What is the Active Prevalence of COVID-19?

Mu-Jeung Yang[a], Nathan Seegert[a], Maclean Gaulin[b], Adam Looney[a], Brian Orleans[c], Andrew T. Pavia[d], Kristina Stratford[e], Matthew Samore[e], Steven Alder[f]

[a]*Department of Finance, David Eccles School of Business, University of Utah*
[b]*Department of Accounting, David Eccles School of Business, University of Utah*
[c]*CCTS Study Design & Biostatistics Center, School of Medicine, University of Utah*
[d]*Division of Pediatric Infectious Diseases, School of Medicine, University of Utah*
[e]*Division of Epidemiology, School of Medicine, University of Utah*
[f]*Division of Public Heath and Division of Epidemiology, School of Medicine, University of Utah*

Abstract

We provide a method to track active prevalence of COVID-19 in real time, correcting for time-varying sample selection in symptom-based testing data and incomplete tracking of recovered cases and fatalities. Our method only requires publicly available data on positive testing rates in combination with one parameter, which we estimate based on a representative randomized sample of nearly 10,000 individuals tested in Utah. The method correctly predicts prevalence in two state-wide, representative randomized testing studies. Applying our method to all 50 states we show that true prevalence is 2–3 times higher than publicly reported.

Keywords: Coronavirus, Testing, Random Sampling, Real-Time Prevalence Tracking
JEL MSC: I15, I18, J68

Email address: Corresponding author: mjyang@eccles.utah.edu. (Mu-Jeung Yang)
This report describes results from a health surveillance project initiated in cooperation with the State of Utah. The authors would like to acknowledge the support from the Governor's Office of Management and Budget (GOMB). The paper also benefited from useful discussion with several biostatisticians connected to the HERO project (https://eccles.utah.edu/utah-health-economic-recovery-outreach/), including Drs. Yue Zhang, Tom Greene, Angela Presson, and Jincheng Shen.

1. Introduction

How prevalent is COVID-19? Accurate measures of the fraction of the population that is currently infected and infectious are crucial for (1) policymakers setting public health and economic policy, (2) private citizens evaluating the risk of getting infected, and (3) researchers trying to understand and predict COVID-19 dynamics. However, estimating prevalence is empirically challenging because the limited testing that is available is typically reserved for individuals meeting certain criteria, like having symptoms or a known exposure. This limitation induces sampling bias, and as a result, causes publicly reported case-counts, like those collected by state governments or the CDC, to underestimate the true number of cases (Stock, 2020; Burger and McLaren, 2017). The gold-standard solution to sample selection is randomized testing. However, a random sampling study can quickly become prohibitively costly and organizationally unwieldy to provide accurate, real-time information as disease dynamics change.[2]

A key mechanism through which COVID-19 impacts economic activity is through voluntary social distancing. Therefore, an analysis that allows for time-varying infection rates is crucial to understand the economic consequences of COVID-19 for changes in demand for local services (Chetty et al., 2020) and changes in unemployment rates (Yang et al., 2020). Understanding current active prevalence is also important for policymakers who face a trade-off between lives and economic livelihoods when considering lockdown policies or other non-pharmaceutical public health measures. However, a problem they face is that publicly reported prevalence provides a biased estimate on the virus's status.

This study attempts to solve these problems by providing an easily applicable model calibrated on randomized testing data. We develop a method for estimating prevalence in local areas based on real-time public data by applying Bayes' Law to a standard SIR (Susceptible, Infected, and Removed) epidemiological model informed by data from a representative randomized testing project with roughly 10,000 participants in Utah.

[2]See the significant changes in prevalence dynamics between late May and early July 2020.

Our method measures latent prevalence in real-time and requires only one parameter—which we estimate from our random testing data—and one publicly available time series: the positive rate of testing.[3] This approach builds on Stock (2020), who shows that under conditional independence of infections and testing, given symptoms, one can use Bayes' Law to utilize positive testing rates for model estimation. We extend his result by showing that one can use this setup to estimate prevalence, even without estimating an SIR model. Additionally, we outline the conditions on the testing regime under which our method is applicable, which allows researchers to apply it to future outbreaks.

We provide estimates for the parameter required to apply this method—the likelihood ratio of symptoms for infected relative to uninfected persons—based on response data from roughly 10,000 randomly selected individuals in Utah between May and July 2020. This parameter allows us to re-weight the symptom-based testing data and extrapolate from positive rates in symptomatic people to the positive rate of the underlying (symptomatic and asymptomatic) population.

We pursue two approaches to validate our method. First, utilizing our randomized testing and health survey data, we directly test the key conditional independence assumption our method requires for valid point estimates of prevalence. We show that we cannot reject the null that the assumption holds in our micro-data (p-value of 0.943). Second, we benchmark our method's prevalence estimate against the true rate derived from randomized testing. Our own randomized viral testing was conducted in Utah between May 4^{th} and July 1^{st}, 2020 and estimates that the prevalence of COVID-19 in Utah was 0.27% with a 95% confidence interval of [0.12%, 0.42%]. At the same time, our method predicts a median viral prevalence of 0.3% during the same period, which is well within the 95% confidence interval of the randomized testing point estimate. To test generalizability of our method beyond Utah, we compare our latent prevalence estimate to prevalence from randomized testing in Indiana between April 25 and April 29, 2020, by Menachemi et al. (2020). They estimate a viral prevalence of 1.7% (95% confidence interval from 1.1% to

[3] Our measurement is related in spirit to sufficient welfare statistics as in Chetty (2009) and Arkolakis et al. (2012), but differs in its focus (i.e., latent disease prevalence).

2.54%), compared to a predicted prevalence of 1.55% from our method.

Given our method's accuracy, we provide estimates for all 50 US states, which imply that prevalence is $2-3$ times higher on average than publicly reported prevalence. Additionally, by comparing the time series of latent and reported prevalence we show that the ratio of these two varies over time. This variation could be driven by the intensity of testing or the selection into testing, both of which affect reported cases but not our estimate of latent prevalence (which is an important insight for modeling purposes, e.g. Yang et al. (2020)).

