What is the Active Prevalence of COVID-19?

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

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

Abstract

We provide a method to track the 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 in May and June 2020. We validate our method using external studies in Indiana in April 2020 and two counties in Utah in March 2021. In all three locations and times, our estimates of latent prevalence are within the 95 percent confidence intervals of prevalence estimates from randomized testing. 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@ou.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 infectious are crucial for policymakers and researchers trying to understand and predict COVID-19 dynamics. Latent prevalence—the unknown fraction of infected individuals in the population—is generally different from the publicly reported prevalence—the reported case counts divided by the total population. However, latent prevalence is empirically challenging to estimate because testing is often reserved for individuals exhibiting symptoms. This limitation induces sampling bias and, as a result, causes publicly reported prevalence, 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]

This study attempts to solve the selection problem by developing a hidden-infection method that is easily applicable and can be calibrated using randomized testing data. This method estimates the prevalence in local areas based on real-time public data by applying Bayes' Law. The method is flexible enough to allow for time-varying infectiousness, which has been shown to be important for understanding the economic consequences of COVID-19 (Chetty et al., 2020; Yang et al., 2020). We validate the predictions of this method using randomized testing data from a large-scale field experiment in Utah of over 10,000 participants in May and June of 2020. We then apply our methods to different locations and time periods: Indiana in April 2020 and two counties in Utah in March of 2021. We find that our estimates of latent prevalence are remarkably similar to the estimates from those randomized field tests.

Our hidden-infection method estimates the 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), which shows that one can

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

[3] Our measurement is related in spirit to sufficient welfare statistics as in Chetty (2009) and Arkolakis et al. (2012), but

use Bayes' Law to translate positive testing rates into latent prevalence, as long as infections are independent of testing conditional on symptoms and tested individuals are symptomatic. We extend this analysis in two ways. First, we incorporate experimental data to calibrate the model and later validate it. Second, we improve the definition of a symptomatic individual to that of an individual with a high risk of infection. We rely on recent refinements of the machine learning method LASSO, which can be shown to select the correct predictor variables reliably (see the discussion of "oracle properties" in Ahrens, Hansen and Schaffer (2020)). LASSO automates the selection of variables that predict the risk of infection and include symptom and non-symptom variables, e.g., employment outside of home or health-care provider. These extensions allow the method to adapt as the virus changes or to be applied for future diseases or pandemics.

The hidden-infection method requires only one parameter: the likelihood ratio of symptomatic individuals among infected relative to uninfected persons. Again, our definition of symptomatic includes all individuals that have a high predicted probability of infection given covariates. We estimate this likelihood ratio using data from a large-scale field experiment of roughly 10,000 randomly selected individuals in Utah between May and July 2020. With this parameter, the method provides a formula to translate publicly available positivity rates, defined as the proportion of tested individuals with a positive COVID result, into latent prevalence.

We pursue two approaches to validate our method. First, we directly test the key conditional independence assumption using our randomized testing and health survey data. We show that we cannot reject the null that the assumption holds in our micro-data (p-value of 0.943). Second, we validate our method's predictions of prevalence by comparing it to prevalence from randomized testing in Utah between May 1^{st} and July 1^{st} 2020, in Indiana between April 25^{th} and April 29^{th} 2020, and in Utah in March 2021 (Samore et al., 2020; Menachemi et al., 2020). Our method's predictions in all three cases are based on the likelihood ratio that we estimate using our first wave of data from Utah in the summer of 2020. For Indiana, we estimate a latent prevalence of 1.8%, in comparison to 1.7% from randomized testing (Menachemi et al., 2020) for the spring of 2020.

differs in its focus (i.e., latent disease prevalence).

For Salt Lake County and Utah County in March 2021, we estimate a latent prevalence of 0.69% and 0.57%, which are approximately the same as their respective randomized testing estimates of 0.74% and 0.47%. In all four cases, the prevalence estimate from our method is within the 95% confidence intervals of the randomized testing estimates. This validation evidence is encouraging for the ability of researchers to apply our method to areas and time periods that are different from those that generated our estimate for the likelihood ratio.[4]

We provide estimates for all 50 US states on July 1^{st} 2020. We find that latent prevalence is $2-3$ times higher on average than publicly reported prevalence. Additionally, we show that sample selection is time-varying by comparing the time series of latent and reported prevalence.

This paper contributes to the fast-growing economics literature analyzing COVID-19 disease dynamics. First, a number of papers have developed different approaches to estimate Susceptible, Infectious, and Recovered (SIR) type of models, such as Atkeson et al. (2020), Korolev (2020), Fernandez-Villaverde and Jones (2020), and Yang et al. (2020). Our method allows researchers to directly measure the time path of COVID-19 prevalence without the use of a SIR model. Second, a literature in epidemiology has pooled international or cross-state data in combination with strong functional form assumptions to correct for sample selection, see Grewelle and Leo (2020), Favero (2020), Fisman and Tuite (2020), and Benatia et al. (2020). Our work offers an alternative approach that is less reliant on functional forms and pooled regression analysis. 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 lower bounds for latent prevalence by relaxing one of our key assumptions.

2. Framework

In this section, we construct a Bayesian updating formula and lay down assumptions that allow researchers to infer the fraction of infected individuals based on positivity rates from testing. The

[4]This test of generalizability is especially important since our sample from Utah might not be representative of other more diverse US States.

derivation using Bayes's rule and Assumption A below follows work by Stock (2020). In addition, we provide an estimation strategy for the key parameter of our formula. We conclude the section with an alternative method that provides a lower bound for the proportion of infected based on weaker assumptions.

2.1. The Hidden-Infection Method

Consider the experiment of drawing an individual randomly out of a population in a geographic location s and time period t, where s and t belong to a certain scope of analysis. We define three events based on the outcomes of this random draw: I_{st}, τ_{st}, and σ_{st}. The first event is I_{st} and occurs when the randomly drawn individual is infected with COVID-19; the second is τ_{st} and occurs when the individual has been recently tested for COVID-19; finally, σ_{st} occurs when the individual has symptoms and characteristics that predict COVID-19 infection, for example, as diagnosed by a medical professional. The event σ_{st} can be understood as an outcome of a unobserved, latent prediction model by a medical professional. We call individuals in the σ_{st} event "symptomatic individuals" for simplicity throughout the text. However, we do not restrict the σ_{st} event only to individuals that exhibit COVID symptoms. Rather, the event occurs whenever the individual's or medical provider's best prediction indicates a high probability of infection. That inference is most heavily based on symptoms but may also depend on other information such as exposure, contact tracing, being a healthcare worker, etc. We formalize a prediction model for the probability of infection in Section 2.2. As we describe in footnote 6, our method could also be used with the event τ_{st} ("being tested") instead of the event of being symptomatic σ_{st}. However, we prefer σ_{st}, as it is less affected by the testing regime, which is likely to vary across locations and over time.

The probability of any event equals the fraction of individuals with the outcome associated with that event in the population. For example, $\mathbb{P}[I_{st}]$ equals the fraction of individuals in location s at

time t that are infected with COVID-19. Define the following quantities,

$$\alpha_{st}^0 = \mathbb{P}[\sigma_{st}|I_{st}^c], \tag{1}$$

$$\alpha_{st}^1 = \mathbb{P}[\sigma_{st}|I_{st}], \tag{2}$$

where I_{st}^c denotes the complement of event I_{st}, namely, when the randomly drawn individual is not infected. Each α represents the fraction of symptomatic individuals among the infected (α_{st}^1) or not infected (α_{st}^0), respectively.

