymptomatic," March 24, 2020 https://www.dw.com/en/up-to-30-of-coronavirus-cases-asymptomatic/a-52900988

Eder, Steve, Twohey, Megan, and Mandavilli, Apoorva, 2020, "Antibody Test, Seen as Key to Reopening Country, Does Not Yet Deliver," *New York Times*, April 19, 2020

Government of the Republic of Korea, 2020, *Flattening the Curve on COVID-19, How Korea Responded to a Pandemic using ICT*, April 15, 2020
http://www.mdon.co.kr/news/download.html?no=26757&atno=92301

Gudbjartsson, Daniel F., Helgason, Agnar, Jonsson, Hakon, Magnusson, Olafur T., Melsted, Pall, Norddahl, Gudmundur L., Saemundsdottir, Jona, Sigurdsson, Asgeir, Sulem, Patrick, Agustsdottir, Arna B., Eiriksdottir, Berglind, Fridriksdottir, Run, et al., 2020, "Spread of SARS-CoV-2 in the Icelandic Population," *The New England Journal of Medicine, DOI:10.1056/NEJMoa2006100*

Jin, Beatrice, 2020, "Live Tracker: How Many Coronavirus Cases Have Been Reported in Each State?," *Politico*, April 10, 2020 6:55 PM, https://www.politico.com/interactives/2020/coronavirus-testing-by-state-chart-of-new-cases/

Maag, Cristopher, 2020, "New Jersey Residents Camp Out All Night for Coronavirus Tests," *USA Today*, March 25, 2020

Mizumoto, Kenji, Kagaya, Katsushi, Zarebski, Alexander, and Chowell, Gerardo, 2020,"Estimating the Asymptomatic Proportion of Coronavirus Disease 2019 (COVID-19) Cases on Board the Diamond Princess Cruise Ship, Yokohama, Japan, 2020, *Eurosurveillance*, 25(10): 2000180.

Mooney, Tom, 2020, "23 Cases of Coronavirus in Rhode Island; Testing Limited by National Supply Shortage," *The Providence Journal*, March 17, 2020

Nishiura,Hiroshi, Kobayashi, Tetsuro, Miyama,Takeshi, Suzuki, Ayako, Jung, Sungmok, Hayashi, Katsuma, Kinoshita, Ryo, Yichi Yang, Yuan, Baoyin, Akhmetzhanov, Andrei R., Linton, Natalie M., 2020, "Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19)," *International Journal of Infectious Diseases,* doi: 10.1016/j.ijid.2020.03.020

Roser, Max, Ritchie, Hannah, Ortiz-Ospina, Esteban, and Hasell, Joe, 2020, "Coronavirus Disease (COVID-19) – Statistics and Research". Published online at *OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]

Silverman, Justin D., Hupert, Nathaniel, and Washburne, Alex D., 2020, "Using Influenza Surveillance Networks to Estimate State-Specific Case Detection Rates and Forecast SARS-CoV-2 Spread in the U.S., medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542

Sonnemaker, Tyler, and Kiersz, Andy, 2020, "Nearly 30% of Coronavirus Cases Have Been Among People 20-44 the CDC Says - Showing that Young People Are Getting Sick Too," *Business Insider*, March 19, 2020

Statista, 2020, "Age Distribution of Coronavirus (COVID-19) in South Korea on April 21, 2020," https://www.statista.com/statistics/1102730/south-korea-coronavirus-cases-by-age/

Strickler, Laura, and Kaplan, Adiel, 2020, "Private Labs Do 85% of Covid-19 Tests, but still struggle with backlogs, shortages," *NBC News online*, April 8, 2020

Trubey, J. Scott, "Georgia Playing Catchup in Coronavirus Testing", *The Atlanta Journal-Constitution*, April 1, 2020

Verity, Robert, Okell, Lucy C., Dorigatti, Ilaria, Winskill, Peter, Whittaker, Charles, Imai, Natsuko, et al. "Estimates of the Severity of the Coronavirus Disease 2019: a Model-based Analysis," *The Lancet Infectious Diseases*, March 30,2020, https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930243-7

Wu, Jin, McCann, Allison, Katz, Josh, and Peltier, Elian, 2020, "28,000 Missing Deaths: Tracking the True Toll of the Coronavirus," *New York Times*, March 22, 2020 https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html

Zimmer, Carl, 2020, "Most New York Coronavirus Cases Come from Europe, Genomes Show," *New York Times* online, April 8, 2020