We add to the fast-growing economics literature analyzing COVID-19 disease dynamics and its implications for health and economic outcomes. First, a number of papers have developed different approaches to estimate SIR type models, such as Atkeson et al. (2020), Korolev (2020), Fernandez-Villaverde and Jones (2020) and Yang et al. (2020). We show below that our method allows researchers to directly measure the time path of COVID-19 prevalence, using publicly available positive testing rates. Second, our work is related to studies of the importance of voluntary social distancing, whether based on rational expectations as in Eichenbaum et al. (2020a), Eichenbaum et al. (2020b), Farboodi et al. (2020), or based on information and learning as in Brzezinski et al. (2020), Allcott et al. (2020), Bursztyn et al. (2020), Simonov et al. (2020), Yang et al. (2020). Our work complements these efforts by providing correct prevalence measures even for models with time-varying infection rates or generalized matching functions. A third strand of the literature utilizes tools from partial identification to provide bounds on prevalence; for example, Aspelund et al. (2020), Manski and Molinari (2020). We add to this work by providing time-varying point estimates of latent prevalence and formally test the necessary conditional independence assumption for field data from Utah.

2. Model

We build on the discussion of inference in Stock (2020) within a canonical SIR model by Kermack and McKendrick (1927), which is the basis of much epidemiological and economic research on COVID-19. The total population N can be compartmentalized into susceptible S_t,

infected (and infectious) I_t and removed, i.e., recovered or deceased R_t. The change in these compartments is determined by the infection rates β_t, the matching function $G_t(.)$, and the inverse of the length of infectiousness γ_t. The discrete time version of a generalized SIR model is given by

$$\begin{aligned} S_t + I_t + R_t &= N && (1) \\ \Delta S_{t+1} &= -\beta_t \cdot G_t(S_t, I_t) && (2) \\ \Delta I_{t+1} &= \beta_t \cdot G_t(S_t, I_t) - \gamma_t \cdot I_t && (3) \\ \Delta R_{t+1} &= \gamma_t \cdot I_t. && (4) \end{aligned}$$

The infection rate β_t allows for arbitrary time variation, such as voluntary social distancing (Farboodi et al., 2020; Fernandez-Villaverde and Jones, 2020; Yang et al., 2020), and the latency period until the virus can be detected through tests, to change the flow of people from susceptible to infected. Furthermore, $G_t(.)$ allows for a general, time-varying, and non-linear matching function that accounts for features such as super-spreading events.[4] Finally, variation in the length of infectiousness $1/\gamma_t$, captures not just clinical disease progression but also factors such as time until an infectious person is identified through testing and quarantined, see Berger et al. (2020) and Yang et al. (2020). It should be noted that while our measurement approach is consistent with this general SIR model in (1)-(4), we only will use a version of equation (1) for our method. For this purpose, we divide (1) by N to transform numbers of people into probabilities of being in different compartments, e.g.,

$$P(S_t) + P(I_t) + P(R_t) = 1, \quad (5)$$

with $P(X_t) = \frac{X_t}{N}$ for $X_t = \{S_t, I_t, R_t\}$. We are interested in measuring the variable $P(I_t)$, which is a stock influenced by inflows (new infections) and outflows (new fatalities and newly recovered cases) as shown in equation (3).

[4] Standard random matching implies that $G_t(I_t, S_t) = \frac{S_t \cdot I_t}{N}$.

As Stock (2020) notes, the barrier to identification of the infection rate is that testing and case-count data is based on symptoms (or other specific risks). Given a person is symptomatic and gets tested, the positive testing rate mostly reflects the probability of testing positive. The positive rate is effectively the only source of usable information, as the number of symptomatic and asymptomatic people is not reported in the data.

We show that if testing is randomly assigned conditional on the presence of symptoms, then we can express prevalence in terms of the positive test rate. Denote I_t as the event of being infected at time t,[5] T_t as the event of being tested at t and σ_t as the event of being symptomatic at t. Then, the conditional independence assumption is given by

$$P(I_t, T_t|\sigma_t) = P(I_t|\sigma_t) \cdot P(T_t|\sigma_t) \tag{6}$$

Although this condition is unlikely to hold exactly in the data, deviations from conditional independence condition (6) are likely to be small in magnitude for most US states. COVID-19 testing is still mostly symptom-based (based on guidance from health professionals and public health officials) and testing of asymptomatic individuals remains rare, even for those with known exposures (who are often encouraged to quarantine instead). Our randomized testing data together with our health survey will also allow us to directly test this assumption for a sample period of about 2 months in Utah. In particular, we use the survey data to calculate $P(I_t, T_t|\sigma_t)$, $P(I_t|\sigma_t)$, and $P(T_t|\sigma_t)$.[6] We report these estimates in panel A of table 2. Importantly, we fail to reject the null that $P(I_t, T_t|\sigma_t) = P(I_t|\sigma_t) \cdot P(T_t|\sigma_t)$. To make sure that the rejection of this null hypothesis is not driven by an underpowered test, we also provide confidence intervals for the odds ratio $\frac{P(I_t, T_t|\sigma_t)}{P(I_t|\sigma_t) \cdot P(T_t|\sigma_t)}$, which are reasonably tight. Additionally, we also show below that our measurement approach is still valuable if conditional independence assumption (6) fails to hold.

[5] This is a slight abuse of notation.

[6] Recall that "t" in the conditional independence test denotes a two month period.

Condition (6) implies

$$P(I_t|T_t,\sigma_t) = P(I_t|\sigma_t)\cdot P(T_t|\sigma_t)\frac{1}{P(T_t|\sigma_t)} = P(I_t|\sigma_t), \tag{7}$$

Equation (7) shows that positive testing rates $P(I_t|T_t,\sigma_t)$ will correctly measure the probability of having the virus, conditional on symptoms, given that testing is random among symptomatic people. Furthermore, the right hand side of (7) can be expressed using Bayes' Law:

$$P(I_t|\sigma_t) = \frac{(1-\alpha)\cdot P(I_t)}{P(\sigma_t)}, \tag{8}$$

where $\alpha = P(\neg\sigma_t|I_t)$ is the probability of being asymptomatic, given that a person is infected. Furthermore, the fraction of symptomatic people is

$$P(\sigma_t) = (1-\alpha)\cdot P(I_t) + s_0\cdot(P(S_t)+P(R_t)), \tag{9}$$

where $s_0 = P(\sigma_t|\neg I_t)$ is the rate of symptoms for not infected persons. Two sets of assumptions facilitate our analysis. First, following Stock (2020), we assume that people in R_t exhibit the same rate of symptoms as people in S_t. Second, we assume that the likelihood ratio $\frac{P(\sigma_t|I_t)}{P(\sigma_t|\neg I_t)} = \frac{1-\alpha}{s_0}$ is a constant across time and locations. This is consistent with a model in which patients select into testing only if their symptoms suggest a subjective likelihood ratio that is greater than $\frac{1-\alpha}{s_0}$. Hence even if symptom rates for non-infected people vary across locations or time (e.g. with the onset of flu season), we assume that doctors adjust their subjective probabilities accordingly, and only recommend taking a test if the individual likelihood ratio is greater than $\frac{1-\alpha}{s_0}$. This is an important assumption as it allows us to generalize the likelihood ratio across time and location. We will test its generalizability in the results section.

