Estimating Undetected COVID-19 Infections

March 29, 2021

Abstract

Having an accurate estimate of total SARS-CoV-2 infections is critical for informing public health decisions, distributing vaccines, and ultimately optimizing social and economic well-being of the country. However, the large number of undetected infections due to testing shortages or infected individuals not seeking a test, makes it challenging to estimate the total number of cases. We specify and estimate a time-varying Markov model of COVID-19 cases. According to our estimation, 22.8% of the US population has been infected as of Nov 29, 2020, which is more than five times the number of officially confirmed and reported cases. The estimated level of undetected infections spiked in March and started to decline beginning in late April though it was not until July that it was exceeded by the detected cases. Our results suggest that the substantial increase in testing capacity in the US has identified a higher percentage of infections. Our model provides estimates of undetected infections that are plausible and consistent with other published estimates, while having the advantage of simplicity and ease of estimation with widely available data.

1 Introduction

Accurate estimates of total coronavirus disease 2019 (COVID-19) cases can help planners make decisions about testing policy and economic openness, let business leaders better understand risks to their workers and customers, and inform economic projections. However, one of the challenges facing policymakers, business leaders, and the general public in understanding the spread of the severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) is the fact that many infected cases go undetected because of testing shortages or infected individuals not seeking a test, for example, asymptomatic individuals may not even consider the need for a test Wu et al. (2020). As vaccines become increasingly available, accurate estimates of past and current undetected COVID-19 cases become more important as it leads to better estimates of cumulative total infections, which is critical for distributing vaccines. Specifically, in the context of a vaccination program, those already effected may already have immunity (if only temporarily) and so-called "herd immunity" could be achieved more quickly by deferring vaccination of those already infected Randolph and Barreiro (2020). Unfortunately, the number of undetected cases, while hard to estimate, is much larger than the confirmed cases due to the vast amount of asymptomatic patients, which significantly undermines estimations of the total number of cases. As shown in a previous study Friedman et al. (2020), most models predict the total number to be at least two to three times larger than the confirmed cases.

In fact, at the early stage of the pandemic, the number of positive tests in the US grew steadily faster than the number of hospitalizations. Likewise, hospitalizations have grown more quickly than deaths attributed to COVID-19. A very simple way to understand the disconnect between deaths and reported new cases is to estimate the total number of cases nationwide using lagged data on the number of deaths and

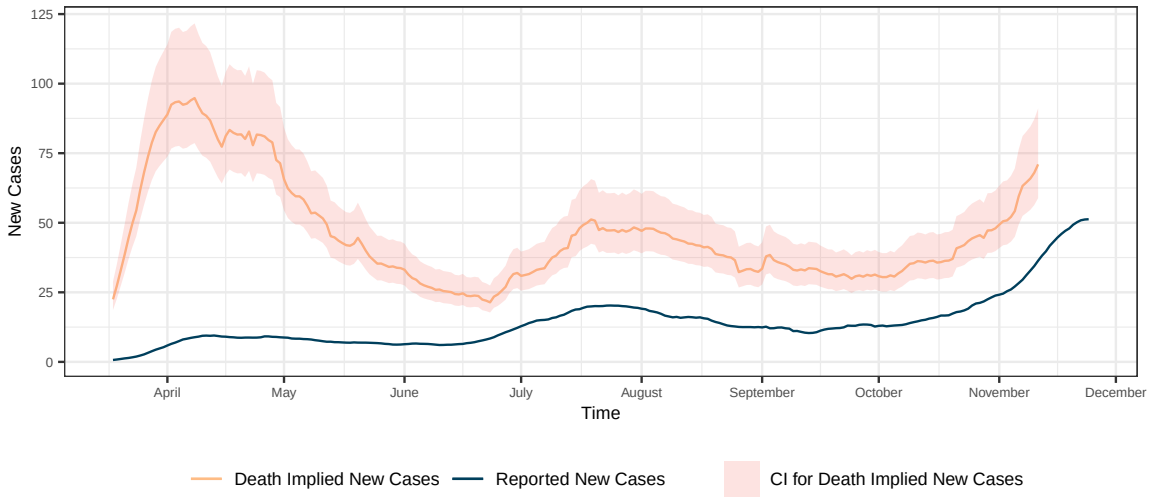


Figure 1: **New Cases in the US (per 100,000, 7-day moving average).**
The red line in the figure shows the death implied new cases. The shaded red area shows the estimated confidence band. The navy line in the figure shows the reported new cases.

the recent estimated infection fatality rates. Figure 1 shows the "death implied" new cases, which are calculated using the 7-day average of new reported deaths in the US lagged by 14 days (to reflect the average time between contracting COVID-19 and death) divided by the infection fatality rate of 0.68% estimated by Meyerowitz-Katz and Merone (2020). The confidence band of the "death-implied" new cases is calculated using the 7-day average of new reported deaths in the US lagged by 14 days divided by the 95% confidence interval of the estimated infection fatality rate. The "death-implied" estimates suggest that the number of new cases in the US rose rapidly in March, then levelled off and started to decrease in April. This pattern is obviously at odds with the number of new positive tests which was quite low comparing to the "death-implied" estimates until late June.