**Appendix**
**Data Used to Estimate the Model on April 7, 2020**

| | Positive | Tests | Positive share | Deaths | Death Rate |
|---|---|---|---|---|---|
| Alabama | 1,968 | 14,765 | 13.3% | 173 | 8.8% |
| Arizona | 2,456 | 32,534 | 7.5% | 191 | 7.8% |
| Arkansas | 875 | 12,845 | 6.8% | 42 | 4.8% |
| California | 14,336 | 117,431 | 12.2% | 1225 | 8.5% |
| Colorado | 4,950 | 25,773 | 19.2% | 448 | 9.1% |
| Connecticut | 6,906 | 26,686 | 25.9% | 1331 | 19.3% |
| Delaware | 673 | 6,994 | 9.6% | 72 | 10.7% |
| Florida | 13,324 | 123,274 | 10.8% | 822 | 6.2% |
| Georgia | 7,314 | 40,012 | 18.3% | 767 | 10.5% |
| Idaho | 1,101 | 10,995 | 10.0% | 48 | 4.4% |
| Illinois | 12,262 | 62,942 | 19.5% | 1359 | 11.1% |
| Indiana | 4,944 | 26,191 | 18.9% | 569 | 11.5% |
| Iowa | 946 | 11,599 | 8.2% | 79 | 8.4% |
| Kansas | 845 | 9,084 | 9.3% | 101 | 12.0% |
| Kentucky | 955 | 18,767 | 5.1% | 155 | 16.2% |
| Louisiana | 14,867 | 69,166 | 21.5% | 1328 | 8.9% |
| Maine | 499 | 6,587 | 7.6% | 35 | 7.0% |
| Maryland | 4,045 | 29,617 | 13.7% | 516 | 12.8% |
| Massachusetts | 13,837 | 76,429 | 18.1% | 1809 | 13.1% |
| Michigan | 17,221 | 47,251 | 36.4% | 2466 | 14.3% |
| Minnesota | 986 | 28,128 | 3.5% | 143 | 14.5% |
| Mississippi | 1,738 | 20,370 | 8.5% | 169 | 9.7% |
| Missouri | 2,722 | 29,835 | 9.1% | 211 | 7.8% |
| Montana | 299 | 6,790 | 4.4% | 12 | 4.0% |
| Nebraska | 409 | 6,787 | 6.0% | 34 | 8.3% |
| Nevada | 1,953 | 17,629 | 11.1% | 163 | 8.3% |
| New Hampshire | 669 | 8,370 | 8.0% | 42 | 6.3% |
| New Jersey | 41,090 | 89,032 | 46.2% | 4377 | 10.7% |
| New Mexico | 624 | 19,136 | 3.3% | 58 | 9.3% |
| New York | 130,689 | 320,811 | 40.7% | 14347 | 11.0% |
| North Carolina | 2,870 | 40,726 | 7.0% | 187 | 6.5% |
| North Dakota | 225 | 7,213 | 3.1% | 13 | 5.8% |
| Ohio | 4,450 | 48,378 | 9.2% | 509 | 11.4% |
| Oklahoma | 1,327 | 16,588 | 8.0% | 143 | 10.8% |
| Oregon | 1,068 | 20,624 | 5.2% | 75 | 7.0% |
| Pennsylvania | 12,980 | 83,854 | 15.5% | 1366 | 10.5% |
| Rhode Island | 1,082 | 8,481 | 12.8% | 155 | 14.3% |
| South Carolina | 2,049 | 18,976 | 10.8% | 124 | 6.1% |
| South Dakota | 288 | 6,020 | 4.8% | 7 | 2.4% |

| | | | | | |
|---|---|---|---|---|---|
| Tennessee | 3,802 | 47,350 | 8.0% | 162 | 4.3% |
| Texas | 7,276 | 85,357 | 8.5% | 523 | 7.2% |
| Utah | 1,675 | 33,394 | 5.0% | 28 | 1.7% |
| Vermont | 543 | 6,633 | 8.2% | 38 | 7.0% |
| Virginia | 2,878 | 24,521 | 11.7% | 300 | 10.4% |
| Washington | 7,984 | 91,375 | 8.7% | 661 | 8.3% |
| West Virginia | 345 | 9,940 | 3.5% | 26 | 7.5% |
| Wisconsin | 2,440 | 29,014 | 8.4% | 231 | 9.5% |
| Wyoming | 210 | 3,929 | 5.3% | 2 | 1.0% |
| Total | 358,995 | 1,898,203 | 18.9% | 37,642 | 10.5% |