Estimating the Fraction of Unreported Infections in Epidemics with a Known Epicenter: an Application to COVID-19*

Ali Hortaçsu† Jiarui Liu‡ Timothy Schwieg§

April 12, 2020

Abstract

We develop a simple analytical method to estimate the fraction of unreported infections in epidemics with a known epicenter and estimate the number of unreported COVID-19 infections in the US during the first half of March 2020. Our method utilizes the covariation in initial reported infections across US regions and the number of travelers to these regions from the epicenter, along with the results of a randomized testing study in Iceland. We estimate that 4-14% (1.5%-10%) of actual infections had been reported in US up to March 16, accounting for an assumed reporting lag of 8 (5) days.

*We thank the Becker Friedman Institute for financial support. We also thank Fernando Alvarez, Susan Athey, Patrick Bayer, Rana Choi, Liran Einav, Jeremy Fox, Mikhail Golosov, Austan Goolsbee, Philip Haile, Jakub Kastl, Magne Mogstad, Casey Mulligan, Derek Neal, Robert Shimer, Jose Scheinkman, Chad Syverson, Harald Uhlig, Theodore Vassilakis, and Alessandra Voena for their helpful comments.

†Kenneth C. Griffin Department of Economics, University of Chicago and NBER

‡Kenneth C. Griffin Department of Economics, University of Chicago

§Becker Friedman Institute

1 Introduction

The global pandemic COVID-19 is here in the United States. The number of confirmed cases is rising rapidly, reaching 398,809 as of April 7 with 12,895 reported deaths. The coronavirus outbreak was declared a national emergency beginning March 1[1]. More than half of U.S. states have imposed various levels of lockdown measures[2]. In addition to the public health crisis, the country is certainly looking at a deep and possibly long-lasting economic recession, according to Ben Bernanke and Janet Yellen in a recent Financial Times article[3].

Given the level of severity of current conditions, we still fail to answer the most basic yet important question: How many people are actually infected with COVID-19 in the U.S. and what is the true fatality rate? Because of the shortage in testing kits, hospitals and disease control centers are only able to test the subsample of people with severe symptoms or travel history. The number of reported infections is much lower than the actual number of infections in the U.S.

These unreported infections can be unrecognized because they often experience mild or no symptoms (Nishiura et al., 2020; Andrei, 2020). If not hospitalized or quarantined, they can infect a large proportion of the population. Thus, estimating the number of unreported infections can inform policy-makers about the proper scale of virus control policies (Alvarez et al., 2020; Eichenbaum et al., 2020), and to assess the effectiveness of public health policies such as social distancing in slowing the spread of the epidemic.

Estimating the number of unreported infections may also give a more accurate measure of the true fatality rate. The current reported fatality rate, which is 3-4% according to WHO[4], is not the true fatality rate. The true fatality rate is the proportion of those *actually infected* who die, not of those *reportedly infected*. The reported fatality rate is a biased estimate of the true rate, because there is selection bias in testing. Since many of the patients who are tested have severe symptoms, they may have a higher true fatality rate than those untested, which would make the reported fatality rate an overestimate of the true rate.

Ideally, a randomized testing experiment will give an unbiased estimate of the true rate. However, given the limited supply of testing kits and surging demand by people with symptoms, randomized testing may be infeasible, especially in the early periods of the outbreak. Therefore, it may be of great value to estimate the fraction of unreported

[1] https://www.whitehouse.gov/presidential-actions/proclamation-declaring-national-emergency-concerning-novel-coronavirus-disease-covid-19-outbreak/

[2] https://www.wsj.com/articles/a-state-by-state-guide-to-coronavirus-lockdowns-11584749351

[3] https://www.ft.com/content/01f267a2-686c-11ea-a3c9-1fe6fedcca75

[4] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_2

infections with observational data at hand. With that knowledge, policy-makers will be better equipped to assess the proper level and duration of virus control policies.

In this paper we develop a simple analytical method to estimate the fraction of unreported infections for situations where the epidemic has a known epicenter. Our methodological strategy, described in Section 3, exploits the covariation between the number of initial reported infections in locations away from the epicenter, and the number of travelers from the epicenter to these locations.

To illustrate the idea, consider a time period when the epicenter is the only location with infections, and that the only way another city/country can be infected is through travelers. Also assume, as in Section 3.2, that any infected travelers can only come from the unreported infected population in the epicenter – an assumption we find reasonable (as reported infected individuals would not be allowed to travel), but are able to relax in Section 3.3. Suppose now the hypothetical situation where we know the reporting rate of infections in the epicenter (the fraction of reported infections to the true number of infections), and that we know the number of travelers from the epicenter to another city/country. Assuming travelers resemble the population of the epicenter, we can calculate the expected number of infected (but unreported) travelers entering other cities/countries. Assuming further that we know the rate of transmission of the disease, we can then calculate the expected number of infections these travelers will have generated in these locations. Comparing the expected number of infections that arise from travelers to reported cases of the infection, we can estimate the reporting rate.

What can we do in the realistic case if the reporting rate in the epicenter is unknown? In Section 3.2, we propose the following: suppose we make the assumption that the reporting rate at the epicenter and the previously uninfected city/country are the same (or a known function of each other). We can then start with a guess on the unknown rate of reporting at the epicenter, which allows us to calculate the implied reporting rate at the previously uninfected city/country, and check whether these are equal (or satisfy the known function). If not, we update our guess, and try again. In other words, we can solve for the reporting rate(s) balancing the expected number of infections from travel and the number of infected that are being reported in both locations.

While the above strategy, outlined in Section 3.2, is in principle implementable, it is crucially dependent on the assumption that reporting rates are the same across the epicenter and destination locations (or a known function of each other). Moreover, its results are very sensitive to knowing the transmission rate of the infection from travelers, as this allows us to project the number of infections in the destination city/country. Suppose now that we have access to the reporting rate of infections from another des-

tination city/country, e.g. through universal or randomized testing, as has been done in Iceland.[5] This allows us to estimate how infectious the travelers from the epicenter are. Assuming that this transmission rate from travelers is the same (or a known function of) as the transmission rate at the destination city/country of interest, we can then calculate the expected number of infections we would expect from travel. Intuitively, the ratio between number of travelers to two destination cities/countries from the epicenter should tell us the ratio of total infections between the two cities/countries. Randomized or universal testing at one of the destinations, Iceland in our case, will give us its number of total infections, so total infections at the other destination can be computed. This strategy is discussed in detail in Section 3.3.