To estimate the gap between observed and total cases, we use a variant of a standard time-varying Markov model to infer the number of undetected cases using easily observable data on reported cases, hospitalizations and deaths at the state and na-

tional level. We examine a standard 5-state time-varying Markov model based on the work by Gouriéroux and Jasiak (2020) (and cites therein) and extend the model by introducing two conditioning variables, testing positivity rate and the intensity of testing (i.e., tests conducted per 100,000 population). In our model the population is either susceptible (S), infected and undetected (IU), infected and detected (ID), hospitalized (H), or deceased (D). States are mutually exclusive so we track hospitalized separately from infected and detected. Recovered cases re-enter the susceptible pool to avoid having another unobservable recovered state but will have little impact on estimation for low levels of overall infection.

The model is estimated on the COVID-19 propagation data of the US and nine individual states over the period of 271 days between March 4 to November 29, 2020. We examine two versions of the model with different assumptions on transition probability to state H and obtain similar results from both. The model provides estimates of undetected infections and total infections and fits observed levels of positive cases, hospitalizations and deaths well.[1] We find that the conditioning variables are important factors in the estimate with intuitive relations to infection probabilities. Of course, other models have been proposed for estimating the number of undetected infections and we compare our results to some of these works. Our estimates of undetected infections and total infections are consistent with other published estimates Friedman et al. (2020) while in comparison, our model has the advantage of simplicity and ease of estimation.[2]

[1] Our estimates are consistent with the pattern of "death-implied" new cases shown in Figure 1.
[2] The Friedman et al. (2020) are available at `https://ourworldindata.org/covid-models`.

2 Model specification

The latent individual history variable $Y_{i,t}$, for individual $i = 1, ..., N$ at time $t = 1, \ldots, T$, is qualitative polytomous with J alternative states denoted by $j = 1, \ldots, J$. As in the work by Gouriéroux and Jasiak (2020), we assume that $Y_{i,t}$ have the same marginal distribution across all individuals $i = 1, ..., N$ at t fixed, which can be summarized by the J-dimensional vector $p(t)$. The j-th component of the marginal distribution is $p_j(t) = P(Y_{i,t} = j)$. In addition, the individual history variable follows a Markov process with time-varying transition matrix $P\left[p(t-1); \theta\right]$, which gives

$$p(t) = P[p(t-1); \theta]' p(t-1), t = 2, ..., T, \tag{1}$$

with θ being a vector of parameters. The data pertaining to the individual history variable $Y_{i,t}$ may not be available in practice, and only aggregate frequencies for some of the states are available. With the assumptions of independent individual histories and homogeneous population of risks, the J-dimensional cross-sectional frequency vector $f(t)$, where $f_j(t)$ is the state j frequency of the population, can be seen as the sample counterpart of $p(t)$.

However, the cross-sectional frequencies are only partially observed. A state aggregation matrix A is used to account for the unobserved states and the observations are $\hat{A}_t = Af(t)$ for $t = 1, ..., T$, where A is a $K \times J$ matrix of full rank K. The parameters of interest, θ and the sequence of the unobserved component of $p(t)$, can then be estimated by solving the following optimization problem (where $\|.\|_2$ is the

Euclidean norm):

$$(\hat{p}(1), ..., \hat{p}(T), \hat{\theta}) = \text{argmin} \sum_{t=2}^{T} \|p(t) - P\left[p(t-1), \theta\right]' p(t-1)\|_2^2 \tag{2}$$

$$\text{s.t. } Ap(t) = Af(t) = \hat{A}_t, t = 1, ..., T.$$

To model the COVID-19 propagation, we consider a Markov process with 5 states: $1 = S$, for susceptible, $2 = IU$, for Infected and Undetected, $3 = ID$, for Infected and Detected, $4 = H$ for Hospitalized, and $5 = D$ for Deceased (due to COVID-19). The sum of the frequencies across all the five states at any given time t equals to the size of the initial population. For simplicity, we assume no immunity in our estimation, hence the recovered cases re-enter the susceptible pool. This assumption lets us avoid having an unobservable recovered state but will have little impact on estimation for low levels of overall infection. The transition matrix $P[p(t-1); \theta]$ of the Markov process is defined as

$$\begin{array}{c} \\ \text{S} \\ \text{IU} \\ \text{ID} \\ \text{H} \\ \text{D} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{S} & \text{IU} & \text{ID} & \text{H} & \text{D} \end{array} \\ \begin{bmatrix} 1-p_{i,t} & p_{i,t}(1-p_{d,t}) & p_{i,t}p_{d,t} & 0 & 0 \\ p_{21} & (1-p_{21}-p_{24})(1-p_{d,t}) & (1-p_{21}-p_{24})p_{d,t} & p_{24} & 0 \\ p_{31} & 0 & 1-p_{31}-p_{34} & p_{34} & 0 \\ p_{41} & 0 & 0 & 1-p_{41}-p_{45} & p_{45} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

with $p_{i,t} = \text{logist}\left(a_1 + a_2\left(p_2(t-1) + p_3(t-1)\right)\right) + a_3 x_t)$, $p_{d,t} = \text{logist}\left(b_1 + b_2 y_t\right)$, where $\text{logist}(x) = 1/[1+\exp(-x)]$ is the logistic function, i.e. the inverse of the logit function. The probability of infected $p_{i,t}$ follows a multinomial logit model for the competing propagation driven by lagged IU and lagged ID, and it also depends on the testing positivity rate x_t. Conditioning on being infected, the probability of being detected $p_{d,t}$ is a function of testing intensity y_t. Each row of the transition matrix sums to one by construction. The structure of zeros indicates that one cannot go backward from ID to IU, patients who died are hospitalized before death, the hospitalized patients

will stay in hospital until they recover or die, and death is considered an absorbing state.