Next, we use Bayes's rule to derive an expression for $\mathbb{P}[I_{st}]$ as a function of $\mathbb{P}[\sigma_{st}]$, α_{st}^1, and $\mathbb{P}[I_{st}|\sigma_{st}]$,

$$\begin{aligned}\mathbb{P}[I_{st}|\sigma_{st}] &= \frac{\mathbb{P}[I_{st} \cap \sigma_{st}]}{\mathbb{P}[\sigma_{st}]} = \frac{\mathbb{P}[\sigma_{st}|I_{st}]\mathbb{P}[I_{st}]}{\mathbb{P}[\sigma_{st}]} = \alpha_{st}^1 \frac{\mathbb{P}[I_{st}]}{\mathbb{P}[\sigma_{st}]}, \\ \mathbb{P}[I_{st}] &= \frac{\mathbb{P}[I_{st}|\sigma_{st}]\mathbb{P}[\sigma_{st}]}{\alpha_{st}^1}.\end{aligned} \tag{3}$$

Similarly, we obtain an expression for $\mathbb{P}[\sigma_{st}]$ as a function of α_{st}^0, α_{st}^1, and $\mathbb{P}[I_{st}]$,

$$\begin{aligned}\mathbb{P}[\sigma_{st}] &= \mathbb{P}[\sigma_{st}|I_{st}]\mathbb{P}[I_{st}] + \mathbb{P}[\sigma_{st}|I_{st}^c]\mathbb{P}[I_{st}^c] \\ &= \alpha_{st}^1 \mathbb{P}[I_{st}] + \alpha_{st}^0(1 - \mathbb{P}[I_{st}]) = (\alpha_{st}^1 - \alpha_{st}^0)\mathbb{P}[I_{st}] + \alpha_{st}^0.\end{aligned} \tag{4}$$

Plug (4) in (3) and solve for $\mathbb{P}[I_{st}]$,

$$\begin{aligned}\mathbb{P}[I_{st}] &= \frac{\mathbb{P}[I_{st}|\sigma_{st}]\left\{(\alpha_{st}^1 - \alpha_{st}^0)\mathbb{P}[I_{st}] + \alpha_{st}^0\right\}}{\alpha_{st}^1} \\ &= \frac{\mathbb{P}[I_{st}|\sigma_{st}]}{\frac{\alpha_{st}^1}{\alpha_{st}^0}(1 - \mathbb{P}[I_{st}|\sigma_{st}]) + \mathbb{P}[I_{st}|\sigma_{st}]}.\end{aligned} \tag{5}$$

Equation 5 gives the proportion of infected people in the population ($\mathbb{P}[I_{st}]$) as a function of the proportion of infected people in the sub-population of symptomatic individuals ($\mathbb{P}[I_{st}|\sigma_{st}]$) and the ratio $\alpha_{st}^1/\alpha_{st}^0$. In what follows, we make assumptions on the joint distribution of these events to

obtain an equation relating $\mathbb{P}[I_{st}]$ to positivity rate, i.e, $\mathbb{P}[I_{st}|\tau_{st}]$. We study deviations from these assumptions in Section 2.3.

Assumption A. *(a) Everyone that is tested is symptomatic,*

$$\mathbb{P}[\sigma_{st}|\tau_{st}] = 1 \quad \forall s,t. \tag{6}$$

(b) Conditional on being symptomatic, being infected and being tested are independent events:

$$\mathbb{P}[I_{st} \cap \tau_{st}|\sigma_{st}] = \mathbb{P}[I_{st}|\sigma_{st}]\mathbb{P}[\tau_{st}|\sigma_{st}] \quad \forall s,t. \tag{7}$$

Assumption A(a) essentially says that tested individuals had a strong reason to be tested: their best prediction indicated a high probability of infection, that is, the σ_{st} event. The σ_{st} event may be a function of not only symptoms but other information such as demographic characteristics. We give full details on the relationship between σ_{st} and individual characteristics in Section 2.2 below.[5] Assumption A(a) implies that $\mathbb{P}[I_{st}|\tau_{st}] = \mathbb{P}[I_{st}|\sigma_{st} \cap \tau_{st}]$.

Assumption A(b) reflects our definition of "best prediction." Part (b) of the assumption says that the mere fact of being tested (τ_{st}) does not improve the prediction power for infection (I_{st}) that the best-prediction event σ_{st} has. In other words, σ_{st} condenses all the information from symptoms and other factors that help predict infection. We describe a prediction model for probability of infection in Section 2.2. Assumption A(b) implies that $\mathbb{P}[I_{st}|\sigma_{st} \cap \tau_{st}] = \mathbb{P}[I_{st}|\sigma_{st}]$. Thus, both parts of Assumption A imply $\mathbb{P}[I_{st}|\tau_{st}] = \mathbb{P}[I_{st}|\sigma_{st}]$ and allow us to retrieve $\mathbb{P}[I_{st}|\sigma_{st}]$ using publicly available data on positivity rates of testing : $\mathbb{P}[I_{st}|\tau_{st}]$. Therefore, if the researcher has access to $\mathbb{P}[I_{st}|\tau_{st}]$ and the ratio $\alpha^1_{st}/\alpha^0_{st}$, the researcher obtains the fraction of infected people under Assumption A.

[5] It is important to emphasize that our definition of σ_{st} does not necessarily mean that a certain symptom or demographic characteristic automatically implies that the individual is "symptomatic." Rather, a symptomatic individual is the one that exhibits any combination of symptoms and demographics, among a large class of combinations, that lead to a predicted probability of infection that is high. This concept is intended to model the process by which testing regimes allow or select individuals to be tested, e.g., a medical professional determining that a confluence of symptoms or conditions add up to a likely infection requiring testing. For example, our model in Section 2.2 is sufficiently flexible to capture testing regimes with the requirement of one or a combination of observable symptoms, as well as different testing thresholds for at-risk demographics.

The ability to obtain latent prevalence $\mathbb{P}[I_{st}]$ as a function of positivity rate of testing $\mathbb{P}[I_{st}|\tau_{st}]$ is extremely valuable to track COVID-19 in real time, but the exact relationship relies on the likelihood ratio $\alpha_{st}^1/\alpha_{st}^0$. Estimation of this ratio requires experimental data and is costly to conduct repeatedly at multiple times and places. Therefore, researchers need to rely on assuming the ratio is relatively constant, at least in a neighborhood of region s at time t of the experimental data. We argue that the assumption of a constant likelihood ratio $\alpha_{st}^1/\alpha_{st}^0$ is reasonable if, for example, differences in local testing regimes or pandemic patterns affect α_{st}^1 and α_{st}^0 proportionately, such that the ratio stays constant. Our empirical evidence in Section 4.2 suggests this simplifying assumption is plausible within a certain scope of time and geography. Of course, researchers will not be able to extrapolate the ratio too far from the location and time period of the experimental data. For example, local and seasonal variation in the prevalence of the flu might imply that α_{st}^0 varies disproportionately relative to α_{st}^1. We discuss this issue in Appendix B and suggest solutions.

Assumption B. *The proportion of symptomatic among infected individuals divided by the proportion of symptomatic among uninfected individuals is constant across time and space:*

$$\frac{\alpha_{st}^1}{\alpha_{st}^0} = \lambda \quad \forall s,t. \tag{8}$$

Our hidden-infection-method obtains the latent prevalence under Assumptions A–B using some estimate of λ and the positivity rate with the following equation,[6]

$$\mathbb{P}[I_{st}] = \frac{\mathbb{P}[I_{st}|\tau_{st}]}{\lambda(1-\mathbb{P}[I_{st}|\tau_{st}]) + \mathbb{P}[I_{st}|\tau_{st}]}. \tag{9}$$

[6]We note that a version of (9) could be derived by starting from $\mathbb{P}[I_{st}|\tau_{st}]$ instead of $\mathbb{P}[I_{st}|\sigma_{st}]$ and leading to an estimate of the latent prevalence of $\mathbb{P}[I_{st}] = \frac{\mathbb{P}[I_{st}|\tau_{st}]}{\Lambda_{st}(1-\mathbb{P}[I_{st}|\tau_{st}])+\mathbb{P}[I_{st}|\tau_{st}]}$. The key difference is that $\Lambda_{st} = \frac{\mathbb{P}[\tau_{st}|I_{st}]}{\mathbb{P}[\tau_{st}|I_{st}^C]}$ is now the likelihood ratio of infected people being tested to non-infected people being tested. Although this alternative formulation of our hidden infection method could also be used to track latent prevalence over time, we argue that it is more reasonable to assume that λ_{st} is constant over time and space than Λ_{st} because the latter varies with changes in testing regimes, which were frequent during the first years of the pandemic.

2.2. Estimation of λ

The symptomatic event σ_{st}, as we have defined it, is unobserved and has to be estimated. The event occurs when a randomly drawn individual displays characteristics that strongly predict infection by COVID. These characteristics may include not only symptoms such as loss of taste or smell, fever, body pain, sore throat, but also variables such as race, gender, sector of employment, etc.

In this section, we specify a prediction model for infection as a function of individual characteristics. Estimation of such a model allows us to construct an indicator variable for the event σ_{st} as a function of individual characteristics. The ultimate goal is to produce an estimate for λ to be used along with Equation 9. The analysis applies to a fixed location s and time period t for which the researcher has experimental data. Thus, we drop the subscripts s and t for ease of notation in the rest of this section.