We would like to be very upfront that the estimation strategies outlined above are dependent on strong assumptions and reliable data on travel patterns, and that any results are *very sensitive to these assumptions*. However, our hope is that our approach is clear in terms of its assumptions and its corresponding limitations; we hope that future research can improve upon these limitations. We have attempted to account for some of the limitations. For example, in Section 3.4, we discuss how to correct for the fact that infections are often reported with a delay, as there is a delay to the outset of symptoms that are often a prerequisite for testing for the infection, as well as a delay in laboratory testing.

Our data consists of detailed daily reported infections for all U.S. states/city/county and Iceland collected by Johns Hopkins University of Medicine Coronavirus Resource Center from January 22 to March 31, 2020; international travel data to U.S. in January and February 2020 from I-94 travels data by National Travel and Tourism Office; international travel data to Iceland by Icelandic Tourist Board in January and February 2020.

Our model generates a range of estimates that depend on the traveler data that is incorporated, the date range considered, and assumptions regarding the lags associated with reported case data. We report this range of estimates in Table 3. Across these estimates, we find that 4% to 14% of cases were reported across the U.S. up to March 16, when social distancing measures began to be applied in major metropolitan areas and travel declined significantly (Thompson et al., 2020) .[6] This suggests that for each case reported in late February/early March, between 6 to 24 cases remained unreported (after accounting for an 8 day reporting lag). Once again, our estimates

[5]Of course, another strategy is to assume that the reporting rate discovered through randomized testing in Iceland is the same in the destination city/country of interest (Section 3.1).

[6]This assumes that cases are reported with a lag of 8 days as in Table 3(b), and incorporates travel data from China, Italy, Spain, Germany, and the UK. A shorter assumed reporting lag of e.g. 5 days generates a range of estimated reporting rates between 1.5% to 10%. We have excluded King county, Washington in these results because this county containing Seattle shows much earlier community infections than other regions in U.S.

are highly dependent on model assumptions, and the data that is used to inform it. We discuss how our results depend on these assumptions in some detail in Section 5.

In the economic literature, Berger et al. (2020) and Stock (2020) study the importance of unreported cases in the context of the coronavirus pandemic. Our paper contributes to the growing literature in epidemiology on estimating the true number of infections using observational data and structural model assumptions. Notably, (Li et al., 2020; Wu et al., 2020; Flaxman et al., 2020; Liu et al., 2020a,b; Nishiura et al., 2020) utilize simulated epidemiological models to estimate the fraction of unreported infections in China and European countries. As Zhao et al. (2020) notes, it is often difficult to identify the fraction of unreported alongside the growth of the infection purely by measures of fit. Our paper complements these extant papers: we provide what we believe is a transparent identification argument and a very light computational strategy that allows researchers to assess the sensitivity of model estimates to modeling and data assumptions. That said, our model may miss important components of disease dynamics that these more sophisticated epidemiological models incorporate. These richer models may also allow one to estimate a richer set of model parameters than we have been able to.[7] Another related recent paper is Imai et al. (2020), who estimate potential total cases in Wuhan China from the confirmed cases in other countries due to international travel, assuming that all cases outside of China are reported correctly [8]. Korolev (2020) discusses non-identification in SEIRD models and proposes estimation strategy conditional on knowing infectious period and incubation period.

Section 2 introduces our model of infection, which describes the early stages of the dynamics of the epidemic. Section 3 presents our two estimation/identification strategies. Section 4 describes the data we are using for estimation. Section 5 lays out the estimation results and our robustness checks.

2 Model

Our model is based on the classic SIR model in epidemiology. We consider the evolution of the virus in both the epicenter c and into target city i over a period of time $T_0 \leq t \leq T_1$. We are considering a relative short period of time in the early stage of the

[7] For example, that Li et al. (2020) estimates different transmission rates for reported vs. unreported infections, which we are unable to identify with our strategy. Li et al. (2020) assume that unreported infected individuals transmit the disease at a slower rate than reported infected individuals. However, since most reported infections are either hospitalized or self-quarantined, it is not clear whether this assumption is an *a priori* reasonable one.

[8] Bogoch et al. (2020) and Lai et al. (2020) calculates how vulnerable countries are to the virus by the magnitude of travelers from Wuhan, and correlate these vulnerability/risk measures with reported cases in these countries.

epidemics. Thus, the "recovered" population at the epicenter, which is a small fraction of the population, is assumed not to play a significant role during this period.

2.1 What happens at the epicenter c

We denote Infected, Reported Infected, and Unreported Infected in time t and epicenter c as $I_{c,t}, R_{c,t}, U_{c,t}$ respectively.

The epicenter starts with some initial infections $I_{c,0}$. We are considering a short period of time in between T_0 and T_1, so the number of susceptibles at the epicenter remain relatively constant throughout this period. There are also no infected cases traveling into epicenter. We assume no recovery. So at time t, the total infections at epicenter with transmission rate β is given by

$$I_{c,t} = I_{c,0} \exp(\beta(t - T_0)) \tag{1}$$

It is worth noting that this β term can include the spread minus recoveries, since we do not model a changing number of susceptibles. It should be viewed as the net spread of infections over time.

Each time t, there is a cohort of travelers $M_{i,t}$ going from epicenter to target city and potentially bringing the virus to target city.

2.2 What happens in target city i

We denote Infected, observed Reported Infected, and Unreported Infected in time t and city i as $I_{i,t}, R_{i,t}, U_{i,t}$ respectively. At period T_0, target city i has zero infections, so $I_{i,T_0} = R_{i,T_0} = U_{i,T_0} = 0$.