In addition, we consider two model specifications for the transition probabilities from state IU and ID to state H. The basic specification assumes constant transition probabilities p_{24} and p_{34}. In this model, there are 11 parameters in $\theta = [a_1, a_2, a_3, b_1, b_2, p_{21}, p_{24}, p_{31}, p_{34}, p_{41}, p_{45}]'$. The full specification assumes time-varying transition probabilities driven by the lagged frequency of the corresponding state with $p_{24} = \text{logist}\left(c_1 + c_2 p_2(t-1)\right)$, $p_{34} = \text{logist}\left(d_1 + d_2 p_3(t-1)\right)$, and in which θ has 13 parameters.[3] The results from these two versions of the model are very similar so we only report the results from the basic model (but results from the full model are available on request).

Empirically, $IU(t)$ and $ID(t)$ represent the state of currently infected excluding those hospitalized. The frequency of $ID(t)$ is observable by assumption, while $IU(t)$ is the unobserved state of unidentified infections and will be considered as additional quantities of interest to be estimated jointly. Also, the frequencies of $H(t)$ and $D(t)$ are both observable. Therefore, we have the state aggregation matrix A expressed as

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

2.1 Data

The historical data of COVID-19 propagation for the US and each state is from *The Covid Tracking Project*, see https://covidtracking.com/data. The frequency of

[3] They are $[a_1, a_2, a_3, b_1, b_2, c_1, c_2, d_1, d_2, p_{21}, p_{31}, p_{41}, p_{45}]'$.

$ID(t)$ is measured by the rolling 2-week sum of the new positive tests in the US, which assumes that a person with positive test will either be hospitalized or recover within 14 days. The frequency of $H(t)$ is the actual number of hospitalized in US on any given date and the frequency of the absorbing state $D(t)$ is measured by the cumulative deaths caused by COVID-19 in the US. In constructing the cross-sectional frequency vector $f(t)$, we express the frequency of each state in per $100,000$ population to facilitate interpretation as well as comparison to estimated infection rates across geographies. For the two conditioning variables, the testing positivity rate x_t is measured by the weekly moving average of the testing positivity rate (i.e., out of all tests) and the test intensity y_t is measured by the rolling 7-day average of tests per day per $100,000$ population as of date t. Online Appendix Figure A.1 show the daily evolution of $ID(t)$, $H(t)$ and $D(t)$ for the US and Online Appendix Figures A.3 - A.5 show the same data for the nine states.

2.2 Estimation of model parameters

The initial frequency is set equal to $100,000$ for state $S(0)$ and 0 for all other states. The model parameters θ and the series of frequencies of the unobserved state $IU(t)$ are then estimated by solving the optimization problem in Equation (1) numerically using the $fminsearch$ function in Matlab. The mean fitted values (%RMSE) are within 2.36% of observed values for the US and the %RMSE for the individual states are shown in Online Appendix Table A.1. The comparisons of fitted and observed frequencies for state ID, state H and state D of the US are shown in Online Appendix Figure A.2. We note that the estimated frequencies track the observations closely.

2.3 Estimation of cumulative total cases

We estimate the cumulative total cases based on the data of detected cases and our estimated undetected cases. The cumulative total cases $CI(T)$ up to date T is $CI(T) = CD(T) + CU(T)$, where $CD(T)$ is the cumulative detected cases and $CU(T)$ is the cumulative undetected cases. We use the cumulative number of positive tests provided in the data set as the measure of cumulative detected cases. Using our estimated model, the cumulative undetected cases is computed as

$$CU(T) = \sum_{t=1}^{T} p_{i,t}(1 - p_{d,t})S(t-1) - \sum_{t=1}^{T}(1 - p_{21} - p_{24})p_{d,t}IU(t-1).$$

The first summation in the above equation is the cumulative number of daily new entrants to state IU, which measures the total number of patients who have been through the infected and undetected state. According to our model specification, a proportion of patients in state IU transit to state ID at each t as they became "detected". These patients were included in the cumulative number of detected cases, therefore, we subtract this portion of patients from the total number of patients who was in state IU to get the cumulative number of undetected cases.