Define the binary random variables, $D^I = \mathbb{I}\{I\}$, $D^\sigma = \mathbb{I}\{\sigma\}$, and $D^\tau = \mathbb{I}\{\tau\}$. Let X be a $1 \times K$ random vector of individual characteristics that predict infection by COVID plus an intercept. For a $K \times 1$ vector of parameters β and a threshold $p \in (0,1)$,

$$D^I = X\beta + U, \qquad \mathbb{E}[U|X] = 0, \tag{10}$$

$$D^\sigma = \mathbb{I}\{X\beta > p\}. \tag{11}$$

The linear probability model in Equation 10 is a flexible specification for $\mathbb{P}[D^I = 1|X]$ to the extent that the vector X may include high-order polynomials and interaction terms of individual characteristics. Equation 11 says that the σ event occurs when the probability of being infected is higher than p, where the researcher specifies p, for example, to maximize the predictability of the model. In our empirical section, we choose p to be the unconditional probability of infection.

We can write the parameter λ in terms of moments of (D^I, D^σ, D^τ),

$$\lambda = \frac{\mathbb{P}[\sigma|I]}{\mathbb{P}[\sigma|I^c]} = \frac{\mathbb{E}\left[D^\sigma D^I\right]\mathbb{E}\left[\left(1-D^I\right)\right]}{\mathbb{E}[D^I]\,\mathbb{E}[D^\sigma(1-D^I)]}. \tag{12}$$

We may also use the model to assess the validity of the independence condition in Assumption A by obtaining an estimate of the following parameter,

$$\mu = \frac{\mathbb{P}[I \cap \tau | \sigma]}{\mathbb{P}[I|\sigma]\,\mathbb{P}[\tau|\sigma]} = \frac{\mathbb{E}\left[D^I D^\tau D^\sigma\right]\mathbb{E}[D^\sigma]}{\mathbb{E}[D^I D^\sigma]\,\mathbb{E}[D^\tau D^\sigma]}. \tag{13}$$

If the independence condition holds, then μ should equal 1.

The researcher has a sample of *iid* data, (X_i, D_i^I, D_i^τ), $i = 1, \ldots, n$, and seeks to estimate β, λ, and μ. In settings with high-dimensional X, an estimate $\widehat{\beta}$ may be obtained by LASSO in Equation 10. Then, the researcher constructs $\widehat{D}_i^\sigma = \mathbb{I}\{X_i\widehat{\beta} > p\}$ for each individual in the sample. This process yields estimates for the parameters of interest as follows,

$$\widehat{\lambda} = \frac{\left(\sum_{i=1}^n \widehat{D}_i^\sigma D_i^I\right)\left(\sum_{i=1}^n (1 - D_i^I)\right)}{\left(\sum_{i=1}^n D_i^I\right)\left(\sum_{i=1}^n \widehat{D}_i^\sigma (1 - D_i^I)\right)},$$

$$\widehat{\mu} = \frac{\left(\sum_{i=1}^n D_i^I D_i^\tau \widehat{D}_i^\sigma\right)\left(\sum_{i=1}^n \widehat{D}_i^\sigma\right)}{\left(\sum_{i=1}^n D_i^I \widehat{D}_i^\sigma\right)\left(\sum_{i=1}^n D_i^\tau \widehat{D}_i^\sigma\right)}.$$

The procedure to obtain standard errors and construct confidence intervals is non-standard for two reasons. First, the distribution of $\widehat{\beta}$ is generally non-Gaussian when the researcher has high-dimensional X and employs LASSO. Second, $D^\sigma = \mathbb{I}\{X\beta > p\}$ is a discontinuous function of β. We address both issues by relying on the residual bootstrap for LASSO developed by Chatterjee and Lahiri (2011) and by approximating the indicator function by a Normal CDF with small variance. Appendix A describes the procedure in detail.

2.3. Relaxing Assumptions

Assumption A may be strong in some contexts. The vector of characteristics X may not be rich enough to produce a best prediction of infection that satisfies both requirements of Assumption A. A much weaker requirement is to assume that the fraction of infected among symptomatic is greater than or equal to the fraction of infected among tested individuals.

Assumption C. $\mathbb{P}[I_{st}|\tau_{st}] \leq \mathbb{P}[I_{st}|\sigma_{st}] \quad \forall s,t.$

Even if the conditional independence Assumption A fails, Assumption C is likely to hold because σ is defined as best predictor of infection, and we expect it to contain more information about infection compared to just the fact the individual was tested. It becomes natural to assume that there is a higher proportion of infected individuals among σ individuals compared to τ individuals. It turns out that researchers may replace Assumption A with Assumption C and still obtain a lower bound on the proportion of infected individuals. To see that, note that Equation 5 is an increasing function of $\mathbb{P}[I_{st}|\sigma_{st}]$. Under Assumptions B and C,

$$\mathbb{P}[I_{st}] \geq \frac{\mathbb{P}[I_{st}|\tau_{st}]}{\lambda(1-\mathbb{P}[I_{st}|\tau_{st}])+\mathbb{P}[I_{st}|\tau_{st}]}. \tag{14}$$

Therefore, researchers may compute a lower bound for $\mathbb{P}[I_{st}]$ using λ and positivity rates in any given region and time.

The extrapolation based on Assumption B may also be considered strong in some contexts. We show in Section 4.2 that extrapolating λ to multiple locations and time periods works well: estimates of latent prevalence based on λ are statistically indistinguishable from estimates obtained by randomized testing. Despite this, researchers must exercise caution in extrapolating λ and should consider either estimating λ in their context or calibrating it using additional data.

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 is from the COVID tracking project[7] and contains the daily rates of positive tests in all state-wide COVID-19 tests. This allows us to estimate $\mathbb{P}(I_{st}|\tau_{st})$. Since the daily data are noisy, we use 7-day moving averages.

[7]Data accessed from covidtracking.com/api.

3.1. Field Experiment

The HERO project was initiated to estimate COVID-19 prevalence in Utah, and its data allow us to estimate the key parameter λ of our method. Randomized testing also provides an estimate of viral prevalence to benchmark our approach.

Between May 1^{st} 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 survey for each household member and get a PCR (viral) and serology (antibody) test if they were older than 12. Individuals were compensated with a $10 USD gift card for completing the survey and being subsequently tested, paid at the test center. 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.

In total, 8,916 addresses ultimately received a visit from our field team and were included in the in-person sample, and 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 proportion 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 proportions, 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.

Gaulin et al. (2021) investigate the important issue of non-response to invitations to participate in COVID-19 testing in our context. Using a large-scale field experiment, they show that response

proportions increase with more incentives (monetary and otherwise). The positivity rates, however, do not vary across these interventions. We also find that estimates are very similar using the in-person and letter-recruitment subsets, alleviating some concern that selection is correlated with infections, testing, or symptoms. Table C.2 in the appendix reports estimates for λ, which we use to estimate latent prevalence of 0.24% (in-person) and 0.33% (letter-recruitment). These estimates are similar to our estimate of 0.24% in the full sample.

3.2. Health Survey and Prediction Model

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). We asked participants, "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*) 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.

We use the machine learning method `rlasso` from the `lassopack` for STATA, which has its theoretical foundations summarized by Ahrens, Hansen and Schaffer (2020), including its ability to consistently select the correct predictors for an optimal prediction model. The survey provides us with a series of characteristics, including whether someone experienced symptoms, such as anosmia and fever, demographics, such as gender and race, and other risk factors such as the sector of employment and known contact with a positive case. Other characteristics include worked outside of the home, general health characteristics, school completed, and ethnicity. The LASSO is particularly helpful in our setting because we have a high number of explanatory variables (including interactions) and are a priori uncertain of which ones predict infection. In our data, infection is measured as a positive COVID-19 serology test. Moreover, we let LASSO choose among high-order interactions of these variables to search across flexible specifications. In practice, we construct $\hat{D}_i^{\sigma} = \mathbb{I}\{X_i\widehat{\beta} > p\}$ using a threshold based on the unconditional probability of infection p, which in our data is close to 1% (Equation 11). It is reassuring that the LASSO consistently selects anosmia

because the epidemological literature considers it a strong predictor of COVID-19 infection, see Menni et al. (2020). Similarly, LASSO consistently selects "known contact" with an infected person as well as its interactions with other variables.

3.3. Descriptive Evidence

Panel A of Table 1 provides descriptive statistics from the US Census and CDC for our sample counties in central Utah. These counties contain two-thirds of Utah's population. 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 proportion is roughly 15 percent.

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. We use these empirical estimates to provide external validity to the hidden-infection-method developed in this study.