Each time $t \in [T_0, T_1]$, target city receives a cohort t of incoming travelers $M_{i,t}$ from the epicenter. Among these travelers, $I^{inc}_{i,t}$ are infected. Each cohort of incoming infected $I^{inc}_{i,t}$ will transmit the virus in target city with rate β for the period of $[t, T_1]$. We assume that the transmission rate at target city is the same as in epicenter. Thus, at period T_1, this cohort will infect $I^{inc}_{i,t} \exp(\beta(T_1 - t))$ people in the city i. The total new infections at target city at T_1 caused by all cohorts of incoming infected travelers will be

$$I_{i,T_1} = \int_{T_0}^{T_1} I^{inc}_{i,t} \exp(\beta(T_1 - t)) dt \tag{2}$$

Let α be fraction of reported case over new infections across periods, so $\alpha = \frac{R_{i,T_1} - R_{i,T_0}}{I_{i,T_1} - I_{i,T_0}}$. Since $I_{i,T_0} = R_{i,T_0} = 0$, we can write $\alpha = \frac{R_{i,T_1}}{I_{i,T_1}}$. In reality at target

city i, we only observe $R_{i,t}$ with some iid measurement error $\epsilon_{i,t}$[9]. Let $\hat{R}_{i,t}$ denote the observed reported cases of city i time t. We have

$$\hat{R}_{i,T_1} = \alpha I_{i,T_1} + \epsilon_{i,T_1} = \alpha \int_{T_0}^{T_1} I_{i,t}^{inc} \exp(\beta(T_1 - t))dt + \epsilon_{i,T_1} \tag{3}$$

3 Estimating the Reporting Rate α

The estimation/identification question is: can we recover α, the reporting rate, when we only observe reported infections $\hat{R}_{i,T_1}$ but not $I_{i,t}^{inc}$, the total incoming infected in equation (3)? In the following Sections 3.1-3.3, we provide a complete treatment of how one can recover α under different scenarios of data availability. We consider two sets of data that could potentially be available: (i) data on travel from epicenter to U.S., and (ii) data from a randomized testing implemented outside of U.S. In Section 3.4, we extend our model and estimation strategy to incorporate reporting lags.

3.1 Travel data unavailable but randomized testing data available

If randomized testing data from some other country is available, the true infection rate in that country can be estimated. Therefore, given the population and number of reported infections, the fraction of reported infections *in that country* can be estimated. If we are willing to believe that the fraction of reported infections in that country is the same as in the U.S. due to similar testing availability or medical systems, then α is trivially recovered. However, in many cases, this assumption is unlikely to hold. In the next sections, we will show how we can recover α relaxing this assumption.

3.2 Travel data available but randomized testing data unavailable

When only travel data is available, we need the assumption that people capable of traveling away from the epicenter would be the uninfected and the unreported infected. This is a reasonable assumption especially in the case of COVID-19 because the great majority of reported infected individuals would be quarantined and not allowed to travel.

Our main assumption in this scenario is:

[9]This measurement error $\epsilon_{i,t}$ could arise from error in collecting the data.

Assumption 3.1.

$$\frac{I^{inc}_{i,t}}{M_{i,t}} = \frac{U_{c,t}}{N_c - \hat{R}_{c,t}} \quad \textit{for any time } t \in [T_0, T_1], \textit{ city } i \textit{ and epicenter } c \tag{4}$$

N_c is the population of epicenter. $\hat{R}_{c,t}$ is the observed reported infections at epicenter c time t, defined analogously to $\hat{R}_{i,t}$ for target city i. In other words, we are assuming that the fraction of unreported infections among incoming travelers from the epicenter is the same as the fraction of unreported infections among people capable of leaving the epicenter. (We will relax this assumption in Section 3.3.) Since α remains constant,[10] we have $U_{c,t} = (1-\alpha)I_{c,t} = \frac{1-\alpha}{\alpha}R_{c,t}$. Therefore, assumption 3.1 becomes:

$$\frac{I^{inc}_{i,t}}{M_{i,t}} = \frac{(1-\alpha)I_{c,t}}{N_c - \hat{R}_{c,t}} \quad \forall t \in [T_0, T_1], \text{city } i, \text{epicenter } c \tag{5}$$

Plugging equation (1) in, we get

$$I^{inc}_{i,t} = \frac{(1-\alpha)I_{c,0}\exp(\beta(t-T_0))}{N_c - \hat{R}_{c,t}} M_{i,t} \quad \forall t \in [T_0, T_1], \text{city } i, \text{epicenter } c \tag{6}$$

Plugging back to equation (3), we get

$$\hat{R}_{i,T_1} = \alpha \int_{T_0}^{T_1} \frac{(1-\alpha)I_{c,0}\exp(\beta(t-T_0))}{N_c - \hat{R}_{c,t}} M_{i,t}\exp(\beta(T_1-t))dt + \epsilon_{i,T_1} \tag{7}$$

$$= \alpha(1-\alpha)I_{c,0}\exp(\beta(T_1-T_0)) \int_{T_0}^{T_1} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1} \tag{8}$$

$$= \alpha(1-\alpha)\frac{R_{c,0}}{\alpha}\exp(\beta(T_1-T_0)) \int_{T_0}^{T_1} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1} \tag{9}$$

$$= (1-\alpha)R_{c,0}\exp(\beta(T_1-T_0)) \int_{T_0}^{T_1} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1} \tag{10}$$

This equation allows us to solve for $(1-\alpha)\exp(\beta(T_1-T_0)$ if we observe $R_{c,0}$. We will allow for observing $R_{c,0}$ with error in Section 3.3. We can estimate β from the growth of reported infections in the epicenter because there is no influx of infected people from other regions. Given that β is now determined, we can solve for α. However, there is much variation in estimation of β within the literature, and our estimate of α varies

[10] In reality α may be varying over time, due to e.g. changes in the extensiveness of testing. If this is the case, as we vary the $[T_0, T_1]$ window, we will obtain window-specific estimates of α, which can be thought of as a weighted average of α during this period.

with point estimates of β (Liu et al., 2020; Read et al., 2020; Shen et al., 2020).

3.3 When travel data and a random testing benchmark are available

In this scenario, we will be leveraging the same fact that the number of incoming unreported infections is informed by the travelers from the epicenter. Now we can also allow for selection in traveling. More specifically, if we think that e.g. urban areas are likely to have a higher infection rate than rural areas and travel abroad more[11], then assumption 3.1 might not hold. Therefore, we introduce a bias correction term γ in the relation between the fraction of infected among travelers and the fraction of unreported infected individuals in the general population. This bias correction term γ can also account for the fact that a fraction of the unreported infected people might be too sick to travel. Our relaxed assumption in this scenario is:

Assumption 3.2.

$$\frac{I_{i,t}^{inc}}{M_{i,t}} = \gamma \frac{U_{c,t}}{N_c - \hat{R}_{c,t}} \quad \textit{for any time } t \in [T_0, T_1], \textit{ city } i \textit{ and epicenter } c, \gamma \neq 0 \tag{11}$$