3 Results

We estimate the time-varying Markov model on COVID-19 propagation data of the US and nine individual states, Arizona, California, Florida, Georgia, North Carolina, New Jersey, New York, Pennsylvania and Texas, of which the total residential population account for nearly half of the US population. We examine two versions of the model, a basic model with static transition probabilities to state H and a full model with time-varying transition probabilities driven by lagged frequency of the corre-

Panel (a)	**Parameter Estimates of the US**				
	a_1	a_2	a_3	b_1	b_2
	-8.4486	-0.0030	25.7573	-5.0047	0.0120
	$1 = S$	$2 = IU$	$3 = ID$	$4 = H$	$5 = D$
$2 = IU$	0.2979	Time-varying	Time-varying	0.0013	0
$3 = ID$	0.0461	0	0.9482	0.0057	0
$4 = H$	0.0805	0	0	0.8970	0.0225
$5 = D$	0	0	0	0	1
Panel (b)	**Average Parameters of Individual States**				
	a_1	a_2	a_3	b_1	b_2
	-8.134	-0.004	31.353	-4.531	0.012
	$1 = S$	$2 = IU$	$3 = ID$	$4 = H$	$5 = D$
$2 = IU$	0.403	Time-varying	Time-varying	0.002	0
$3 = ID$	0.077	0	0.917	0.006	0
$4 = H$	0.077	0	0	0.8970	0.026
$5 = D$	0	0	0	0	1

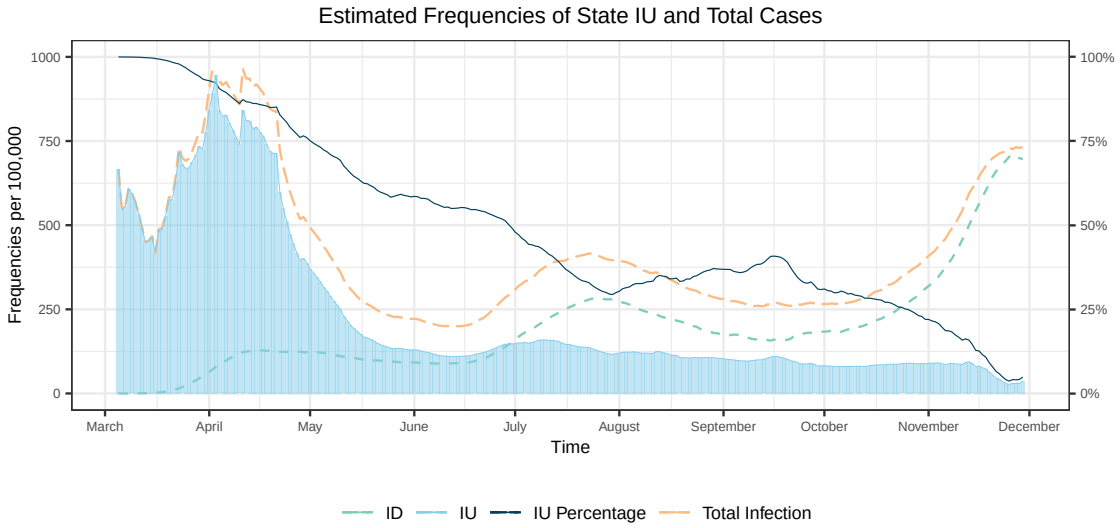
Table 1: Parameter Estimates of the Basic Model

sponding state. The key results (e.g., estimates of infected and undetected cases) are very similar, so we only report the results from the basic model and the results from the other specification are available on request. The parameter estimates for the US are reported in Table 1a and the average of state parameters are in Table 1b. The model parameters for individual states are given in Online Appendix Table A.1.

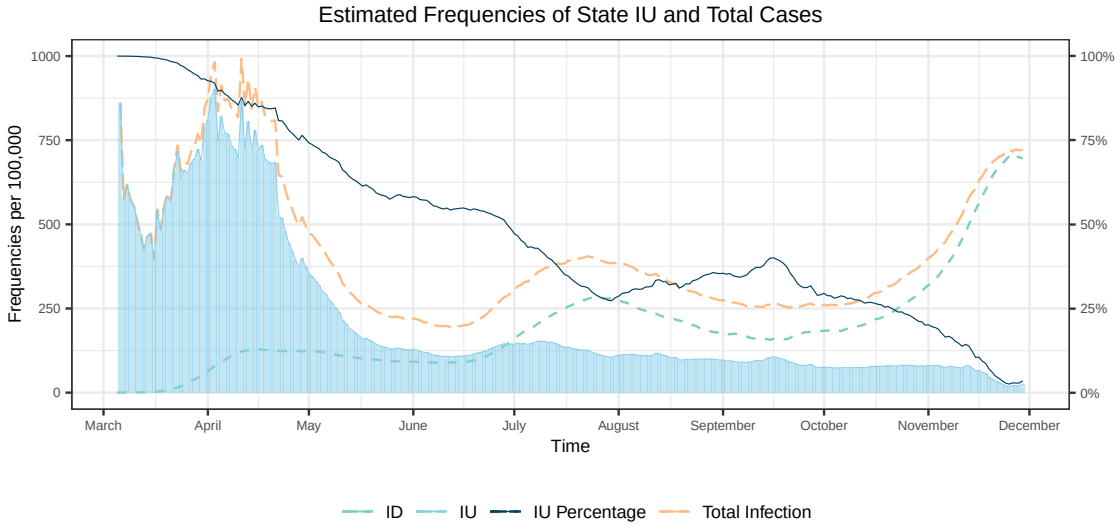
3.1 Estimates of infected and undetected cases

The time series of the frequencies of state $IU(t)$ are the quantities of primary interest. Figure 2a shows the estimated frequencies of the state IU (blue bars) and the observed frequencies of the state ID (dashed green line) for the US. We find that the estimated undetected cases grew rapidly since mid-March until peaking in early-April. After that, estimated undetected cases has declined substantially but it was not un-

til the beginning of July that the number of detected cases exceeded the number of undetected cases.