Using (5) in (9) we obtain

$$P(\sigma_t) = s_0 + (1-\alpha-s_0)\cdot P(I_t). \tag{10}$$

Then we can use (10) in (8) and solve for $P(I_t)$ to get

$$P(I_t) = \frac{P(I_t|\sigma_t)}{\frac{1-\alpha}{s_0}(1-P(I_t|\sigma_t)) + P(I_t|\sigma_t)}. \tag{11}$$

Equation (11) is our central measurement tool, henceforth referred to as "hidden-infection-method." It states that given an estimate of the likelihood ratio $\frac{1-\alpha}{s_0}$ and a time series of the fraction of tests that are positive $P(I_t|T_t,\sigma_t)$, which proxies for $P(I_t|\sigma_t)$, one can measure the time path of the fraction of the population that is currently infectious $P(I_t)$. Furthermore, the validity of the point estimate of (11) depends on conditional independence (6) as well as the absence of widespread asymptomatic testing, which we believe is a good approximation for many states.

We add four observations on the robustness of the hidden-infection-method (11). First, note that it is not necessary to estimate any of the hidden objects $\beta_t, \gamma_t, G_t(\cdot)$ or the latent number of initially infected I_0. Second, note that misreporting of symptoms only matters to the degree that it impacts the likelihood ratio $\frac{1-\alpha}{s_0}$. If infected and uninfected people are equally likely to over-report symptoms the ratio $\frac{1-\alpha}{s_0}$ stays unaffected. Third, even if conditional independence (6), fails, equation (11) can provide useful bounds on the variable $P(I_t)$, see Aspelund et al. (2020) and Manski and Molinari (2020) for similar results. Specifically, suppose the following inequalities holds

$$\begin{aligned} P(I_t|\sigma_t) &\geq P(I_t|T_t,\sigma_t) \\ P(I_t|\sigma_t) &\geq P(I_t|T_t,\neg\sigma_t) \end{aligned} \tag{12}$$

The first inequality in equation (12) reflects selection into testing and is influenced by two sample selection effects. On the one hand, some people with information beyond symptoms, such as exposure to COVID-19 positive persons might be more likely to get tested, biasing $P(I_t|T_t,\sigma_t)$ up relative to $P(I_t|\sigma_t)$. On the other hand, very health conscious people might get tested even though they aggressively socially distance and only exhibit mild flu-like symptoms, biasing $P(I_t|T_t,\sigma_t)$ down relative to $P(I_t|\sigma_t)$. The first condition in (12) states that the first effect does not dominate the second effect.

In contrast, the second inequality states that positive rates for tested asymptomatic people are lower than the infection probability for symptomatic people in the population. This assumption will hold, if for example most asymptomatic testing is precautionary testing, for example in preparation of major surgery. It will even hold for moderate amounts of contact-tracing but will fail if all tested asymptomatic people had know exposures to COVID-19 and are therefore highly likely to be infected. If (12) holds, then

$$P(I_t|T_t) = P(I_t|T_t,\sigma_t)\cdot P(\sigma_t|T_t) + P(I_t|T_t,\neg\sigma_t)\cdot P(\neg\sigma_t|T_t) \leq P(I_t|\sigma_t) \tag{13}$$

since $P(\sigma_t|T_t) + P(\neg\sigma_t|T_t) = 1$. In other words, reported positive rates are lower than infection rates of symptomatic people in the population. In this case, it can be shown that

$$P(I_t) \geq \frac{P(I_t|T_t)}{\frac{1-\alpha}{s_0}(1-P(I_t|T_t)) + P(I_t|T_t)}. \tag{14}$$

In other words, under condition (12), use of reported positive rates $P(I_t|T_t)$, will still enable the hidden-infection-method to provide a valid lower bound on current COVID-19 prevalence in the population. We emphasize that the bound (14) is valid even without our conditional independence assumption (6) and without the assumption of no asymptomatic testing.

3. Data

We combine publicly available data with data from the Health and Economic Recovery Outreach (HERO) project, a large COVID surveillance program conducted in Utah (Samore et al., 2020). The public data we use reports the fraction of positives in all tests by state from the COVID tracking project.[7] This data contains the daily rates of positive tests in all state-wide COVID-19 tests to measure $P(I_t|T_t,\sigma_t)$. Since this daily data is noisy, we use a 7-day moving average.

[7]Data accessed from covidtracking.com/api

3.1. Field Experiment

The HERO project was initiated to estimate COVID-19 prevalence in Utah, understand indicators and risk factors for COVID, and improve decision-making. In the current application, randomized testing provides estimates of the key parameters for the hidden-infection-method, the baseline symptom rate s_0 and the asymptomatic rate α.[8] The associated survey instrument, which collected information about symptoms, allows us to estimate the latent probability that infected individuals are asymptomatic and the probability an uninfected individual has symptoms. Randomized testing also provides an estimate of viral prevalence, which provides an ideal benchmark for the estimate from our hidden-infection-method.

Between May 4^{th} and July 1^{st}, we contacted 25,438 households in central Utah (Davis, Salt Lake, Summit and Utah Counties). To recruit a representative sample, we randomly selected households from a public list of 657,870 addresses (provided by Utah municipalities) using a stratified sampling approach. Each address was encouraged to fill out a household survey, have each household member fill out an individual survey, and all members over the age of 12 were encouraged to get a PCR (viral) and serology (antibody) test. Individuals were compensated with a $10 USD gift card for completing the survey and being subsequently tested. Households in our first recruitment strategy ("in-person" recruitment) received a postcard, a letter, and a field team visited their address three times. The remaining addresses ("letter only") received a letter but were not contacted by our field team.

We lowered frictions for getting tested by parking a 'testing bus' in the center of the geographically compact area we selected. These areas consist of two or more adjacent Census tracts (which we refer to as tract-groups).[9] We stratified the tract-groups based on publicly reported case prevalence, the portion of the population identified as Hispanic, and the population's median age. Due to variance in county size across our sample, we used only subsets of these stratifying variables for the

[8] See Samore et al. (2020) for the estimates of the latent infectious rate estimates from the Utah HERO project.
[9] For reference, there are 212 Census tracts in Salt Lake County for which we defined 131 Census tract-groups. These tract-groups consist of roughly 4,000 addresses on average.

smaller counties, with only the largest, Salt Lake County, using all three. We then over-sampled some of the strata resulting in 26 tract-groups representing the 15 strata.[10]

From each selected tract-group, we sampled 30 Census blocks to be visited by field teams.[11] We chose seven primary addresses from each block and provided up to seven additional backup addresses if any of the first seven were vacant or not available after three attempts.[12] In total, 8,916 addresses ultimately received a visit from our field team and were included in the in-person sample.