4. Results

4.1. Prevalence of COVID in Utah

Panel A of Table 2 reports estimates for α^1, α^0, and the likelihood ratio λ. In all four columns, we use the LASSO machine learning method to select variables that predict symptomatic individuals according to our definition in Section 2.2. The difference across columns is the initial set of variables given to the LASSO procedure. In column 1, we limit the set of variables to those about symptoms, like headache, fever, anosmia, etc. Of these symptoms, LASSO selects anosmia as the symptom that predicts infection risk. This selection is consistent with the medical evidence in Menni et al. (2020), which show that anosmia is a particularly strong predictor of COVID-19 infection in patients. In column 2, we include nonsymptom variables, including working outside of the home, industry of work, and known exposure. Column 3 includes up to three-way interactions of symptom variables, while column 4 does three-way interactions of symptom and nonsymptom variables.

The estimate of the likelihood ratio λ is 16.35 in column 1 and is remarkably similar across all four columns. The median positivity rate in Utah from May 1^{st} to July 1^{st} 2020 is 6.38%. Plugging these estimates into Equation 9, we obtain an estimate for the latent prevalence during this period of 0.42%.[8] Said differently, our hidden-infection-method estimates that 0.42% of the population in Utah during this period were infected.

$$\begin{aligned}\mathbb{P}[I_{st}] &= \frac{\mathbb{P}[I_{st}|\tau_{st}]}{\lambda(1-\mathbb{P}[I_{st}|\tau_{st}])+\mathbb{P}[I_{st}|\tau_{st}]} \\ &= \frac{6.38\%}{16.35*(100\%-6.38\%)+6.38\%} \\ &= 0.42\%.\end{aligned} \tag{15}$$

Our estimate using the hidden-infection-method is similar to our estimate using randomized testing in Utah during the same period of time and within the 95% confidence interval of 0.12% and 0.42% for the randomized testing estimate. Specifically, using randomized testing, we estimate that, on average, 0.27% of the population was infected in Utah from May 1^{st} to July 1^{st} 2020.[9] In comparison, the publicly reported prevalence calculated as the ratio of confirmed cases minus fatalities and recoveries relative to state population is 0.109%, roughly a third of the prevalence estimate from randomized testing, and outside of the 95% confidence interval. Our hidden-infection method, therefore, provides a reasonable estimate of the latent prevalence (as estimated by randomized testing) and a better estimate than the publicly reported prevalence.

4.2. Benchmarking the Hidden-Infection-Method

In this section, we use our estimate of the likelihood ratio λ from Utah between May 1^{st} and July 1^{st} 2020, in combination with publicly reported data to out-of-sample predict prevalence from actual randomized testing in Indiana from April 25–29, 2020 and Utah from March 3^{rd} to March 13^{th} 2021.

[8] The Utah State COVID-19 dashboard reported a median positivity rate of testing of 6.38%, as reported in daily tracking by covidtracking.com.

[9] This estimate is weighted using sampling weights to account for nonresponse and sampling design.

The estimates from Utah from March 3^{rd} to March 13^{th} 2021 provide an interesting test because we targeted locations in Utah and Salt Lake County that were experiencing higher case counts, and it was significantly later in the pandemic. This test is particularly interesting because the conditional independence assumption may be less plausible in locations experiencing higher than normal case counts and because one may think that α^0 could be substantially different in 2021 than in 2020 due to seasonal trends in influenza-like illnesses. However, in practice, we recommend that researchers with access to context-specific data use our method with a local estimate of λ.

Across location and time, our hidden-infection-method provides consistent estimates of the viral prevalence $\mathbb{P}[I_{st}]$. In Indiana from April 25^{th}–29^{th} 2020, Menachemi et al. (2020) 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 was 23.1% (data from covidtracking.com). Using our likelihood ratio of 16.35, we obtain a median latent prevalence estimate of 1.80% over that time period, again close to the actual randomized testing estimate and well within the 95% confidence interval. In Utah, from March 3^{rd} to March 13^{th} 2021, we estimated prevalence in two hot spots in Utah: one in Utah county and one in Salt Lake county. In Utah County, we estimate a prevalence of 0.69% using the hidden-infection-method and 0.74% (95% confidence interval 0.24% to 1.25%) using our randomized testing. In Salt Lake County, we estimate a prevalence of 0.57% using the hidden-infection-method and 0.47% (95% confidence interval 0% to 1.00%) using our randomized testing. Figure 1 shows the point estimates and confidence intervals for prevalence from randomized testing. It should be noted that our latent prevalence estimates fall within the confidence intervals from randomized testing and are often quite close the exact point estimate. Said differently, in practice, the estimate approximates random testing well even in different time periods and locations and in hot spots with higher reported cases and potentially different testing regimes.

4.3. Test of Conditional Independence

The previous section validated our hidden-infection-method using out-of-sample predictions. This alleviates some concern of the strength of Assumptions A and B in practice. In addition, we

report in panel B of Table 2 that we do not find evidence in the randomized testing data against the independence condition in Assumption A. We report estimates and confidence intervals for μ, which should equal to one if the conditional independence assumption holds (Equation 13). We report these estimates using different sets of variables that predict being at risk of COVID-19 infection (Columns 1–4 of panel B of Table 2). We fail to reject the null of conditional independence in all four columns. However, we also note that confidence intervals might be considered wide, which potentially indicates insufficient power. In the case that Assumption A is considered too strong, researchers may replace it with the weaker Assumption C, and our hidden-infection-method estimates become estimates for a lower bound for $\mathbb{P}[I_{st}]$.

4.4. Estimates Across All States

In this section, we report our estimates of latent prevalence for the rest of the United States as of July 1^{st}, 2020 in Table 3. These estimates are reported in the first column of Table 3, with the 95% confidence interval in the second column, based on bootstrapped estimates of the likelihood ratio λ. Details of the procedure are given in Appendix A. The third column reports the current reported positive testing rate that is used in our hidden-infection method to calculate $\mathbb{P}[I_{st}]$ in the first column. This table also provides an estimate of reported prevalence from publicly available data, that is, $(O_{st} - F_{st} - C_{st})/N_{st}$, where O_{st} is the number of confirmed cases in state s at time t , F_{st} is total fatalities, C_{st} is recovered cases, and N_{st} is the population. Compared to our method, this reported prevalence estimate suffers from selection bias. It also requires tracking confirmed cases until recovery or a fatality, which is often incomplete and hard to measure.

One of the advantages of our hidden-infection method is that it relies only on publicly available data on positivity—which is easily measurable. In contrast, data on total fatalities and recovered cases are relatively poor quality because of incomplete tracking. Several states do not report these numbers or report questionable numbers (e.g., California, Florida, and Massachusetts). For states that do not report recoveries, we impute recovered cases as the 21 day lag of new cases minus

fatalities for all states.[10]

The last column in Table 3 shows that latent prevalence is, on average, 2.89 times higher than reported prevalence numbers. In Figure 2, panel (A), we show this difference using a scatter plot of our estimated latent and reported prevalence for July 1^{st} 2020. If reported prevalence truly captures all latent prevalence, then the two measures should line up around the provided 45-degree line. Instead, most data points are above the 45-degree line, indicating that our estimated latent prevalence is substantially higher than the 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) uses 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.

Benatia et al. (2020) use the pooled estimation of a sample selection model across states to estimate latent prevalence. They report ratios of estimated cases to confirmed cases of 9.7 for Utah and 16.2 for Indiana for April 12. Almost all states have ratios of latent to confirmed infections of over ten and are substantially higher than estimates either from our hidden infection method or from actual randomized testing.

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, for that same date, our hidden-infection method implies a latent prevalence of 3.04% or a ratio of latent to reported prevalence of 3.5.

[10]This calculation provides similar numbers for states that report recoveries and, in some cases, are exactly those numbers.

4.5. 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 in Figure 2, panel (B). This figure graphs reported and latent prevalence on two different axes to highlight their co-movements. 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 the latent prevalence rate from the hidden-infection method and the reported prevalence depend on different data inputs. Our estimated latent prevalence relies on publicly reported positivity, which accounts for changes in testing availability. Therefore, changes in positivity are more likely to reflect disease spread than changes in testing. In contrast, reported prevalence is impacted by testing, recovery reporting, and fatality reporting, all of which introduce their own sample selection biases, which can change over time. Fourth, our estimated latent prevalence generally leads reported prevalence. For example, the latent prevalence in Utah peaks on June 25^{th}, almost a month before the reported prevalence peak on July 24^{th}. The lag in reported prevalence is most likely driven by reporting delays or imputations in fatalities and recoveries.

An important limitation for any user of our hidden-infection method to be aware of is that it will yield poor approximations if testing is highly rationed. For example, if only people with information on likely exposure to COVID-19 through contact tracing are tested. For example, several states exhibited values of $\mathbb{P}(I_{st}|\tau_{st}) = 1$ 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 with known exposure to COVID-19.