(a) **US Parameters**



(b) **Average State Parameters**

Figure 2: **The Estimated Frequencies of State IU and Total Infections.** The figure shows the estimated frequencies of state IU and total infections. The dashed green line is the observed frequencies of state ID and the blue bars are the estimated frequencies of state IU. The dashed orange is the estimated frequencies of total cases (e.g. total frequencies of state IU and state ID). The solid navy line with corresponding values on the right y-axis shows the evolution of unidentified percentage of the total cases.

We calculate the total infections per 100,000 population at time t (dashed orange line) by the summation of $ID(t)$ and $IU(t)$. We observe that the total number of infections are driven by the undetected cases before June, while the confirmed cases dominate the trend in total infections thereafter.[4] The evolution of the unidentified percentage of the total cases are shown by the navy line of Figure 2a with corresponding values on the right y-axis and we observe that this percentage is trending down over time.

We estimate undetected cases and total infections for each individual state. The results are shown in Online Appendix Figures A.6a - A.14a. Based on our estimation, all states in our model experienced a rapid increase in the total infections during April, while five out of nine states in our data only show a mild spread of SARS-CoV-19 according to the confirmed cases. We will discuss later that this is in fact a result of insufficient testing. Around three weeks after most states issued Stay-at-Home orders, our estimation suggests that both the undetected cases and the total infections peaked around the middle of April and started to decline in all nine states. As each state began to lift restrictions around early June, the estimated total infections started to climb again except for New York and New Jersey. In contrast to the early outbreaks when the majority of total infections were undetected cases, the total infections during this period were driven by the detected cases in all states. In the most recent surge beginning at late October, the percentage of undetected cases, although varies across states, were below 20% for all states, suggesting that detected cases are becoming representative enough for total cases.

[4]This result is also consistent with the converging pattern between the "death-implied" new cases and the reported new cases shown in Figure 1.

3.2 Estimates of cumulative cases

We compute the cumulative total cases based on the data of detected cases and our estimated undetected cases. The estimated cumulative infections for the US and individual states as of November 29, 2020 are shown in Figure 3 and Table 2. Based on our estimation, the number of cumulative cases in the US is $74,667,047$, which is more than 5 times the reported confirmed cases and accounts for 22.75% of the US population. Our estimated percentage of undetected infections out of cumulative total infections is 82.34% for the US comparing to the estimates of 79.21% and 74.23% in the works by Gu (2020) and Friedman et al. (2020) respectively. For individual states, both the percentage of population infected and the percentage of undetected cases out of total infections vary a lot across states. New York and New Jersey, which are the early epicenters of the Covid-19 pandemic, have 56.84% and 39.81% of population being infected respectively. In contrast, Georgia and North Carolina have 21.79% and 20.10% of population being infected respectively. New York and New Jersey also have the highest percentage of undetected infections (94.20% and 90.55% respectively), while Georgia and North Carolina have the lowest percentage of undetected infections (81.82% and 82.86%). This result is consistent with the fact that New York and New Jersey experienced worse early outbreaks than other states during which the test capacity was limited.

3.3 Conditioning variables

As conditioning variables in our analysis we include both the testing positivity rate and the intensity of testing. A high testing positivity rate is a sign of higher virus transmission accompanied with insufficient testing, which points to higher level of total infections even with a relatively low number of detected cases. The intensity

Panel (a)	**Number of Infections**			
State	Confirmed	Undetected	Total	% Undetected Infections
AZ	325,995	1,606,040	1,932,035	83.13%
CA	1,198,934	7,674,817	8,873,751	86.49%
FL	976,944	5,534,536	6,511,480	84.99%
GA	420,601	1,893,227	2,313,828	81.82%
NC	361,778	1,748,355	2,110,133	82.86%
NJ	334,114	3,200,884	3,534,998	90.55%
NY	641,161	10,414,697	11,055,858	94.20%
PA	357,196	2,747,136	3,104,332	88.49%
TX	1,157,273	7,018,290	8,175,563	85.84%
State Sum	5,773,996	41,837,985	47,611,981	87.87%
US	13,188,777	61,488,270	74,677,047	82.34%

Panel (b)	**Infections per 100,000 Population**			
State	Confirmed	Undetected	Total	Population
AZ	4,477	22,061	26,538	7,280,000
CA	3,034	19,425	22,459	39,510,000
FL	4,548	25,766	30,314	21,480,000
GA	3,960	17,827	21,787	10,620,000
NC	3,445	16,651	20,096	10,500,000
NJ	3,762	36,046	39,808	8,880,000
NY	3,296	53,546	56,842	19,450,000
PA	2,790	21,462	24,252	12,800,000
TX	3,990	24,201	28,191	29,000,000
State Sum	3,619	26,227	29,847	159,520,000
US	4,018	18,735	22,753	328,200,000