Our second recruitment strategy (letter-only) received a letter with instructions on how to fill out an online survey and an invitation to get tested. This recruitment strategy allowed us to sample a wider geographic area. We selected addresses across all the tract-groups in each stratum in the same proportions as the in-person sample (omitting those tract-groups selected for in-person sampling). The primary sampling unit in this design is Census block-groups, and we selected at least 19 addresses within those block-groups.[13] In total, we sampled 13,997 addresses for letter-only contact. We supplemented this with 2,078 addresses which were uncontacted backup blocks in the in-person tract-groups, for a total of 16,076 letters.

Of the 8,916 addresses our field team approached, 2,975 responded by completing at least one survey, resulting in an average response rate of 33.4%. In the in-person sample, 1,752 (19.7%) visited the testing bus and completed a PCR test, and 2,154 (24.2%) completed a serology test. The sample of letters-only households yielded lower response rates, with only 2,091 (13.0%) households completing at least one survey and 1,851 (11.5%) being ultimately tested. On average 2.0 people per household from the in-person sample were tested and 1.8 people from the letter-only sample. In total, 8,221 people completed a viral test and 6,451 people completed a serology test.[14]

[10] We then used this stratification strategy to construct sampling weights to undo this deliberate sample selection and provide representative moments.

[11] In certain areas, we selected more than 30 blocks: in Park City, which is low density and defined as one tract-group, we selected 63 blocks, and, an area in Davis county we selected 45 blocks.

[12] These backup addresses received letters and are counted as "letter only" if the field team did not physically visit the addresses.

[13] The large majority of stratum had 19 addresses per block-group to achieve the target number of addresses in that stratum, but due to omitting the tract-group from the in-person sample entirely, this was not always possible. When it was not, we increased the number of addresses sampled from each block-group. A notable example of this is Park City, where we sampled up to 519 addresses from a single block-group.

[14] Test samples were collected by medical professions from the University of Utah Medical Center and analyzed by

3.2. *Health Survey*

We gathered extensive participant information through a health survey as part of the HERO project. The survey was completed by the field team during their visits to the address or completed online by the individuals in the household (for both in-person and letters-only samples, based on directions sent in the letters). The survey took approximately 15 minutes to complete by the field team for the first household member, and less time for any subsequent household members. The questions focused on the individual's current and past health, daily activities and social distancing behaviors, employment, and demographic questions.

This study uses a survey question regarding symptoms experienced by participants. Specifically, we asked, "Over the last 7 to 10 days, have you experienced any of the following symptoms? Select all that apply" with multiple-choice answers including "New loss of taste or smell" (hereafter referred to as *anosmia*).[15] From these, we calculate $(1-\alpha)$ by calculating the sampling-probability weighted average amount of positive answers for anosmia among those respondents who had positive COVID-19 serology tests. We then calculated s_0 as the sampling-probability weighted average number of positive answers for anosmia among those respondents who had negative COVID-19 serology tests. We use anosmia for our tabulated results because it is the most discriminating. Our method also works with different symptoms or different combinations of symptoms. For example, in columns (2) and (3) of Table 2 we consider someone as experiencing symptoms if they have anosmia and either 1 or 3 additional symptoms.

3.3. *Descriptive Evidence*

Roughly two-thirds of Utah's population resides in the four counties in our sample. Panel A of Table 1 provides descriptive statistics of these four counties from the US Census and CDC. Salt

ARUP Laboratories, the University's national reference laboratory. The sampling design, nonresponse corrections, bounding response rates, information treatments, and response rates by distance to the bus can be found in Mclaren et al. (2020) and Samore et al. (2020).

[15] Other options included, Fever, New or worsened cough, New or increased shortness of breath or difficulty breathing, Chills, Repeated shaking with chills, Muscle pain, Headache, Sore throat, and none of the above. See Mclaren et al. (2020) for more details on the survey.

Lake County, containing Salt Lake City, is the largest county in our sample, followed by Utah County to the south and Davis County to the north. Summit County, which contains the ski resort destination Park City, is the smallest county sampled but had the earliest cases of COVID in Utah and the highest reported case count in April when we began organizing this project.

We sampled the three larger counties roughly equally in proportion to their population; Summit County was over-sampled. Panel B of Table 1 reports the number of households we sampled and the households and individuals that participated in our sample by county. Our overall response rate is roughly 15 percent. The response rate in Summit County is lower because many addresses were vacant vacation homes and many households (in the letter-only sample) do not receive mail at their physical address. Panel C of Table 1 reports estimates of survey responses regarding characteristics, mobility, and Covid-19 concern, as well as viral and antibody prevalence for those that were ultimately tested. The median age of individuals in our sample is similar to the median age in the census data, albeit systematically older because we exclude individuals younger than 12. Our study also has an under-representation of individuals who self-identify as Hispanic relative to the census data, despite over-sampling Census tracts with an above-median proportion of Hispanic residents. Across the four counties, between 8 and 13 percent of participants indicated they had no concern of COVID, with the remainder having some concern or substantial concern. The estimates for viral and antibody prevalence are raw and do not include corrections for sampling design, nonresponse, or other population corrections (for estimates with these corrections see Samore et al. (2020)). Both viral and antibody prevalence are higher in Summit County than in the other three counties. This higher prevalence is likely due to being the county with the first known case. Importantly, the antibody prevalence rate is substantially smaller than previous estimates from other studies with sample selection issues that suggest anywhere between 2% and 30% positive rates.[16] We use these empirical estimates to provide external validity to the hidden-infection-method developed in this study.

[16] https://www.sciencemag.org/news/2020/04/antibody-surveys-suggesting-vast-undercount-coronavirus-infections-may-be-unreliable

4. Results

4.1. Prevalence of COVID in Utah

Panel B of Table 2 reports estimates for $(1-\alpha)$, s_0, and the likelihood ratio of these two variables. This table presents unweighted and sampling-probability weighted estimates. We focus on anosmia as the key symptom as it is the most informative symptom with a likelihood ratio of 16.35. Anosmia is more informative than the other symptoms because many people who do not have COVID-19 still experience other symptoms such as fevers and coughs, while people who do not have COVID-19 are much less likely to have anosmia (0.39%). This evidence is consistent with the medical evidence in Menni et al. (2020), who show that anosmia is a particularly strong predictor of COVID-19 infection in patients. Furthermore, Lampos et al. (2020) show that Google searches for anosmia predict the growth in confirmed COVID-19 cases.