We can further allow for the fact that the reporting rate in epicenter α_c can be different from that of region i, so $U_{c,t} = (1-\alpha_c)I_{c,t} = \frac{1-\alpha_c}{\alpha_c} R_{c,t}$. We can now rewrite assumption 3.2 as:

$$\frac{I_{i,t}^{inc}}{M_{i,t}} = \gamma \frac{(1-\alpha_c)I_{c,t}}{N_c - \hat{R}_{c,t}} \quad \forall t \in [T_0, T_1], \gamma \neq 0, \text{city } i, \text{epicenter } c \tag{12}$$

Plugging into equation (1) and (3), we get

$$\hat{R}_{i,T_1} = \alpha \frac{1-\alpha_c}{\alpha_c} \gamma \exp(\beta(T_1 - T_0)) R_{c,0} \int_{T_0}^{T_1} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1} \tag{13}$$

The additional parameters for the bias correction term γ and different reporting rate for the epicenter complicate the estimation of α using travel data alone. However, having data from a country that has done randomized or complete testing greatly helps overcome this challenge. In our case, we are able to identify α using additional information given by the randomized testing benchmark provided by Iceland. Since the Iceland company deCODE genetics implemented random testing of COVID-19 for

[11] This is likely to be true in epicenters like China.

a representative sample of the island population[12], we are able to observe true infection rate of Iceland at time T_1. Multiplying this true infection rate by the population of region j in Iceland will give us the actual number of infections in region j at time T_1, which is I_{j,T_1}. Thus, for any region j in Iceland we observe $I_{j,T_1} - I_{j,T_0}$, which in turn, equals the infections generated by travelers from the epicenter:

$$I_{j,T_1} - I_{j,T_0} = \frac{1-\alpha_c}{\alpha_c}\gamma \exp(\beta(T_1 - T_0))R_{c,0} \int_{T_0}^{T_1} \frac{M_{j,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_j \tag{14}$$

Note that if we allow for iid measurement error ϵ_j in observing $I_{j,T_1} - I_{j,T_0}$, we can get a consistent estimate of $\frac{1-\alpha_c}{\alpha_c}\gamma \exp(\beta(T_1 - T_0))R_{c,0}$ by estimating equation (14). If we don't allow for measurement error ϵ_j, then we can estimate $\frac{1-\alpha_c}{\alpha_c}\gamma \exp(\beta(T_1 - T_0))R_{c,0}$ with no error. Estimating equation (13) gives consistent estimate of $\alpha\frac{1-\alpha_c}{\alpha_c}\gamma \exp(\beta(T_1 - T_0))R_{c,0}$. Taking the ratio, we have identified α.

One intuition for this strategy is the following: the ratio between travel to U.S. and travel to Iceland from the epicenter should tell us the ratio of total infections between U.S. and Iceland. Iceland's randomized testing gives us its number of total infections, so U.S. total infections can be computed. In other words, we observe the outcome in U.S. with under-reporting, and the unobserved counterfactual outcome with full reporting is given by the benchmark Iceland. An additional advantage of this estimation/identification strategy, as opposed to the previous strategy in section 3.2, is that now we don't need an estimate of β in order to recover α. We also allow for the fact that $R_{c,0}$ could be observed with error. Identifying α does not require observing $R_{c,0}$ perfectly because $R_{c,0}$ appears identically in both equations.

We should be clear that for terms with β to cancel out, the argument does assume that β is the same across Iceland and the U.S. We believe this might be a reasonable assumption for the early periods of the infection when social distancing or other widespread measures had not yet been implemented (in a potentially differential fashion). We also need the bias term γ to be the same for US and Iceland; this means that proportion of (unreported) infected travelers from China to the U.S. and Iceland are the same. More detailed micro-data on travelers may be used to assess the validity of this assumption.

This estimation strategy also works when a complete testing benchmark exists. If the whole population of region j is tested, then we observe $I_{j,T_1} - I_{j,T_0}$ trivially. Equation (14) still gives consistent estimate of $\frac{1-\alpha_c}{\alpha_c}\gamma \exp(\beta(T_1 - T_0))R_{c,0}$ and the rest of the argument follows.

Note that if in the model reported infections and unreported infections have different transmission rates, then our strategy would not be able to capture the differential rates.

[12]We will describe the randomized testing in detail in Section 4

We would need other sources of information to help us pin down these differential rates.

3.4 Incorporating Reporting Lags

In this section, we show how our model can incorporate a fixed reporting lag in reported infections and derive identification equations. Reporting lags are important, because if people are tested for the virus only after symptoms show up, there will be a lag in reported infections. Another major reason for reporting lag is the lag in testing results. The turnaround time for testing results in U.S. major laboratory companies could be 2 to 3 days (Kaplan and Thomas, 2020).

We denote true infected, true reported infected, and true unreported infected in time t and target city i as $I_{i,t}, R_{i,t}, U_{i,t}$ respectively. Those for epicenter c as $I_{c,t}, R_{c,t}, U_{c,t}$. Let k be the lagged report period. At time t city i denote the lagged reported infected $LR_{i,t} = R_{i,t-k}$. For epicenter c, lagged reported infected is $LR_{c,t} = R_{c,t-k}$. But the observed lagged reported infected of target city $\widehat{LR}_{i,t}$ is with iid measurement error $\epsilon_{i,t}$.

Define reporting rate at city i as $\alpha = \frac{R_{i,t-k}}{I_{i,t-k}} = \frac{LR_{i,t}}{I_{i,t-k}}$ and at epicenter c as $\alpha_c = \frac{R_{c,t-k}}{I_{c,t-k}} = \frac{LR_{c,t}}{I_{c,t-k}}$. This means that we are considering the reporting rate of lagged reported cases as a fraction of the lagged total infections.

When travel data are available but randomized testing data unavailable, we still maintain assumption 3.1. In city i time T_1, the estimating equation is

$$\widehat{LR}_{i,T_1} = \alpha \frac{1-\alpha_c}{\alpha_c} \exp(\beta(T_1 - T_0 - k)) LR_{c,k} \int_{T_0}^{T_1-k} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1} \quad (15)$$

If $\alpha = \alpha_c$, then both of them are identified if β and k are identified and $LR_{c,k}$ is observed without error.