Table 2: Estimated Cumulative Infections as of Nov 29, 2020 (state parameters)

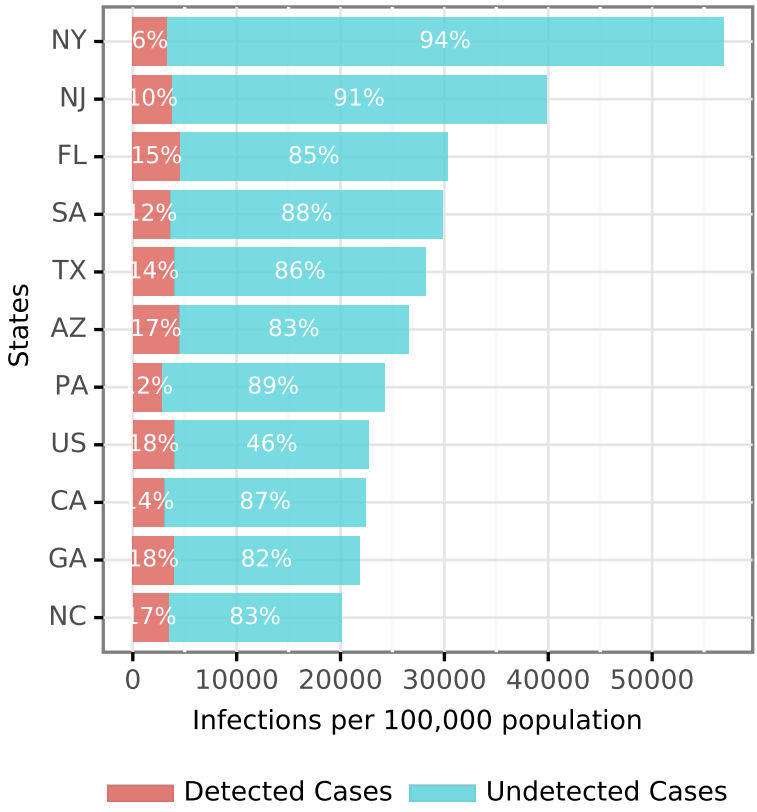


Figure 3: **The estimated cumulative infections (per 100,000 population) for the US and individual states as of November 29, 2020.**
The figure shows the estimated cumulative infections (per 100,000 population) for the US and nine individual states (SA represents the aggregate of the nine states) as of November 29, 2020. The red bar shows the proportion of detected infections and the blue bar shows the proportion of undetected infections.

Figure 4: **The Conditioning Variables (x_t and y_t).**
The top panel shows the time series data of the test positivity rates x_t, which is measured by the weekly moving average of the rates of positivity in testing (i.e., out of all tests). The bottom panel shows the time series data of test intensity y_t, which is measured by the rolling 7-day average of tests per day per $100,000$ population as of date t.

of testing measures tests conducted daily per $100,000$ population. As testing plays a key role in identifying infected individuals, we believe that the intensity of testing is an important factor in determining the proportion of total infections being detected. The two conditioning variables for the US and individual states are shown in Figure 4.

In the transition matrix of the Markov process, we model the time-varying probability of a susceptible individual being infected in the next period as a function of testing positivity rate and the probability of being detected conditioning on being

infected as a function of testing intensity. Our estimates of the coefficient of testing positivity are 25.75 for the US and ranging from 15.30 to 39.45 for the nine individual states with an average of 31.35. The positive coefficients imply that a higher testing positivity rate is associated with a higher probability of a susceptible individual being infected holding other factors constant. We get positive estimates of coefficients for intensity of testing, which is 0.012 for the US and ranging from 0.006 to 0.018 for the nine individual states with an average equal to the US coefficient. The positive estimates is in line with our intuition that the detection rate is increasing with the intensity of testing. From Figure 2a, our results suggest that the substantial increase in testing capacity over time has been successful in identifying a much higher percentage of infections in the US. Cross-sectionally, we observe in Figure 5 that states with high testing positivity rate experience worse outbreaks with higher total infections. From the three snapshots, we find that the testing intensity is increasing in all states over time, so does the percentage of detected cases. In most of the cases, states with higher testing intensity have higher detection rate.

4 Discussion

We estimate a time-varying Markov model to infer the number of undetected cases and total infections from easily observable data on reported cases, hospitalizations and deaths at the state and national level. The results are intuitive and in line with other published estimates. Our model is capable of providing timely estimates of cumulative total cases as well as the evolution of total infections over time, which is a critical for assessing the burden of COVID-19 on healthcare system and informing public health decisions. Nonetheless, our model is fairly simple and can be applied easily to other regions.