We report estimates for anosmia, and anosmia in combination with other symptoms, such as fever, nausea, stuffy nose, etc. in the three columns of Table 2. The likelihood ratio for the combination of anosmia and other variables is smaller than for anosmia alone, as reported in columns (1)–(3), driven by the fact that the addition of these symptoms increases the noise in symptom data used to diagnose infection. Note that a smaller likelihood ratio would imply a larger latent prevalence rate, as shown in (11).

We emphasize that our usage of anosmia does not literally suggest that anosmia is the only symptom used to screen people into testing. Rather, the likelihood ratio measured using anosmia is unusually informative about how well a potential virus case can be predicted based on symptoms and what a plausible threshold of the likelihood ratio threshold for selection into testing is.

Combining our estimates of α and s_0, and the publicly reported seven-day average of the positive

rate[17] on July 1^{st} of 11.48%, we obtain

$$\begin{aligned} P(I_t) &= \frac{P(I_t|\sigma_t)}{\frac{1-\alpha}{s_0}(1-P(I_t|\sigma_t))+P(I_t|\sigma_t)} \\ &= \frac{11.48\%}{16.35*(100\%-11.48\%)+11.48\%} \\ &= 0.79\%. \end{aligned} \tag{15}$$

Table 3 compares our hidden-infection-method prevalence with reported prevalence for Utah as of July 1^{st}, 2020, and shows that latent prevalence is 2.62 times higher than in publicly reported data. As a result, the latent infection risk in the population is higher than suggested by the publicly reported data.

4.2. Benchmarking the Hidden-Infection-Method

Our hidden-infection-method provides estimates of viral prevalence $P(I_t)$ that can be compared to estimates of viral prevalence from randomized PCR testing. Our estimate for viral infection from our randomized testing is 0.27% (95% confidence interval 0.12% to 0.42%) for the period from May 4^{th} to July 1^{st}. In comparison, we estimate a prevalence of 0.3% from the hidden-infection-method that combines the likelihood ratio of 16.35% and the 5.0% publicly reported median positivity rate from our sample period May 4^{th} to July 1^{st}.[18] This estimate from the hidden-infection-method is similar to and within the 95% confidence interval for our estimate from randomized testing. The similarity in estimates provides external validity to the hidden-infection-method. In comparison, the median of reported prevalence, calculated as the ratio of confirmed cases minus fatalities and recoveries relative to the state population, is only 0.09%, which is only a third of the prevalence estimate from randomized testing and outside of the 95% confidence interval.

As mentioned in section 2, a key assumption for generalizability of our method beyond Utah

[17] Note that this assumes that daily positive rates provide an estimate of the stock of currently active infections among the symptomatic population, which in turn assumes that people get tested at a random (idiosyncratic) time after the first symptoms appear.

[18] The Utah State COVID-19 dashboard reported a median positivity rate of testing of 5.0%, as reported in daily tracking by covidtracking.com.

is that the likelihood ratio $\frac{1-\alpha}{s_0}$ is constant across locations and time. We therefore compare the estimates from our hidden-infection-method with another representative prevalence estimate for the state of Indiana, by Menachemi et al. (2020). For the period from April 25-29, 2020 they report a viral prevalence (using PCR tests) of 1.7% with a 95% confidence interval from 1.1% to 2.54%. The median reported positive rate for Indiana during that same period is 17.0% (data from covidtracking.com). Using our likelihood ratio of 16.35, we obtain a latent prevalence estimate of 1.55%, again close to the actual randomized testing estimate and well within the 95% confidence interval.

4.3. Estimates Across All States

Given the validation of our hidden-infection-method on random testing data, we turn to estimating our model for the rest of the United States. Table 3 shows our main results for $P(I_t)$ from equation (11), for all 50 states as of July 1^{st}, 2020. These estimates are reported in the first column of Table 3. The second column reports the current reported positive testing rate that is used in our hidden-infection-method to calculate $P(I_t)$ in the first column. This table also provides an estimate of reported prevalence from publicly available data on confirmed cases O_t, total fatalities F_t, and recovered cases C_t; $\frac{O_t-F_t-C_t}{N}$. Compared to our method, this reported prevalence estimate suffers from selection bias and requires tracking confirmed cases until they are recovered or a fatality. Such tracking is logistically challenging and, therefore, often incomplete.

One of the advantages of the hidden-infection-method is that it relies only on the positive rate—which is readily available and high quality because it is easily measurable. In contrast, data on total fatalities, and recovered cases are not always available and are relatively poor quality because of incomplete tracking. In fact, several states do not report these numbers or report questionable numbers (e.g., California, Florida, and Massachusetts). We impute recovered cases as the 21 day lag of new cases minus fatalities for all states.[19] Our approach to estimating latent prevalence is even more valuable for these states because it provides the only evidence on current viral prevalence,

[19]This calculation provides similar numbers for states that report recoveries and in some cases, seem to be exactly those numbers.

short of a randomized testing study.

The last column in Table 3 shows that latent prevalence is, on average, 2.89 times higher than reported prevalence numbers. This difference can also be seen in figure 1, which is a scatter plot of latent and reported prevalence for July 1st, 2020. If reported prevalence would truly capture all prevalence, then the two measures should line up around the provided 45-degree line. Instead, most data points line up above the 45-degree line, indicating that actual prevalence is substantially higher than reported prevalence.

Our estimates of the ratio of latent prevalence to reported prevalence are also substantially smaller than those of other studies that use alternative methods for different contexts. For example, Li et al. (2020) use Bayesian estimation with Kalman-filtering of daily confirmed case counts in China to estimate the number of latent infection cases, which is over seven times larger than reported case counts. Additionally, Aspelund et al. (2020) use partial identification methods to establish that latent infections were 5 to 10 times larger than reported infections in the early stages of COVID-19 in Iceland. Other partial identification studies such as Manski and Molinari (2020) find very large bounds for latent prevalence, such as 14.1%-61.8% for New York on April 24^{th}, 2020. Since the reported prevalence for New York on that date is 0.87%, the implied ratio of latent to reported prevalence is between 16 and over 71. In comparison, our hidden-infection-method implies a latent prevalence of 3.04% or a ratio of latent to reported prevalence of 3.5 for New York on April 24^{th}, 2020.