5. Conclusion

This paper provides a method to measure the latent prevalence of COVID-19, correcting for sample selection in symptom-based testing data and incomplete tracking of recovered cases and fatalities. 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.

Our methodology demonstrates surprising out-of-sample generalizability. First, we calibrate and validate the model using randomized testing in Utah from May to June 2020. Second, we show it predicts Indiana's active prevalence in April 2020 and Utah's active prevalence in March 2021. Since conditions on the ground, such as testing regime and epidemiological environment, can differ, we also provide boundary conditions for the applications of different variants of our method, either by itself or in combination with randomized testing. We hope our method can be a useful starting point to track COVID-19 and other future potential outbreaks in real-time. Future extensions of our approach include (i) integrating it to an economic model that predicts the impact of policies such as stay-at-home orders; and relaxing the invariance assumption on λ by allowing it to vary as a function of covariates.

References

Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer, "lassopack: Model Selection and Prediction with Regularized Regression in Stata," *The Stata Journal*, 2020, *20* (1), 176–235.

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.

Benatia, David, Raphael Godefroy, and Joshua Lewis, "Estimating COVID-19 Prevalence in the United States: A Sample Selection Model Approach," *medRxiv preprint, https://www.medrxiv.org/content/10.1101/2020.04.20.20072942v1*, 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.

Chatterjee, Arindam and Soumendra Nath Lahiri, "Bootstrapping Lasso Estimators," *Journal of the American Statistical Association*, 2011, *106* (494), 608–625.

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.

Favero, Nathan, "Adjusting confirmed COVID-19 case counts for testing volume," *medRxiv preprint, https://www.medrxiv.org/content/10.1101/2020.06.26.20141135v1*, 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.

Fisman, David and Ashleigh Tuite, "Simple Accurate Regression-Based Forecasting of Intensive Care Unit Admissions due to COVID-19 in Ontario, Canada," *medRxiv preprint, https://www.medrxiv.org/content/10.1101/2020.11.16.20231399v1.full.pdf*, 2020.

Gaulin, Maclean, Nathan Seegert, and Mu-Jeung Yang, "Doing Good rather than Doing Well: What Stimulates Personal Data Sharing and why?," *Working Paper, University of Utah*, 2021.

Grewelle, Richard and Giulio Leo, "Estimating the Global Infection Fatality Rate of COVID-19," *medRxiv preprint, https://www.medRxiv.org/content/10.1101/2020.05.11.20098780v1*, 2020.

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

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.

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.

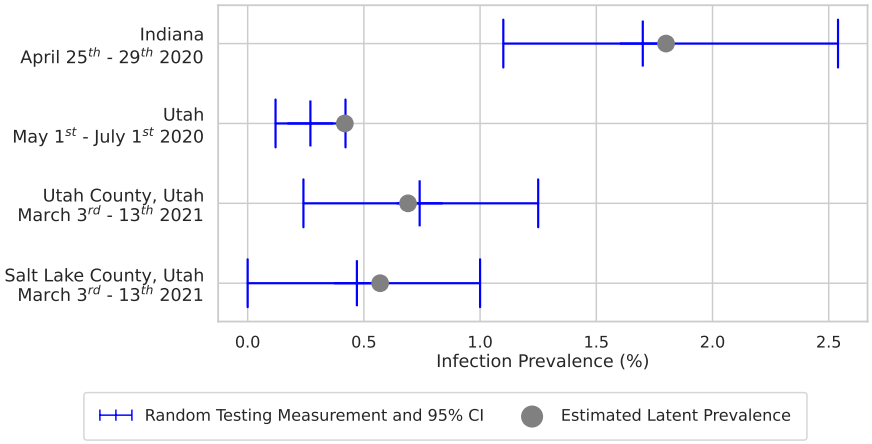
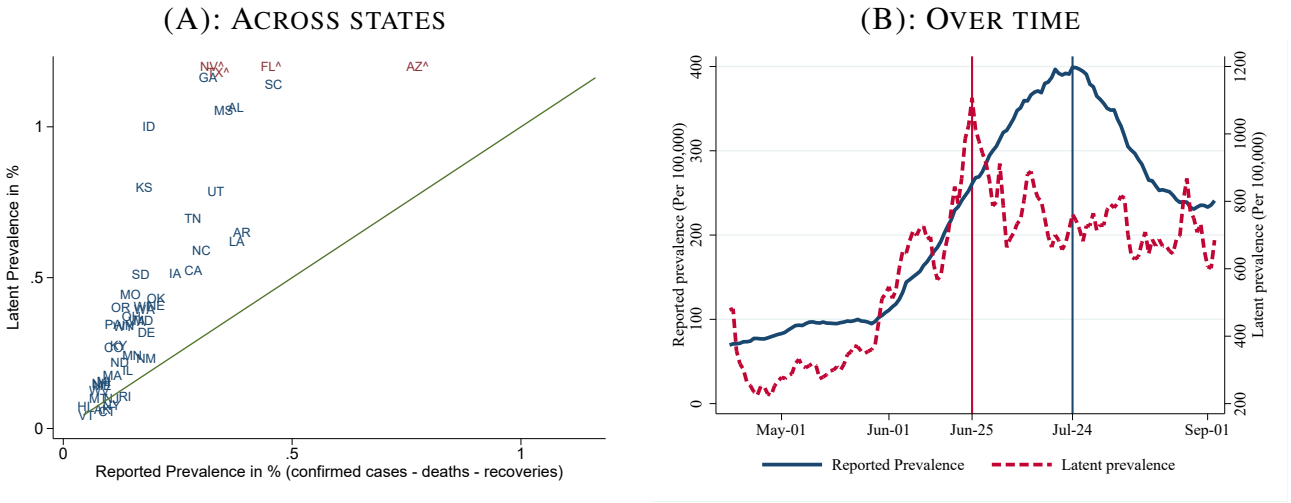


Figure 1: Random testing data are plotted with their calculated 95% confidence intervals in blue, with the latent prevalence calculated estimate shown with a grey circle. Data on positivity rates to calculate the latent prevalence is from the COVID tracking project.

Figure 2: Latent and reported prevalence



Panel (A): Latent prevalence is calculated using equation (9). Reported prevalence is calculated as $(O_{st} - F_{st} - C_{st})/(N_{st})$, where O_{st} are confirmed cases in state s at time t, F_{st} are total fatalities, C_{st} are recovered cases, and N_{st} is population. States with ˆ and red text are displayed lower on the Y-axis then their true values for convenience. Data is from the COVID tracking project. The estimates and data are displayed for July 1^{st} 2020.

Panel (B): Time path of latent and reported prevalence in Utah. Latent prevalence is defined as the fraction of currently infected in the state. Reported prevalence is calculated as $(O_{st} - F_{st} - C_{st})/(N_{st})$, where O_{st} are confirmed cases in state s (i.e., Utah) at time t, F_{st} are total fatalities, C_{st} are recovered cases, and N_{st} is population. Note that the two prevalences are plotted on different scales for ease of functional comparison. 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 Size	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

In panel A, we report estimates of α^0, α^1, and $\lambda = \alpha^1/\alpha^0$, using our experimental data from Utah from May to July 2020. The estimates for viral and antibody prevalence include corrections for nonresponse and sampling design (see Samore et al. (2020)). All columns use LASSO to select variables that predict being infected using different sets of variables, as described in Sections 2.2 and 3.2. Bootstrapped standard errors and 95 percent confidence intervals are reported in parentheses and square brackets, respectively, following the procedure described in Appendix A. In Column 1, we report estimates allowing the LASSO to choose from all clinical symptom variables in our data, including *anosmia*, fever, cough, shortness of breath, chills, muscle pain, headache, and sore throat. In Column 2, we report estimates allowing LASSO to choose from all symptomatic variables, as well as nonsymptomatic variables, e.g., a binary variable for working outside of the home. We give the full list of variables in Section 3.2. Column 3 includes up to three-way interactions of symptomatic variables, and Column 4 allows for three-way interactions of symptomatic and nonsymptomatic variables. In panel B, we report estimates for μ as means of assessing the independence condition in Assumption A (Equation 13). All estimates are weighted for sampling weights to account for nonresponse and sampling design. Estimates without weighting are reported in Table C.1 (Appendix C).