When both travel data and randomized testing data are available, we maintain assumption 3.2. In US city i time T_1, the estimating equation is

$$\widehat{LR}_{i,T_1} = \alpha\gamma \frac{1-\alpha_c}{\alpha_c} \exp(\beta(T_1 - T_0 - k)) LR_{c,k} \int_{T_0}^{T_1-k} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1} \quad (16)$$

In Iceland region j time T_1, the estimating equation is

$$\hat{I}_{j,T_1-k} = \gamma \frac{1-\alpha_c}{\alpha_c} \exp(\beta(T_1 - T_0 - k)) LR_{c,k} \int_{T_0}^{T_1-k} \frac{M_{j,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{j,T_1} \quad (17)$$

If we know k, then we can compute $\hat{I}_{j,T_1-k}$ from the randomized testing data. Regressing equation (17) gives consistent estimate of $\gamma\frac{1-\alpha_c}{\alpha_c}\exp(\beta(T_1-T_0-k)LR_{c,k}$. Regressing equation (16) gives a consistent estimate of $\alpha\gamma\frac{1-\alpha_c}{\alpha_c}\exp(\beta(T_1-T_0-k))LR_{c,k}$. Taking the ratio, we can identify α if we know k. Again here we can also allow for the situation where we don't observe $LR_{c,k}$ perfectly. Details of how we derive the estimating equations are in the appendix.

4 Data

4.1 COVID-19 Data

Daily reported infections and recovery data are collected by Johns Hopkins University of Medicine Coronavirus Resource Center from January 22 to March 31, 2020. We use data for all U.S. states/counties and Iceland.

Randomized testing data in Iceland is obtained from the website maintained by the Directorate of Health and the Department of Civil Protection and Emergency Management in Iceland[13]. We have daily number of tests conducted by deCODE genetics and daily number of confirmed cases. We use the first wave of randomized testing by deCODE which spans March 15 - 19, 2020. During the first wave they performed 5490 tests and confirmed 48 cases, which implies an infection rate of .874%. The randomized testing conducted was random over non-confirmed individuals which made up a very tiny portion of the Icelandic population at this time. This could lead to slightly downward biases in our Iceland confirmed data, slightly biasing our US alpha estimates upward.

In our main estimation we consider February 23 as T_0 and March 10 as T_1 with a 5 day lag, and T_1 as Mar 13 for an 8 day lag. This is because there were very few infections in January and early February. We check robustness of different time periods in Section 5.3.

4.2 Travel Data

We obtain monthly data of international arrivals to U.S. by port of entry and country of origin from I-94 Arrivals by National Travel and Tourism Office. We use the number of visitors from China, Italy, Spain, UK, and Germany in January and February 2020 as the measure for incoming travelers to U.S. states. For international arrivals to Iceland, we get the number of visitors from China, Italy, Spain, UK, and Germany in January

[13] https://www.covid.is/data

and February 2020 from the Icelandic Tourist Board. We have not been able to obtain March travel data into either country.

The National Travel and tourism office of the United States provides monthly data for entry by port of entry, as well as a separate data set for country of origin. We construct the number of visitors from China, Italy, Spain, Germany, and the UK by scaling the port-of-entry data by the percentage of total visitors that are from these countries. This introduces error, as we cannot observe directly the number of e.g. Chinese travelers into a particular city or state. It is also important to note that we do not observe inter-state travel. While this may not be important for the immediate infections caused by travelers from the epicenter, our projections for the number of infections for T_1 that are far removed from T_0 will be less accurate due to interstate travel.

For Icelandic data, 99% of international travelers arrive through Keflavik airport into Iceland. The data contains a breakdown of arrival by country of origin, broken down by month of arrival. We use January and February arrival data from China, Italy, Spain, UK, and Germany for estimation.

Our travel data for both countries does not control for connecting flights. However the United States data is limited to the top-30 port of entries, many of which are large urban cities for which there will be less connecting flights. Further work that can obtain more precise estimates of entry may be able to control for this. In the case of Iceland: a survey conducted by the Icelandic Tourist Board suggests that 2-5% of international travelers are aboard connecting flights, suggesting it is less of a problem for this set of data.

4.3 Population Data

Estimates of U.S. State and county population data come from the U.S. Census Bureau. Data for the populations of China, Iceland, Italy, Spain, UK, and Germany as of 2020 are obtained from the United Nations Population Division.

5 Empirical Application

5.1 Implementation

We now consider estimation of α_{US} using randomized sampling in Iceland, as described in Section 3.3. Randomized sampling done by deCODE genetics gives a percentage of the population that has contracted the virus. We estimate equation (14) using Randomized Testing to construct I_{j,T_1-k}. We do not have city-level travel data into

Iceland. 99% of all international travel arrives through a single airport, and while the data provided is accurate, this gives only a single data point for estimation. As a result, equation (14) relies on a single data point of travel and infection, but $\exp(\beta(T_1 - T_0))\gamma\frac{1-\alpha_c}{\alpha_c}R_{c,0}$ is estimated without error by the ratio of $I_{j,T_1} - I_{j,T_0}$ and $\int_{T_0}^{T_1}\frac{M_{j,t}}{N_c - \hat{R}_{c,t}}dt$.

For estimation of reporting rates in the U.S., we need estimates of several figures: Firstly $M_{j,t}$, and secondly of $I_{j,T_1} - I_{j,T_0}$. We discuss the estimation of these here. We observe only monthly travel data to construct $M_{j,t}$, and to maintain robustness to January travels and infections, we average February and January travel into both the United States and Iceland. We assume that $M_{j,t}$ is uniform over the entire time period such that $\int_{Jan1}^{Feb29} M_{j,t}$ is equal to the sum of all travel into the city from January and February. Thus the integral $\int_{T_0}^{T_1}\frac{M_{j,t}}{N_c - \hat{R}_{c,t}}$ varies only due to the confirmed infections increasing over time. Estimation of I_{j,T_1} is complicated due to randomized testing by Iceland only being conducted at certain dates. To resolve this problem, we scale the Iceland randomized results by the scale of the confirmed cases against March 15. This means that if there were half the confirmed cases in March 5 as in March 15, the total infections would be half of the randomized testing percent times the population of Iceland. This allows for us to consider T_1 closer to the onset of the infection than the randomized testing dates. We also remove the number of infected from Wuhan China from our data on confirmed infected in China due to the lock-down restrictions placed on this city. We use the first wave of deCODE testing to determine the percentage of the population that has contracted the disease. This testing took place during Mar 15 through Mar 19. The results show that .874% of the population of Iceland have contracted the disease as of Mar 15[14].