4.4. Tracking Prevalence

One key advantage of using the hidden-infection-method is that it allows us to track prevalence in real-time. We highlight four key features of how the hidden-infection-method captures high-frequency dynamics of disease spread in Figure 2. First, latent prevalence is 2 to 4 times higher than the reported prevalence (note the different vertical axis scales). Second, the ratio between latent and reported prevalence changes over time because our method accounts for the changes in sample selection. Sample selection changes as the set of cases accounted for in the publicly reported case counts, fatalities, or recovered cases varies over time. Third—and related to the changes in sample

selection—it is worth highlighting that latent prevalence rate from the hidden-infection-method and the reported prevalence depend on different data inputs. Our latent prevalence measure relies on positive testing rates, which account for changes in testing availability. Therefore, changes in positive testing rates are more likely to reflect disease spread instead of changes in testing rates. In contrast, reported prevalence is impacted by testing rates, recovery reporting rates, and fatality reporting rates, all of which introduce their own sample selection biases, which themselves can change over time. Fourth, our latent prevalence measure generally leads reported prevalence. For example, in Utah the latent prevalence peaks on June 25th almost a month before reported prevalence on July 24th. The lag in reported prevalence is most likely driven by reporting delays or imputations in fatalities and recoveries.

An important limitation that any user of our hidden-infection-method (11) should keep in mind is that it will yield poor approximations if testing is extremely rationed such that only cases with information on likely exposure to COVID-19 beyond symptoms are tested (e.g., contact tracing). For example, several states exhibited values of $P(I_t|T_t,\sigma_t) = 1$ for several weeks after the first confirmed case. This value was mainly driven by the fact that the only tests being conducted were on highly symptomatic people, which at the same time had known exposure to COVID-19, e.g., due to travel. However, if the rationing of tests is more like a lottery conditional on symptoms, our assumptions will hold, as we showed in our field data from Utah. Going forward, with increased testing capabilities and more widespread asymptomatic testing, our method can still be useful in providing lower bounds on COVID-19 prevalence, as shown in equation (14).

5. Conclusion

This paper provides a method to measure COVID-19 prevalence, correcting for sample selection in symptom-based testing data, and incomplete tracking of recovered cases and fatalities. We show that our statistic measures the latent prevalence of COVID-19 correctly for any generalized SIR model with time-varying infection and removal rates as well as general time-varying matching functions. Importantly, we provide supporting evidence that the required conditional independence

assumption holds in our data and that our method is able to correctly predict prevalence as measured by representative randomized testing studies. We calculate latent prevalence for all 50 US states, showing that latent prevalence is likely 2-3 times higher than reported, and that sample selection of prevalence is time-varying.

The methods developed here link economic responses of individuals through voluntary social distancing, publicly available positivity rates, and latent prevalence. Our methods are not only timely for policymakers during this pandemic, but also can be applied to future outbreaks where sample selection and behavioral responses can provide biased and time-delayed estimates.

References

Allcott, H., L. Boxell, J. Conway, M. Gentzkow, M. Thaler, and D. Yang, "Differences in Social Distancing During the Coronavirus Pandemic," *NBER Working Paper*, 2020.

Arkolakis, C., A. Costinot, and A. Rodriguez-Clare, "New Trade Models, Same Old Gains?," *American Economic Review*, 2012.

Aspelund, K., M. Droste, J. Stock, and C. Walker, "Identification and Estimation of Undetected COVID-19 Cases using Testing Data from Iceland," *NBER Working Paper*, 2020.

Atkeson, A., K. Kopecky, and T. Zha, "Estimating and Forecasting Disease Scenarios for COVID-19 with an SIR Model," *NBER Working Paper*, 2020.

Berger, D., K Herkenhoff, and S. Mongey, "An SEIR Infectious Disease Model with Testing and Conditional Quarantine," *NBER Working Paper*, 2020.

Brzezinski, A., V. Kecht, and D. Dijcke, "The Cost of Staying Open: Voluntary Social Distancing and Lockdowns in the US," *Working Paper, Oxford University*, 2020.

Burger, R. and Z. McLaren, "An Econometric Method for Estimating Population Parameters from Non-random Samples: An application to Clinical Case Finding," *Health Economics*, 2017, *26*, 1110–1122.

Bursztyn, L., A. Rao, C. Roth, and D. Yanagizawa-Drott, "Misinformation During a Pandemic," *NBER Working Paper*, 2020.

Chetty, R., "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, 2009.

Chetty, R., J. Friedman, N. Hendren, and M. Stepner, "How Did COVID-19 and Stabilization Policies Affect Spending and Employment? A New Real-Time Economic Tracker Based on Private Sector Data," *NBER Working Paper*, 2020.

Eichenbaum, M., S. Rebelo, and M. Trabandt, "The Macroeconomics of Epidemics," *NBER Working Paper*, 2020.

Eichenbaum, M., S. Rebelo, and M. Trabandt, "The Macroeconomics of Testing and Quarantining," *NBER Working Paper*, 2020.

Farboodi, M., G. Jarosch, and R. Shimer, "Internal and External Effects of Social Distancing in a Pandemic," *Working Paper, Becker Friedman Institute*, 2020.

Fernandez-Villaverde, J. and C. Jones, "Estimating and Simulating a SIRD Model of COVID-19 for Many Countries, States, and Cities," *Working Paper, Stanford University*, 2020.

Kermack, W and A. McKendrick, "A contribution to the mathematical theory of epidemics, part I," *Proceedings of the Royal Society of London*, 1927.

Korolev, I., "Identification and Estimation of the SEIRD Epidemic Model for COVID-19," *Working Paper, Binghamton University*, 2020.

Lampos, V., M. Majumder, E. Yom-Tov, M. Edelstein, S. Moura, Y. Hamada, M. Rangaka, R. McKendry, and I. Cox, "Tracking COVID-19 using online search," *arXiv:2003.08086*, 2020.

Li, R., S. Pei, B. Chen, Y. Song, T. Zhang, and W. Yang, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," *Science*, 2020.

Manski, C. and F. Molinari, "Estimating the COVID-19 Infection Rate: Anatomy of an Inference Problem," *Journal of Econometrics*, 2020.

Mclaren, Zoe, Nathan Seegert, Mac Gaulin, Mu-Jeung Yang, Matthew Samore, Steve Alder, Adam Looney, Andy Pavia, Tom Greene, Brian Orleans, Angela Presson, and Kristina Stratford, "Field Study Design Choices In Practice," *Working Paper*, 2020.

Menachemi, N., C. Yiannoutsos, B. Dixon, T. Duszynski, W. Fadel, K. Wools-Kaloustian, N. Needleman, K. Box, V. Caine, C. Norwood, L. Weaver, and P. Halverson, "Population Point Prevalence of SARS-CoV-2 Infection Based on a Statewide Random Sample — Indiana, April 25–29, 2020," *CDC Morbidity and Mortality Weekly Report*, 2020.