	Variables to Predict Infection			
	Level		Three-way interactions	
	Symptom	Symptom and nonsymptom	Symptom	Symptom and nonsymptom
	(1)	(2)	(3)	(4)
	A: Parameter estimates			
α^1	0.057	0.057	0.057	0.288
	(0.001)	(0.004)	(0.005)	(0.023)
	[0.054,0.059]	[0.049,0.065]	[0.047,0.067]	[0.243,0.333]
α^0	0.004	0.004	0.004	0.017
	(0.001)	(0.004)	(0.001)	(0.014)
	[0.002,0.005]	[0.000,0.011]	[0.001,0.006]	[0.000,0.044]
λ	16.354	16.501	16.354	16.582
	(0.058)	(0.022)	(0.461)	(0.109)
	[16.241,16.468]	[16.457,16.544]	[15.450,17.259]	[16.369,16.795]
	B: Test of conditional independence			
μ	1.022	1.091	0.612	0.546
	(0.515)	(0.432)	(0.537)	(0.529)
	[0.013,2.032]	[0.245,1.937]	[0.000,1.665]	[0.000,1.584]

Table 3: Latent vs Reported Prevalence

This table presents state level estimates of our model parameters. Positive rate is the fraction of tests reported that are positive for COVID-19 from the COVID tracking project. $\mathbb{P}(I_{st})$ is latent prevalence from equation (9) in state s at time t. 95% CI is the 95 percent confidence interval calculated based on the most conservative CI for λ calculated in Table 2 column (3). Cases-deaths, is calculated as the number of confirmed cases minus fatalities as a fraction of state population. Rep. Prev. is baseline reported prevalence, calculated as number of confirmed cases minus fatalities and recoveries as a fraction of state population. The ratio in the last column is the ratio of estimated latent prevalence to the baseline reported prevalence. Estimates are calibrated on July 1^{st}, 2020.

State	$\mathbb{P}(I_{st})$ (1)	95% CI (2)	Positive rate (3)	Cases - deaths (4)	Rep. Prev. (5)	Ratio (6)
			A: Utah			
UT	**0.79%**	**[0.75, 0.83]**	**11.48%**	**0.69%**	**0.30%**	**2.62**
			B: All other states			
AK	0.06%	[0.06, 0.07]	1.04%	0.13%	0.05%	1.22
AL	1.06%	[1.01, 1.12]	14.95%	0.77%	0.35%	3.08
AR	0.65%	[0.62, 0.69]	9.67%	0.69%	0.36%	1.83
AZ	2.35%	[2.23, 2.48]	28.22%	1.12%	0.74%	3.19
CA	0.52%	[0.50, 0.55]	7.92%	0.58%	0.25%	2.09
CO	0.27%	[0.26, 0.28]	4.23%	0.53%	0.07%	3.60
CT	0.05%	[0.05, 0.06]	0.89%	1.19%	0.06%	0.88
DE	0.32%	[0.30, 0.34]	4.97%	1.12%	0.15%	2.15
FL	1.35%	[1.28, 1.42]	18.24%	0.71%	0.42%	3.23
GA	1.16%	[1.10, 1.23]	16.14%	0.76%	0.28%	4.13
HI	0.07%	[0.07, 0.08]	1.20%	0.06%	0.02%	4.45
IA	0.51%	[0.49, 0.54]	7.80%	0.90%	0.22%	2.37
ID	1.00%	[0.95, 1.06]	14.18%	0.33%	0.16%	6.30
IL	0.19%	[0.18, 0.20]	3.07%	1.09%	0.11%	1.68
IN	0.34%	[0.33, 0.36]	5.33%	0.64%	0.11%	3.04
KS	0.80%	[0.76, 0.84]	11.63%	0.51%	0.14%	5.56
KY	0.27%	[0.26, 0.29]	4.29%	0.34%	0.09%	3.11
LA	0.62%	[0.59, 0.66]	9.27%	1.23%	0.35%	1.79
MA	0.18%	[0.17, 0.19]	2.82%	1.45%	0.07%	2.47
MD	0.36%	[0.34, 0.38]	5.56%	1.06%	0.14%	2.58
ME	0.14%	[0.13, 0.15]	2.27%	0.24%	0.05%	2.90
MI	0.16%	[0.15, 0.16]	2.48%	0.65%	0.06%	2.64
MN	0.24%	[0.23, 0.26]	3.81%	0.62%	0.11%	2.11
MO	0.44%	[0.42, 0.47]	6.80%	0.34%	0.11%	4.07
MS	1.06%	[1.00, 1.12]	14.85%	0.90%	0.32%	3.35
MT	0.10%	[0.09, 0.11]	1.61%	0.09%	0.04%	2.39
NC	0.59%	[0.56, 0.62]	8.85%	0.61%	0.27%	2.21
ND	0.22%	[0.21, 0.23]	3.44%	0.46%	0.09%	2.46
NE	0.41%	[0.38, 0.43]	6.25%	0.97%	0.17%	2.41
NH	0.15%	[0.14, 0.16]	2.38%	0.39%	0.05%	3.13
NJ	0.10%	[0.09, 0.10]	1.58%	1.76%	0.07%	1.33
NM	0.23%	[0.22, 0.25]	3.67%	0.56%	0.15%	1.60
NV	1.29%	[1.22, 1.36]	17.57%	0.59%	0.28%	4.52
NY	0.07%	[0.07, 0.08]	1.21%	1.90%	0.07%	1.04
OH	0.37%	[0.35, 0.39]	5.73%	0.43%	0.11%	3.28
OK	0.43%	[0.41, 0.46]	6.63%	0.35%	0.17%	2.58
OR	0.40%	[0.38, 0.42]	6.15%	0.20%	0.09%	4.43
PA	0.34%	[0.33, 0.36]	5.34%	0.63%	0.08%	4.51
RI	0.10%	[0.10, 0.11]	1.68%	1.52%	0.11%	0.98
SC	1.14%	[1.08, 1.20]	15.85%	0.71%	0.43%	2.68
SD	0.51%	[0.48, 0.54]	7.75%	0.75%	0.14%	3.78
TN	0.70%	[0.66, 0.74]	10.28%	0.65%	0.25%	2.78
TX	1.28%	[1.22, 1.36]	17.53%	0.56%	0.30%	4.28
VA	0.35%	[0.34, 0.37]	5.49%	0.71%	0.13%	2.77
VT	0.04%	[0.04, 0.05]	0.71%	0.18%	0.02%	2.40
WA	0.39%	[0.37, 0.42]	6.06%	0.46%	0.14%	2.80
WI	0.41%	[0.38, 0.43]	6.24%	0.54%	0.14%	2.91
WV	0.13%	[0.12, 0.13]	2.01%	0.16%	0.04%	2.99
WY	0.34%	[0.32, 0.36]	5.27%	0.26%	0.09%	3.60
AVERAGE	**0.50%**	[0.45, 0.50]	**7.25%**	**0.68%**	**0.17%**	**2.89**

Appendix: For Online Publication Only

Appendix A. Standard Errors and Confidence Intervals

The first goal of this section is to describe a procedure to obtain valid standard errors for the estimators $(\widehat{\lambda},\widehat{\mu})$ and a valid confidence set for (λ,μ). The estimators are described in Section 2.2 and applied to the data in Section 4. The second goal of this section is to extend this procedure for inference on $\mathbb{P}[I_{st}]$ using Equation 9.

We start with two main ingredients. First, $\Phi\left(\frac{x-p}{h}\right)$ is approximately equal to $\mathbb{I}\{x>p\}$ as $h\downarrow 0$, where Φ is the CDF of a standard Gaussian distribution. Second, the data in Section 3 is such that X is a vector of discrete random variables. These two ingredients allow us to write (λ,μ) as a differentiable function of β and moments of the data on (X,D^I,D^τ).

To see that, let $\{x_1,\ldots,x_P\}$ be the discrete support of the random vector X and choose a small $h>0$. Define

$$W=\left[\mathbb{I}\{X=x_1\},\ldots,\mathbb{I}\{X=x_P\}\right], \tag{A.1}$$

$$F(\beta)=\begin{bmatrix}\Phi\left(\frac{x_1\beta-p}{h}\right)\\ \vdots\\ \Phi\left(\frac{x_P\beta-p}{h}\right)\end{bmatrix}. \tag{A.2}$$

It turns out that $\mathbb{I}\{X\beta>p\}$ is approximately equal to $WF(\beta)$ for small h. We use $h=0.1$ in our routine.