We estimate equation (13) using multiple data points from U.S. states and counties. We obtain our estimate of $\alpha\exp(\beta(T_1 - T_0))\gamma\frac{1-\alpha_c}{\alpha_c}R_{c,0}$ via OLS without a constant term. One important note is that if the magnitude of measurement error in travel data were high, this problem may be alleviated via instrumental variables strategy using other travel data measured with error.

We then construct our estimate of α by dividing the two estimates. It is important to note that as a result of the division, this method is not reliant on population data from the epicenter of infection. As long as γ and β are the same between Iceland and the United States we will have identified α. It is likely that at the onset of the infection similar preventative measures have been taken in these two countries, meaning that β will be reasonably close for each country.

Is China the only epicenter for the United States? While the first confirmed infection in Seattle occurred from a visitor from China, our data on The United States and Iceland occurs later in the global progression of the virus than our Chinese data.

[14] Stock et al. (2020) also estimates the undetected rate and total infection rate in the Iceland study.

By the time these countries were experiencing infections, Italy had also experienced an outbreak. To this end, we also allow for a second epicenter: Italy. Italy is located much closer to Iceland and constitutes a substantial amount of travel to the country. However, to maintain identification, we require that α, β and T_0 be same for both China and Italy, and we observe $LR_{c,k}$ for both epicenters with no error. However we find that allowing T_0 to vary does not affect our estimates by much. We also consider a broader collection of epicenters of China, Italy, Spain, Germany and the UK. For some collection of epicenters L: Our estimation equation for the United States is given below.

$$\widehat{LR}_{i,T_1} = \alpha \frac{1-\alpha_c}{\alpha_c} \gamma \exp(\beta(T_1 - T_0 - k)) \left(\sum_{\ell \in L} \left[\int_{T_0}^{T_1-k} \frac{LR^{\ell}_{c,k} M^{\ell}_{i,t}}{N^{\ell}_c - \hat{R}^{\ell}_{c,t}} dt \right] \right) + \epsilon_{i,T_1} \quad (18)$$

A similar equation is also estimated with multiple epicenters (China and Italy, also with Spain, Germany, and UK) for Iceland.

5.2 Results: Illustration

As a first illustration of our approach, we first estimate α using February 23 as T_0 and March 10 as T_1. We consider other dates for robustness later. We consider two lag models, 5 days, the median time for symptoms to appear, and 8 days, to capture the testing lag in addition to symptom onset (Lauer et al. (2020); Kaplan and Thomas (2020); Li et al. (2020)). For the 8 day lag, T_1 was set to March 13 for comparison. We estimate, for the 8 day lag, a value of $\alpha = 0.0416$ (s.e. 0.00984).[15] This would mean that for every case confirmed in the United States in early March, there are still $\frac{1-\alpha}{\alpha} = 23$ unconfirmed cases (assuming a reporting lag of 8 days).

There is one city present in our data that is a huge outlier. Seattle featured very early infections, and was unable to contain the spread of early infections unlike other cities in the United States. We believe that for King County, T_0 may be much earlier than for the other cities. This means that within our time interval, there are substantial amounts of infections caused by residents of the city, not only visitors. As a result, this city has a substantially higher (3700%) amount of confirmed cases per visitor than any other city at the current time, and we exclude it from the data.

Correct estimation of the reporting lag parameter is essential, as our estimates of α are sensitive to this. We consider its robustness in the following section.

Our approach is also sensitive to the travel data magnitudes, which may not be

[15] We are reporting a naive OLS standard error here. We have not fully explored the sampling properties of this estimation method.

well estimated for the United States due to data limitations. In particular, connecting flights after port of entry may lead to underestimates of international arrivals into smaller cities and counties. We also lack inter-state travel between the United States, which would be important for estimating α later into the spread of the virus.

Have we considered all epicenters of the virus for the United States and Iceland? There were other countries which had seen substantial infections such as South Korea. Their exclusion biases both the estimates from both Iceland as well as the United States, and as long as the magnitudes of travel were even between the two will not bias alpha. If these other epicenters had more travel to the United States relative to Iceland United States than Iceland, this would downward bias our estimates of α, and vice versa. However, we see little change in our estimates by adding in Spain, Germany and UK. If the travel patterns between the United States and Iceland to and from an omitted set of epicenter countries are not very different, we do not believe their omission will substantially alter our results.

5.3 Results: Range of Estimates and Robustness Checks

Our dates for T_0 and T_1 are chosen such that they capture the onset of the infection for the United States. As table 3 shows, our α estimate is reasonably stable along choices of T_1, and very stable among choices of T_0 all throughout February. We estimate a range of $1.5\% - 10\%$ for the average reporting rate across the US with a reporting lag of 5 days and $4\% - 14\%$ reporting rates when there is a lag of 8 days. Using only China as the epicenter, we observe similar patterns in α. For early March we note a relatively stable α over T_1. For very early choices for T_1, our Iceland estimates of confirmed are very small, and this could create very noisy estimates of α (the first case in Iceland was confirmed February 28). As we increase T_1, we see an increase in α. This may be due to increases in the availability of test kits, which lead to higher reporting rates. However, this result may in part be due to unobserved/unaccounted travel, particularly within the United States, along with the fact that we do not have data on March travel into the US. Both of these factors would lead to under-reporting of travel for late March, and cause estimates of α to be upward biased. Moreover, as we progress later into March, social distancing/health policy measures across Iceland and U.S. began to be applied, leading to differential changes in the transmission rate.

How would these estimates affect our estimate of total infections, as opposed to reported infections? As an illustration, with our estimated average reporting rates from table 3, we compute the estimated total infections for different U.S. counties as of March 15 and March 20 in table 5. To give a range of our estimates, we take the 10th and 90th percentile of the estimated reporting rates over all cutoffs to compute

a bound on estimated total infections. As of March 20, our results suggest, for our average α estimate in table 3, that there were 41,205 (85,937) Infected residents of New York City, with a 10th percentile of 24,366 (47,257) and a 90th percentile of 114,849 (306,607) for an assumed 8 (5) day lag. These estimates imply that, as of March 20, with an assumed 8(5) day reporting lag, about 0.5% (1%) of New York City population was infected for our average estimate in table 2, and 1.4% (3.6%) at the 90th percentile of our estimates.

Throughout the analysis above, we have excluded King County, Washington which contains Seattle. Table 7 displays our estimates including this county, which heavily skews the data. We believe this may be due to significant community infections occurring in the county during our time period, as the city was infected much earlier than other cities.

We consider our estimates robustness to reporting lags in table 8. Our estimates of α appear reasonably robust to a range of lengths of the lag, with an increase as the lag becomes longer.