Menni, C., A. Valdes, M. Freydin, S. Ganesh, J. Moustafa, A. Visconti, P. Hysi, R. Bowyer, M. Mangino, M. Falchi, J. Wolf, C. Steves, and T. Spector, "Loss of smell and taste in combination with other symptoms is a strong predictor of COVID-19 infection," *Nature Medicine*, 2020.

Samore, Matthew, Steve Alder, Adam Looney, Andy Pavia, Tom Greene, Nathan Seegert, Mac Gaulin, Mu-Jeung Yang, Brian Orleans, Angela Presson, and Kristina Stratford, "Seroprevalence of SARS-CoV-2–Specific Antibodies Among Central-Utah Residents," *Working Paper*, 2020.

Simonov, A., S. Sacher, J. Dubé, and S. Biswas, "The Persuasive Effect of Fox News: Non-Compliance with Social Distancing During the COVID-19 Pandemic," *NBER Working Paper*, 2020.

Stock, J., "Data Gaps and the Policy Response to the Novel Coronavirus," *NBER Working Paper*, 2020.

Yang, M., A. Looney, M. Gaulin, and N. Seegert, "What Drives the Effectiveness of Social Distancing in Combatting COVID-19 across U.S. States?," *Working Paper, University of Utah*, 2020.

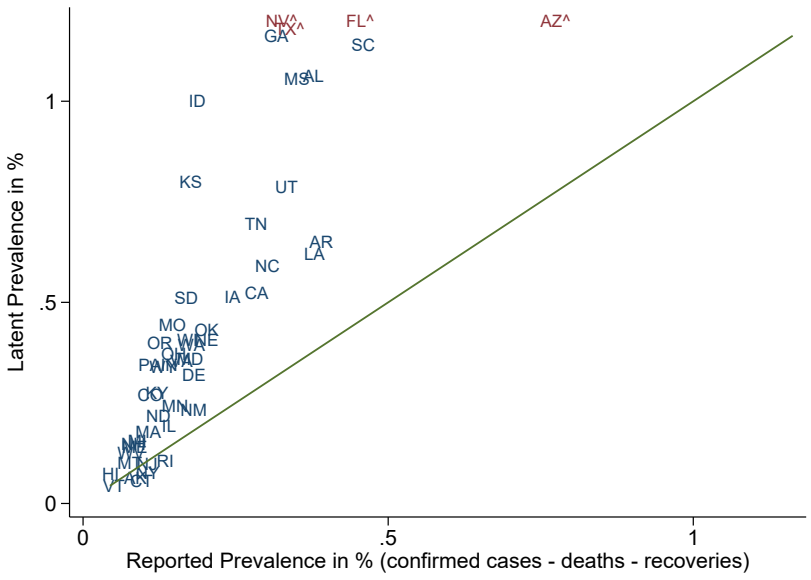

Figure 1: Latent prevalence is calculated using equation (11). Reported prevalence is calculated as $\frac{O_t - F_t - C_t}{N}$, where O_t are confirmed cases, F_t are total fatalities and C_t are recovered cases. States with ^ and red text are displayed lower on the Y-axis then their true values for convenience. Data is from the COVID tracking project.

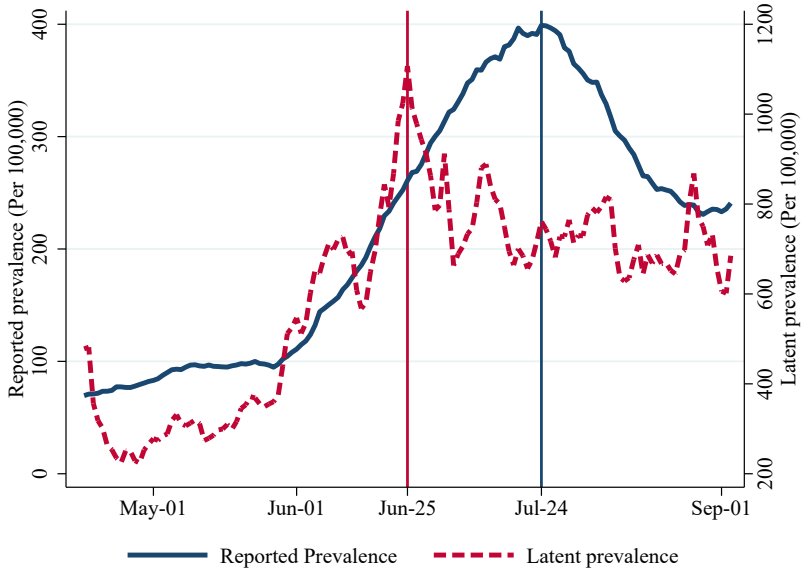

Figure 2: Time path of latent and reported prevalence in Utah. Latent prevalence is defined as fraction of currently infected state population. Reported prevalence is calculated as $\frac{O_t - F_t - C_t}{N}$, where O_t are confirmed cases, F_t are total fatalities and C_t are recovered cases. Data is from the COVID tracking project, which takes daily snapshots of the Utah state COVID dashboard.

Table 1: Sample Characteristics

Notes: This table provides descriptive statistics from the US Census and our survey that provides an overview of our sample.

	Salt Lake	Utah	Davis	Summit
Panel A: Aggregate Data from Census and CDC				
Population	1,120,805	590,440	340,621	40,511
Household Population	3.0	3.6	3.2	2.7
Median Age	34.7	27.2	32.5	39.9
% Hispanic	18.1	11.4	9.1	6.0
Reported Prevalence (5/7/2020)	268	206	91	913
Reported Deaths (5/7/2020)	39	11	2	0
Panel B: Sample Characteristics				
Households Sampled	12,138	5,202	4,023	4,075
Households In Sample	2,673	1,130	1,029	280
Households with Antibody Test	2,068	890	816	217
Households with Viral Test	1,589	715	706	144
Individuals In Sample	5,500	2,684	2,303	480
Individuals with Antibody Test	4,060	2,060	1,750	351
Individuals with Viral Test	3,129	1,603	1,487	232
Panel C: Individual Survey and Testing Characteristics				
Median Age	42.2	40.2	43.4	51.5
% Hispanic	8.97	8.85	3.46	5.70
% Female	52.5	52.9	51.7	54.6
% Very Concerned	9.89	12.5	8.26	9.62
Viral Prevalence	0.286	0.187	0.202	0.851
Antibody Prevalence	0.98	1.26	0.91	2.81