Define the vector of parameters θ as follows.

$$\theta=\begin{bmatrix}\beta\\ \mathbb{E}[D^I]\\ \mathbb{E}[W']\\ \mathbb{E}[D^IW']\\ \mathbb{E}[D^\tau W']\\ \mathbb{E}\left[D^ID^\tau W'\right]\end{bmatrix},$$

where W' denotes the transpose of the W vector. Following the definitions in Section 2.2,

$$\begin{bmatrix}\lambda\\ \mu\end{bmatrix}=\begin{bmatrix}\frac{\mathbb{E}[D^\sigma D^I](1-\mathbb{E}[D^I])}{\mathbb{E}[D^I](\mathbb{E}[D^\sigma]-\mathbb{E}[D^\sigma D^I])}\\ \\ \frac{\mathbb{E}[D^ID^\tau D^\sigma]\mathbb{E}[D^\sigma]}{\mathbb{E}[D^ID^\sigma]\mathbb{E}[D^\tau D^\sigma]}\end{bmatrix}=\begin{bmatrix}\frac{\{\mathbb{E}[D^IW]F(\beta)\}(1-\mathbb{E}[D^I])}{\mathbb{E}[D^I](\{\mathbb{E}[W]F(\beta)\}-\{\mathbb{E}[D^IW]F(\beta)\})}\\ \\ \frac{\{\mathbb{E}[D^ID^\tau W]F(\beta)\}\{\mathbb{E}[W]F(\beta)\}}{\{\mathbb{E}[D^IW]F(\beta)\}\{\mathbb{E}[D^\tau W]F(\beta)\}}\end{bmatrix}\doteq G(\theta), \tag{A.3}$$

where $G(\theta)$ is a differentiable function of θ.

The next step is to construct an estimate for the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta}-\theta\right)$ using the residual bootstrap. This naturally gives an estimate for the asymptotic distribution of $\sqrt{n}\left(G(\widehat{\theta})-G(\theta)\right)$ via the Delta method. Chatterjee and Lahiri (2011) describe a residual bootstrap procedure that is consistent for the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$, where $\widehat{\beta}$ is computed

using LASSO. They also demonstrate that the moments of the bootstrap distribution are consistent for their relevant counterparts. The bootstrap procedure that we propose below is a straightforward modification of their procedure, because the non-β components of θ are sample averages. The steps of the procedure are described below.

Step 1. Construct an estimate $\widehat{\theta}$ for θ. For the β components of θ, use LASSO with penalization parameter λ_n that satisfies $\lambda_n/\sqrt{n} \to \lambda_0 \in [0,\infty)$. That gives $\widehat{\beta}$. The remaining components in $\widehat{\theta}$ are just sample averages: $\overline{D^I}$, $\overline{W}$, $\overline{D^I W}$, $\overline{D^\tau W}$, and $\overline{D^I D^\tau W}$;

Step 2. Construct a vector of residuals $\widetilde{\varepsilon}_i$. First, we need to regularize the LASSO estimate $\widehat{\beta}$ according to Chatterjee and Lahiri (2011). To do that, choose a sequence $a_n = cn^{-\delta}$ with $c \in (0,\infty)$ and $\delta \in (0,0.5)$ and construct $\widetilde{\beta}$ by making $\widetilde{\beta}_j = \widehat{\beta}_j \mathbb{I}\{|\widehat{\beta}_j| \geq a_n\}$ for every $j = 1,\ldots,K$. In practice, Chatterjee and Lahiri (2011) find that $a_n \in \{0.0125, 0.05, 0.125, 0.25\}$ work well in simulations with $n = 250$, X_j drawn from a standard Gaussian, and $|\beta_j| \leq 6$. For $\delta = 0.25$, these values of a_n correspond to $c \in \{0.05, 0.2, 0.5, 0.99\}$ in the rule $a_n = cn^{-\delta}$. We use $a_n = 0.05$ for our results in section 4, but find that our results are consistent across $a_n \in \{0.05, 0.1, 0.2\}$. Take $\widehat{\theta}$ from the previous step, replace $\widehat{\beta}$ with $\widetilde{\beta}$, and call the resulting vector $\widetilde{\theta}$. The vector of residuals $\widetilde{\varepsilon}_i$ is

$$\widetilde{\varepsilon}_i = \begin{bmatrix} D_i^I - X_i\widetilde{\beta} \\ D_i^I - \overline{D^I} \\ W_i' - \overline{W}' \\ D_i^I W_i' - \overline{D^I W}' \\ D_i^\tau W_i' - \overline{D^\tau W}' \\ D_i^I D_i^\tau W_i' - \overline{D^I D^\tau W}' \end{bmatrix}, \tag{A.4}$$

for $i = 1,\ldots,n$;

Step 3. Simulate bootstrap residuals ε_i^*. Define $\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \widetilde{\varepsilon}_i$. Construct a sample of residuals $\{\varepsilon_1^*,\ldots,\varepsilon_n^*\}$ by drawing n observations randomly and uniformly with replacement from the set of centered residuals $\{\widetilde{\varepsilon}_1 - \overline{\varepsilon},\ldots,\widetilde{\varepsilon}_n - \overline{\varepsilon}\}$;

Step 4. Simulate bootstrap data. Take the sample of bootstrap residuals of the previous step and construct the bootstrap sample of data,

$$\begin{bmatrix} D_i^{I*} \\ D_i^{I**} \\ W_i^{*\prime} \\ D^I W_i^{*\prime} \\ D^\tau W_i^{*\prime} \\ D^I D^\tau W_i^{*\prime} \end{bmatrix} = \begin{bmatrix} X_i\widetilde{\beta} \\ \overline{D^I} \\ \overline{W}' \\ \overline{D^I W}' \\ \overline{D^\tau W}' \\ \overline{D^I D^\tau W}' \end{bmatrix} + \varepsilon_i^*, \tag{A.5}$$

for $i = 1,\ldots,n$;

Step 5. Compute bootstrap estimates θ^*. Apply the same LASSO estimator of Step 1 to the sample of D_i^{I*} and X_i to compute β^*. Take sample averages of D_i^{I**}, W_i^*, $D^I W_i^*$, $D^\tau W_i^*$, and $D^I D^\tau W_i^*$. Call

these sample averages $\overline{D^I}^*$, $\overline{W}^*$, $\overline{D^I W}^*$, $\overline{D^\tau W}^*$, and $\overline{D^I D^\tau W}^*$, respectively. Stack all these estimates in the vector θ^*;

Step 6. Compute the bootstrap distribution. Repeat Steps 3–5 B times, e.g., $B = 1{,}000$. For each repetition, record bootstrap estimate θ_b^*, $b = 1, \ldots, B$. Compute $T_{n,b}^* = \sqrt{n}\left(G(\theta_b^*) - G(\widetilde{\theta})\right)$ for each b. The bootstrap distribution of T_n^* is approximately equal to the empirical distribution of $T_{n,b}^*$ over $b = 1, \ldots, B$, given the original sample. This empirical distribution is a consistent estimator for the distribution of $\sqrt{n}\left(G(\widehat{\theta}) - G(\theta)\right)$, which interests us;

Step 7. Compute standard errors. Compute the variance-covariance matrix of the bootstrap distribution of T_n^* from the previous step. This is a consistent estimator for the asymptotic variance-covariance matrix of $\sqrt{n}\left(G(\widehat{\theta}) - G(\theta)\right)$. Take this matrix, divide by n, and take the square root. The elements of the main diagonal are valid standard errors for $(\widehat{\lambda}, \widehat{\mu})$;

Step 8. Compute confidence set. Let $1-\alpha$ be the desired asymptotic confidence level. Compute the $1-\alpha$ quantile of the distribution of $\|T_n^*\|$, where $\|\cdot\|$ is the Euclidean norm. Call that $1-\alpha$ quantile $\widehat{t_n}(1-\alpha)$. The joint confidence set for (λ, μ) is given by

$$I_{n,1-\alpha} = \left\{ g \in \mathbb{R}^3 \ : \ \|g - G(\widehat{\theta})\| \leq n^{-1/2} \widehat{t_n}(1-\alpha) \right\}.$$

If the researcher only desires a confidence set for one individual parameter, e.g., λ, simply repeat the procedure above for the first coordinate of T_n^*.