We note that large lags ($k > 10$) pose a problem for estimation in our model. For estimation purposes, we maintain $T_1 - k$ to be a constant date as we consider changes in the lag parameter. This means that for large lags, we must consider T_1 dates deep into March. However, the further we get into March, the more interstate travel and carrying of infections between cities and states matters, which may lead to overstating α. To complicate matters, Icelandic and the US policies for handling the spread of the virus may have diverged significantly. This means that our assumption of β constant between countries also may not hold. When it does not, our estimate of α identifies $\alpha \times \exp\left[\left(\beta_{US} - \beta_I\right)\left(T_1 - T_0\right)\right]$. $\beta_{US} > \beta_I$ implies that we are overestimating α for large T_1 values. Evidence from Kucharski et al. (2020) suggests that β is very sensitive to changes in policy, leading to this upward bias in α.

6 Conclusion

In this paper, we lay out a simple model of disease transmission across a known epicenter and target cities. Using this model, we provide simple analytical arguments to allow the estimation/identification of reporting rates in target cities away from the epicenter. Our preferred estimation strategy utilizes variation of travel patterns from epicenter to destination cities and available randomized testing results from elsewhere in the world. The empirical implementation of our model generates a range of estimates for the percentage of infections that have been reported. Using international travel data to the U.S. and randomized testing data from Iceland, for the February to early March window, our estimates of the average reporting rate in the U.S. lie in a

range of $4-14\%$, accounting for an assumed reporting lag of 8 days. (The range is $1.5-10\%$, accounting for a reporting lag of 5 days.) Our estimates suggest that a large number of infections in the U.S. have not been reported in this early period.

We should be very clear that we are not offering or endorsing any policy recommendations based on our estimates. Nor do we suggest that any of our analysis should be taken as a substitute for randomized/complete testing, which will provide the most reliable estimates of the true infection rate in the population. Our primary aim in this paper has been to obtain tractable analytic results showing how to identify the reporting rate from available data. We also note that our model is a substantially stripped down version of epidemiological models considered by (Li et al., 2020; Wu et al., 2020; Flaxman et al., 2020). These more complex models may allow additional sources of variation in the data to pin down the key parameters of interest. Importantly, we do want to emphasize that our identification and estimation results rely quite sensitively on model assumptions and the (un)availability of high quality data on travel. We hope future research can improve on these important limitations.

Figures and Tables

Table 1: Summary Statistics of Fraction of Reported Infections by County

Version	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
China Travel Only	0.001404	0.027434	0.037345	0.060896	0.100897	0.203500
China and Italian Travel	0.001254	0.025142	0.034784	0.055746	0.094906	0.183048
China and EU Travel	0.001397	0.024920	0.038928	0.051934	0.067098	0.193740

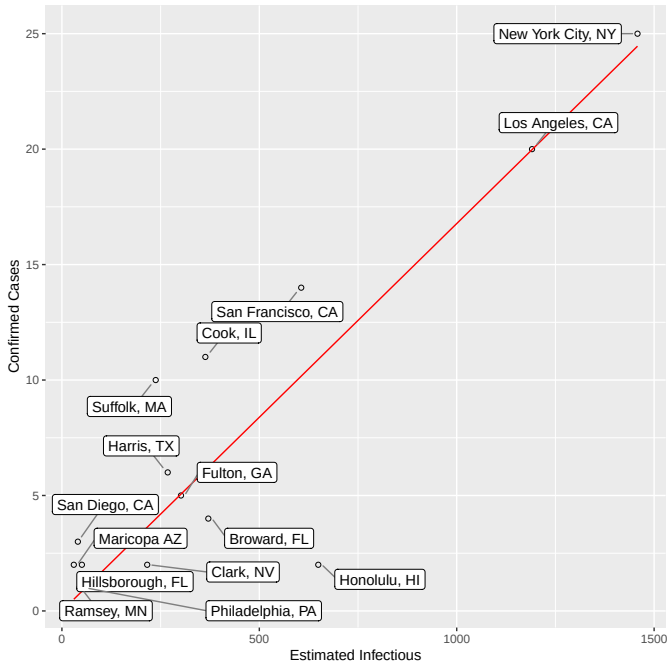
Summary statistics reported on the distribution of α for each county in the data. α is estimated for T_0 Feb 23, T_1 Mar 13, and a Lag of eight days. EU travel includes traveler data from Italy, Spain, UK, and Germany.

Table 2: Estimated average fraction of reported infections

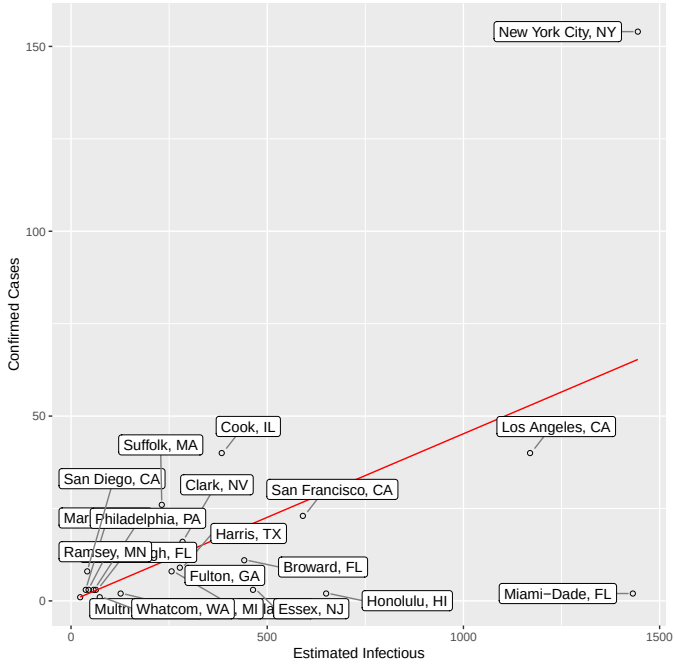
	α_{ice}	α_{US}	$\frac{1-\alpha_{US}}{\alpha_{US}}$
	5 Day Lag		
China and Italy Travel Data	0.0231	.0161	61.2
Only Chinese Travel Data	0.0231	0.0169	58.3
China, Italy, Spain, Germany, UK	0.0231	0.0168	58.6
	8 Day Lag		
China and Italy Travel Data	0.0231	0.0416	22.5
Only Chinese Travel Data	0.0231	0.0458	20.8
China, Italy, Spain, Germany, UK	0.0231	0.0452	21.1

We report estimated α by OLS without a constant for several specifications of the model. We use T_0 as Feb 23 and T_1 as March 10,13 for each Lag respectively. For the versions including European data, European travel to both Iceland and the United States is considered. King County, WA is omitted from the calculation.