Table 2: Key parameters from Randomized Testing

Panel A displays estimates of $P(I_t, T_t|\sigma_t), P(I_t|\sigma_t), P(T_t|\sigma_t)$ for different symptoms and reports the t-Statistic for a test for the null hypothesis of conditional independence as in (6): $P(I_t, T_t|\sigma_t) = P(I_t|\sigma_t) \cdot P(T_t|\sigma_t)$. This t-statistic is calculated as a test of nonlinear combinations of estimators from a seemingly unrelated regression of infection, tested, and their union on an indicator for symptomatic individuals. We also report the odds ratio and the confidence interval in square brackets below. Column (1) reports estimates using the loss of smell or taste (anosmia) as a single symptom. In Columns (2) and (3), we consider someone to be symptomatic if they had anosmia and at least one more or three more symptoms, respectively, from the list: stuffy nose, diarrhea, abdominal pain, nausea, and fever. Panel B displays estimates of fraction of infected persons who display symptoms $(1-\alpha)$, listed on top and fraction of uninfected persons who display symptoms s_0. Point estimates are presented with estimated standard errors below in parentheses. Weighted estimates use sampling weights. Data are from representative state-wide testing in Utah from May to July 2020.

	Symptoms		
	Anosmia	Anosmia + at least 1 symptom	Anosmia + at least 3 symptoms
	(1)	(2)	(3)
	A: Test of Conditional Independence		
$P(I_t, T_t\|\sigma_t)$	2.30%	2.21%	3.75%
$P(I_t\|\sigma_t)P(T_t\|\sigma_t)$	2.25%	2.38%	3.23%
t-Statistic	0.0893	-0.2094	0.5232
$\frac{P(I_t,T_t\|\sigma_t)}{P(I_t\|\sigma_t)P(T_t\|\sigma_t)}$	1.0225	0.9286	1.1613
	[0.518, 1.526]	[0.304, 1.554]	[0.482, 1.841]
	B: Parameter estimates		
Unweighted			
$1-\alpha$	5.43	3.26	1.09
	(2.38)	(1.86)	(1.09)
s_0	0.39	0.48	0.20
	(0.07)	(0.08)	(0.05)
Likelihood Ratio	13.85	6.82	5.54
	(6.50)	(4.02)	(5.68)
Weighted			
$1-\alpha$	5.68	2.75	0.75
	(2.45)	(1.73)	(0.91)
s_0	0.35	0.50	0.23
	(0.07)	(0.08)	(0.05)
Likelihood Ratio	16.35	5.53	3.29
	(8.89)	(3.78)	(3.40)

Table 3: Latent vs Reported Prevalence

This table presents state level estimates of our model parameters. Positive rate is fraction of tests reported that are positive for COVID-19 from the COVID tracking project. $P(I_t)$ is latent prevalence from equation (11). Cases-deaths, is calculated as number of confirmed cases minus fatalities as fraction of state population. Rep.Prev. is baseline reported prevalence, calculated as number of confirmed cases minus fatalities and recoveries as fraction of state population. Ratio is the ratio of estimated latent prevalence to baseline reported prevalence. Estimates are calibrated on July 1^{st}, 2020.

State	$P(I_t)$ (1)	Positive rate (2)	Cases - deaths (3)	Rep. Prev. (4)	Ratio (5)
			A: Utah		
UT	**0.79%**	**11.48%**	**0.69%**	**0.30%**	**2.62**
			B: All other states		
AK	0.06%	1.04%	0.13%	0.05%	1.22
AL	1.06%	14.95%	0.77%	0.35%	3.07
AR	0.65%	9.67%	0.69%	0.36%	1.83
AZ	2.35%	28.22%	1.12%	0.74%	3.19
CA	0.52%	7.92%	0.58%	0.25%	2.09
CO	0.27%	4.23%	0.53%	0.07%	3.60
CT	0.05%	0.89%	1.19%	0.06%	0.88
DE	0.32%	4.97%	1.12%	0.15%	2.15
FL	1.35%	18.24%	0.71%	0.42%	3.23
GA	1.16%	16.14%	0.76%	0.28%	4.13
HI	0.07%	1.20%	0.06%	0.02%	4.45
IA	0.51%	7.80%	0.90%	0.22%	2.37
ID	1.00%	14.18%	0.33%	0.16%	6.30
IL	0.19%	3.07%	1.09%	0.11%	1.68
IN	0.34%	5.33%	0.64%	0.11%	3.04
KS	0.80%	11.63%	0.51%	0.14%	5.56
KY	0.27%	4.29%	0.34%	0.09%	3.11
LA	0.62%	9.27%	1.23%	0.35%	1.79
MA	0.18%	2.82%	1.45%	0.07%	2.47
MD	0.36%	5.56%	1.06%	0.14%	2.58
ME	0.14%	2.27%	0.24%	0.05%	2.90
MI	0.16%	2.48%	0.65%	0.06%	2.64
MN	0.24%	3.81%	0.62%	0.11%	2.11
MO	0.44%	6.80%	0.34%	0.11%	4.07
MS	1.06%	14.85%	0.90%	0.32%	3.35
MT	0.10%	1.61%	0.09%	0.04%	2.39
NC	0.59%	8.85%	0.61%	0.27%	2.21
ND	0.22%	3.44%	0.46%	0.09%	2.46
NE	0.41%	6.25%	0.97%	0.17%	2.41
NH	0.15%	2.38%	0.39%	0.05%	3.13
NJ	0.10%	1.58%	1.76%	0.07%	1.33
NM	0.23%	3.67%	0.56%	0.15%	1.60
NV	1.29%	17.57%	0.59%	0.28%	4.52
NY	0.07%	1.21%	1.90%	0.07%	1.04
OH	0.37%	5.73%	0.43%	0.11%	3.28
OK	0.43%	6.63%	0.35%	0.17%	2.58
OR	0.40%	6.15%	0.20%	0.09%	4.43
PA	0.34%	5.34%	0.63%	0.08%	4.51
RI	0.10%	1.68%	1.52%	0.11%	0.98
SC	1.14%	15.85%	0.71%	0.43%	2.68
SD	0.51%	7.75%	0.75%	0.14%	3.78
TN	0.70%	10.28%	0.65%	0.25%	2.78
TX	1.28%	17.53%	0.56%	0.30%	4.28
VA	0.35%	5.49%	0.71%	0.13%	2.77
VT	0.04%	0.71%	0.18%	0.02%	2.40
WA	0.39%	6.06%	0.46%	0.14%	2.80
WI	0.41%	6.24%	0.54%	0.14%	2.91
WV	0.13%	2.01%	0.16%	0.04%	2.99
WY	0.34%	5.27%	0.26%	0.09%	3.60
AVERAGE	**0.50%**	**7.25%**	**0.68%**	**0.17%**	**2.89**