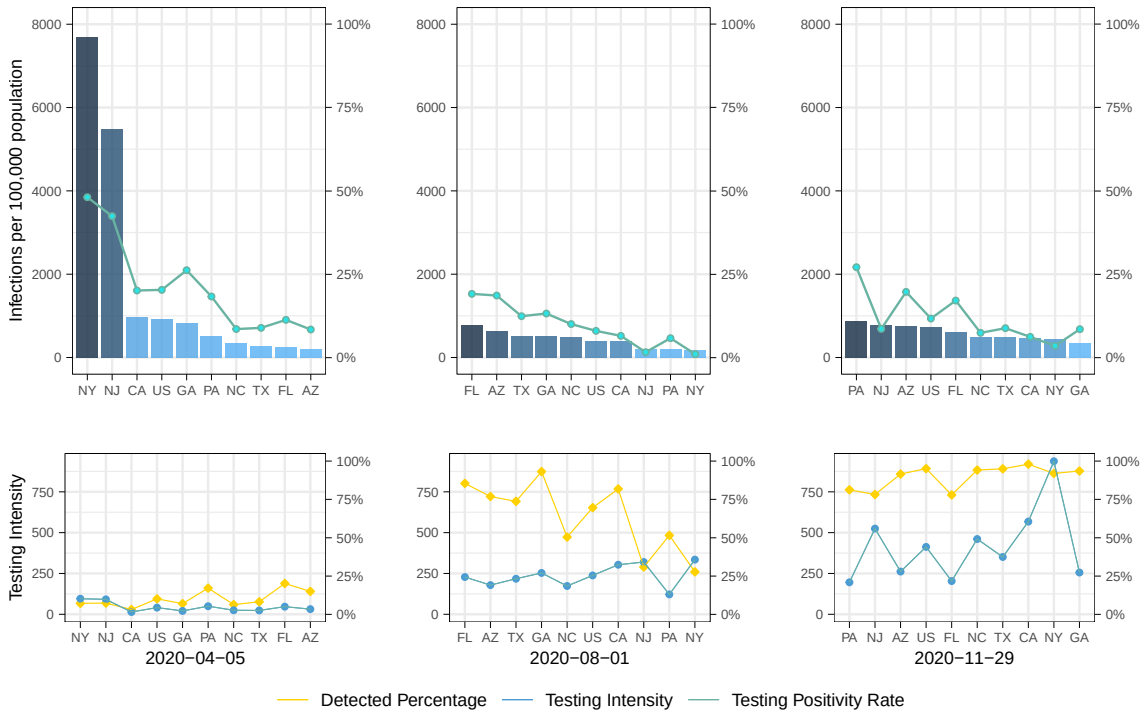


Figure 5: **Snapshots of total infections, detected rate and two conditioning variables.**
The figure shows three snapshots of our estimated total infections, detected rate with two conditioning variables on April 5 (left panel), August 1 (middle panel), and November 29 (right panel). The three dates are taken to reflect three peaks in COVID-19 cases. The blue bars show the total infections per $100,000$ population and the yellow line shows the percentage of detected cases out of total infections. The teal line shows the testing positivity rate and the blue line shows the testing intensity on the selected date. All percentage values (detected percentage and testing positivity rate) are corresponds to the right y-axis.

Our estimates indicate a high percentage of undetected cases early in our sample period followed by a decline to a much lower percentage of undetected cases by July, which is consistent with the converging trend between "death implied" new cases and reported new cases[5]. The substantial increase in testing capacity has been successful in identifying a much higher percentage of infections. Taken at face value, our results show that reported confirmed cases in the US increasingly reflect the true number of infections. The bad news from these results is that the recent surge in positive tests since October is in fact an increase in new cases as opposed to an increase related to higher number of tests.

The evolution of estimated total infections also demonstrates the effectiveness of the non-pharmaceutical interventions. Most states in the US issued Stay-at-Home orders in late March. After three weeks, we see a significant decrease of total infections in the US and all individual states, which suggests the strong impact of the interventions on containing the spread of SARS-CoV-19. This result is consistent with previous studies Korevaar et al. (2020); Unwin et al. (2020). However, the effectiveness of the interventions is not necessarily reflected by the reported confirmed cases as we still observe steady increase in detected cases in AZ, CA, NC and TX.

Estimation of the cumulative total infections may provide better information on population immunity and thus enable planners to better distribute vaccines, make more informed public health decisions, and ultimately optimize social and economic well-being of the country. The number of cumulative infections may suggest what percentage of the population is immune to the virus because of previous infection and a very low reinfection rate, or likewise how many people need to get a vaccine before we reach herd immunity. According to our estimation, by the end of November, about 23% of the US population has been infected by SARS-CoV-2, which suggests that

[5]See Figure 1 for details

we are still a long way from herd immunity. Yet, if vaccinations could be prioritized based on past infections, our estimated total infections represents a substantial base of the population which may already be immune.