We now consider inference on $\mathbb{P}[I_{st}]$. The researcher first obtains $\widehat{\lambda}$ using experimental data in a certain region s_0 and time period t_0. For a different region s or time period t, the researcher obtains a publicly reported positivity rate, that is, $\mathbb{P}[I_{st}|\tau_{st}]$. Combining the two numbers with Equation 9 gives an estimate for $\mathbb{P}[I_{st}]$,

$$\widehat{\mathbb{P}}[I_{st}] = \frac{\mathbb{P}[I_{st}|\tau_{st}]}{\widehat{\lambda}(1 - \mathbb{P}[I_{st}|\tau_{st}]) + \mathbb{P}[I_{st}|\tau_{st}]}.$$

Note that, given $\mathbb{P}[I_{st}|\tau_{st}]$, $\widehat{\mathbb{P}}[I_{st}]$ a decreasing function of $\widehat{\lambda}$. Thus, if the $1-\alpha$ confidence interval for λ is $[\widehat{\lambda}_l, \widehat{\lambda}_u]$, then a $1-\alpha$ confidence interval for $\mathbb{P}[I_{st}]$ can be constructed by

$$\left[\frac{\mathbb{P}[I_{st}|\tau_{st}]}{\widehat{\lambda}_u(1 - \mathbb{P}[I_{st}|\tau_{st}]) + \mathbb{P}[I_{st}|\tau_{st}]} \ ; \ \frac{\mathbb{P}[I_{st}|\tau_{st}]}{\widehat{\lambda}_l(1 - \mathbb{P}[I_{st}|\tau_{st}]) + \mathbb{P}[I_{st}|\tau_{st}]} \right].$$

Appendix B. Accounting for seasonal variation in α^0

In the main text, we assumed that the likelihood ratio $\lambda = \frac{\alpha^1}{\alpha^0}$ is constant across time and space. This assumption might be considered especially problematic for $\alpha_{st}^0 = \mathbb{P}[\sigma_{st}|I_{st}^c]$, which is the

fraction of symptomatic individuals among uninfected. In fact, clinical symptoms, such as cough and runny nose, might exhibit strong seasonal variation due to the flu season.

One way to accomplish such an adjustment is to rely on data on "influenza-like illnesses" (ILI) from the CDC, which are defined as exhibiting "fever plus cough or sore throat" and no other known causes of these symptoms, other than influenza. Specifically, the CDC publishes weekly data on the ratio of outpatient visits that exhibit ILI symptoms during the flu season, relative to the non-flu-season baseline. We denote this ILI ratio as r_t^{ili}.

To utilize such ILI data, one will need to make the following additional assumption:

Assumption D. *Time variation in the probability of being symptomatic for uninfected persons* α_t^0 *is proportional to variation in the ILI ratio:* $\alpha_t^0 = \alpha^0 \cdot r_t^{ili}$

With this assumption, one can use ILI ratios from the CDC for the 2020-2021 flu season to adjust our latent COVID-19 prevalence measures, see www.cdc.gov/flu/weekly/weeklyarchives2020-2021/data/senAllregt11.html. This data measures the percentage of outpatient visits due to ILI, during the flu season, relative to the non-flu season, denoted by r_t^{ili}. We use this adjustment to re-calculate our latent COVID-19 prevalence measures and create the time series in Figure B.3 below.

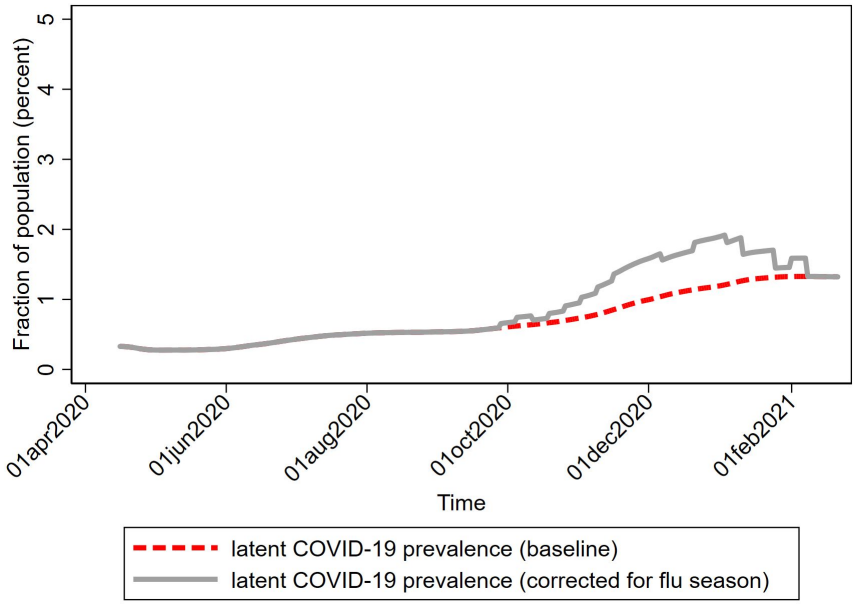


Figure B.3: Time series of latent COVID-19 prevalence in Utah, from April 2020 to February 2021. The red dashed line displays our baseline latent COVID-19 prevalence $\mathbb{P}(I_{st})$, using a constant α^0. The grey line uses adjustments for the 2020-2021 flu season, using $\alpha_t^0 = \alpha^0 \cdot r_t^{ili}$, where r_t^{ili} is the ratio of influenza-related illness (ILI) outpatient visits during a specific week in the 2020-2021 flu season, relative to visits for ILI during non-flu season months.

We draw two conclusions from this figure. First, our method can relatively easily be adjusted to take account for time variation of influenza-like illnesses during the flu season, if one is willing to use Assumption D. Second, even without adjustments for the flu season, our COVID-19 prevalence

estimate provides a lower bound to the latent prevalence of COVID-19 after adjustment for the flu season.

Appendix C. Additional Estimates

This appendix produces unweighted estimates and estimates from two subsamples of recruitment from our field study.

Table C.1: Unweighted Key parameters from Randomized Testing

In panel A, we report estimates of α^0, α^1, and $\lambda = \alpha^1/\alpha^0$, using our experimental data from Utah from May to July 2020. All columns use LASSO to select variables that predict being infected using different sets of variables to select. In Column 1, we report estimates allowing the LASSO to choose from all symptomatic variables in our data, such as fever and anosmia. In Column 2, we report estimates allowing LASSO to choose from all symptomatic and nonsymptomatic variables, such as working out side of the home. Columns 3 and 4 include interactions up to three-way interactions of symptomatic and symptomatic and nonsymptomatic variables, respectively. All estimates are unweighted for sampling and estimates with weighting are reported in 2. Bootstrapped standard errors and 95 percent confidence intervals are reported in parentheses and square brackets following the procedure in Appendix A.

	Symptoms			
	Level		Three-way interactions	
	Symptom	Symptom and nonsymptom	Symptom	Symptom and nonsymptom
	(1)	(2)	(3)	(4)
	A: Parameter estimates (unweighted)			
α^1	0.054	0.056	0.054	0.307
	(0.001)	(0.004)	(0.005)	(0.023)
	[0.052,0.057]	[0.048,0.064]	[0.044,0.065]	[0.262,0.352]
α^0	0.004	0.004	0.004	0.019
	(0.001)	(0.004)	(0.001)	(0.014)
	[0.002,0.006]	[0.000,0.011]	[0.002,0.006]	[0.000,0.046]
λ	13.845	14.321	13.845	16.168
	(0.058)	(0.022)	(0.461)	(0.109)
	[13.732,13.959]	[14.277,14.364]	[12.941,14.749]	[15.955,16.382]

Table C.2: Key parameters from Randomized Testing - In Person and Letter Recruitment Separately

This table replicates Column 1 of Table 2, with the full sample (column 1), the in person recruitment (Column 2) and the letter recruitment (Column 3) samples. In panel A, we report estimates of α^0, α^1, and $\lambda = \alpha^1/\alpha^0$, using our experimental data from Utah from May to July 2020. All columns use LASSO to select variables that predict being infected using all symptomatic variables in our data, such as fever and anosmia. In panel B, we report a test of our main assumption A, given in equation (7). All estimates are weighted for sampling. Bootstrapped standard errors and 95 percent confidence intervals are reported in parentheses and square brackets following the procedure in Appendix A.

	Symptoms		
	Level		
	All Obs.	In Person	Letter Recruitment
	(1)	(2)	(3)
	A: Parameter estimates (sampling weighted)		
α^1	0.057	0.085	0.055
	(0.001)	(0.002)	(0.002)
	[0.054,0.059]	[0.082,0.088]	[0.051,0.060]
α^0	0.004	0.004	0.003
	(0.001)	(0.001)	(0.002)
	[0.002,0.005]	[0.002,0.007]	[0.000,0.007]
λ	16.354	20.333	16.039
	(0.058)	(0.100)	(0.058)
	[16.241,16.468]	[20.137,20.529]	[15.925,16.152]
	B: Test of conditional independence		
μ	1.022	1.143	0.625
	(0.515)	(0.691)	(0.871)
	[0.013,2.032]	[0.000,2.496]	[0.000,2.332]