The model parameters are easy to estimate and have intuitive explanations. In Table 1, $p_{21} = 0.2979$, which corresponds to a less than 1 week average recovery time of for state IU, and $p_{31} = 0.0461$, which represents an average recovery time around 20 days for state ID. The model estimates that it takes longer for a patient in the detected state to recover, which is reasonable considering it is more likely that patients with severe cases will get tested (and be detected) thus the overall health condition of state ID is worse than state IU. This finding is also consistent with the estimated transition probabilities to state H. The probability of transition to state H is 0.0057 from state ID, which is higher than the probability of 0.0013 from state IU. The estimate of p_{33} is 0.9482, which means that people stay in the state ID for an average around 18 days and are then either hospitalized or recover. This estimate is roughly consistent with how we construct the variable representing state ID (i.e. rolling 2-week sum of the positive tests). The mortality rate conditional on being hospitalized is 2.25%, which is higher than the estimated value of 0.68% for the overall infection-fatality rate of COVID-19 in the work by Meyerowitz-Katz and Merone (2020). This result is not surprising considering that the severity of the illness is higher for the hospitalized patients than the average severity of all cases.

We estimate the model using data of the US and nine individual states. The states in our sample constitute a good representation of the overall demographic not only because their population account for nearly half of the US population, but also since they have experienced the pandemic in very different ways since March in terms of the trends of detected cases, hospitalizations, and deaths. Despite the similar recent surge in detected cases and hospitalizations, these data peaked in late-April

for New York, New Jersey and Pennsylvania, while other states have their peaks in mid-July. We estimate the undetected cases and total infections in the US using the average state parameters. The estimates of cumulative total infections are given in Online Appendix Table A.2, which is similar to the estimates using the US parameters shown in Table 2. Figure 2b shows the estimated frequencies of the state IU and total infections using the average state parameters. We observe that the estimates follow similar trend as the results in Figure 2a using the US parameters.

One concern about our analysis is that we are not able to condition on the age of those with detected cases or who are hospitalized due to data limitation. There are published studies (see O'Driscoll et al. (2020)) that show the estimated infection fatality rate increasing progressively with age. Given anecdotal evidence that age of detected cases is changing through time, the estimation is likely to benefit by conditioning estimates on other variables such as the average age of hospitalized patients or the average age of those testing positive. Our estimates could also underestimate total infections if the quality of care has improved over time and reduced hospitalization and death rates in a way the model does not capture.

References

Friedman, J., P. Liu, C. E. Troeger, A. Carter, R. C. Reiner, R. M. Barber, J. Collins, S. S. Lim, D. M. Pigott, T. Vos, S. I. Hay, C. J. Murray, and E. Gakidou (2020). Predictive performance of international COVID-19 mortality forecasting models. medRxiv 2020.07.13.20151233.

Gouriéroux, C. and J. Jasiak (2020). Time Varying Markov Process with Partially Observed Aggregate Data: An Application to Coronavirus. *Journal of Econometrics*, (forthcoming).

Gu, Y. (2020). Estimating True Infections Revisited: A Simple Nowcasting Model to Estimate Prevalent Cases in the US. https://covid19-projections.com/estimating-true-infections-revisited/. Accessed: 2020-11-30.

Korevaar, H. M., A. D. Becker, I. F. Miller, B. T. Grenfell, C. J. E. Metcalf, and M. J. Mina (2020). Quantifying the impact of us state non-pharmaceutical interventions on covid-19 transmission. *medRxiv*.

Meyerowitz-Katz, G. and L. Merone (2020). A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. medRxiv 2020.05.03.20089854.

O'Driscoll, M., G. Ribeiro Dos Santos, L. Wang, D. A. T. Cummings, A. S. Azman, J. Paireau, A. Fontanet, S. Cauchemez, and H. Salje (2020, 11). Age-secific mortality and immunity patterns of SARS-CoV-2. *Nature*.

Randolph, H. E. and L. B. Barreiro (2020). Herd immunity: Understanding COVID-19. *Immunity 52*(5), 737 – 741.

Unwin, H. J. T., S. Mishra, V. C. Bradley, A. Gandy, T. A. Mellan, H. Coupland, J. Ish-Horowicz, M. A. C. Vollmer, C. Whittaker, S. L. Filippi, X. Xi, M. Monod, O. Ratmann, M. Hutchinson, F. Valka, H. Zhu, I. Hawryluk, P. Milton, K. E. C. Ainslie, M. Baguelin, A. Boonyasiri, N. F. Brazeau, L. Cattarino, Z. Cucunuba, G. Cuomo-Dannenburg, I. Dorigatti, O. D. Eales, J. W. Eaton, S. L. van Elsland, R. G. FitzJohn, K. A. M. Gaythorpe, W. Green, W. Hinsley, B. Jeffrey, E. Knock, D. J. Laydon, J. Lees, G. Nedjati-Gilani, P. Nouvellet, L. Okell, K. V. Parag, I. Siveroni, H. A. Thompson, P. Walker, C. E. Walters, O. J. Watson, L. K. Whittles, A. C. Ghani, N. M. Ferguson, S. Riley, C. A. Donnelly, S. Bhatt, and S. Flaxman (2020, 12). State-level tracking of COVID-19 in the United State. *Nature Communications 11*, 6189.

Wu, S., A. Mertens, Y. Crider, A. Nguyen, N. Pokpongkiat, S. Djajadi, A. Seth, M. Hsiang, J. Colford, A. Reingold, B. Arnold, A. Hubbard, and J. Benjamin-Chung (2020, 09). Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications 11*, 4507.