Figure 1: Estimated reported infections by county



(a) 5 Day Lag



(b) 8 Day Lag

This plot shows the ratio of Confirmed Cases to Estimated Cases in each County on T_1 = March 10 for 5 day lag, and 13 for 8 day lag. T_0 is February 23rd. We use the full European Entry. Estimated Cases are computed as the ratio of the right hand side of equation 16 and the ratio of equation 17

Table 3: Mean Fraction of Unreported Infections with Different Cutoffs

		T_0 Date						
		Feb 1	Feb 5	Feb 10	Feb 15	Feb 20	Feb 25	Feb 29
T_1 Date	Mar 6	0.09759	0.09758	0.09756	0.09762	0.09767	0.09774	0.09767
	Mar 7	0.05279	0.0528	0.05282	0.05275	0.05266	0.05249	0.05196
	Mar 8	0.03083	0.03083	0.03084	0.03081	0.03076	0.03068	0.03047
	Mar 9	0.01611	0.01611	0.01612	0.0161	0.01609	0.01605	0.01598
	Mar 10	0.01683	0.01683	0.01684	0.01682	0.01679	0.01676	0.01668
	Mar 11	0.02283	0.02284	0.02285	0.02282	0.02278	0.02273	0.02263
	Mar 12	0.02043	0.02044	0.02046	0.02041	0.02037	0.02031	0.02022
	Mar 13	0.03101	0.03102	0.03104	0.03099	0.03093	0.03086	0.03075
	Mar 14	0.04247	0.04248	0.04252	0.04244	0.04236	0.04227	0.04212
	Mar 15	0.03682	0.03683	0.03686	0.0368	0.03675	0.03669	0.0366
	Mar 16	0.04809	0.0481	0.04815	0.04806	0.04798	0.0479	0.04777
	Mar 17	0.06773	0.06776	0.06784	0.06769	0.06756	0.06743	0.06723
	Mar 18	0.109	0.109	0.1092	0.1089	0.1086	0.1084	0.1081
	Mar 19	0.2479	0.248	0.2483	0.2477	0.2472	0.2466	0.2459

(a) 5 Day Reporting Lag

		T_0 Date						
		Feb 1	Feb 5	Feb 10	Feb 15	Feb 20	Feb 25	Feb 29
T_1 Date	Mar 9	0.1397	0.1397	0.1397	0.1395	0.1391	0.1382	0.1354
	Mar 10	0.09538	0.09542	0.0954	0.09517	0.09485	0.09422	0.09265
	Mar 11	0.08921	0.08927	0.08925	0.08898	0.08864	0.08804	0.08678
	Mar 12	0.03928	0.03931	0.0393	0.03915	0.03899	0.03875	0.03832
	Mar 13	0.0456	0.04563	0.04562	0.04548	0.04532	0.0451	0.04472
	Mar 14	0.05725	0.0573	0.05729	0.05708	0.05686	0.05658	0.05613
	Mar 15	0.05082	0.05086	0.05085	0.05071	0.05058	0.05039	0.0501
	Mar 16	0.08169	0.08178	0.08176	0.08147	0.08121	0.08088	0.0804
	Mar 17	0.1201	0.1203	0.1202	0.1197	0.1192	0.1186	0.1178
	Mar 18	0.211	0.2114	0.2113	0.2102	0.2092	0.2081	0.2066
	Mar 19	0.4528	0.4538	0.4536	0.4505	0.448	0.4452	0.4414

(b) 8 Day Reporting Lag

This table displays α value for different dates for both T_0 and T_1. We vary T_0 across the month of February, and T_1 across early March. Very early March and February dates are not available since Iceland confirmed infections only begin February 28. Travel data is assumed uniform across days throughout and is not weighted as T_0 or T_1 change. We include Italy, Spain, Germany, and the United Kingdom as epicenters as well as China.

Table 5: Estimated Total Infected By County - Lag 5

	Mar 15				Mar 20			
County	Rep.	10Pct.	Mean	90Pct.	Rep	10Pct.	Mean	90Pct.
Broward, FL	36	330	601	2143	128	1174	2135	7619
Clark, NV	16	147	267	952	126	1156	2102	7500
Cook, IL	50	459	834	2976	278	2550	4638	16548
Fulton, GA	20	183	334	1190	88	807	1468	5238
Harris, TX	11	101	184	655	30	275	501	1786
Hillsborough, FL	4	37	67	238	32	294	534	1905
Honolulu, HI	3	28	50	179	28	257	467	1667
Los Angeles, CA	53	486	884	3155	292	2679	4872	17381
Maricopa AZ	4	37	67	238	34	312	567	2024
New York City, NY	269	2468	4488	16012	5151	47257	85937	306607
Philadelphia, PA	4	37	67	238	67	615	1118	3988
Ramsey, MN	5	46	83	298	16	147	267	952
San Diego, CA	16	147	267	952	127	1165	2119	7560
San Francisco, CA	28	257	467	1667	76	697	1268	4524
Suffolk, MA	27	248	450	1607	86	789	1435	5119

Estimated Total Infected in each county at March 15 and March 20 For a 5 day lag in reporting. We compute the 90th and 10th percentiles based on the range of α given in Table 3. Total Infected is computed as $\frac{R_{c,t}}{\alpha}$. Since we are using the same range of α for each county, row entries with the same number of reported cases are identical. Harris County is missing Mar 20, so Mar 19 data is used for Harris.

Table 6: Estimated Total Infected By County - Lag 8

	Mar 15				Mar 20			
County	Rep	10Pct.	Mean	90Pct.	Rep	10Pct.	Mean	90Pct.
Broward, FL	36	170	288	803	128	605	1024	2854
Clark, NV	16	76	128	357	126	596	1008	2809
Cook, IL	50	237	400	1115	278	1315	2224	6198
Dallas, TX	11	52	88	245	74	350	592	1650
Essex, NJ	7	33	56	156	73	345	584	1628
Fulton, GA	20	95	160	446	88	416	704	1962
Harris, TX	11	52	88	245	30	142	240	669
Hillsborough, FL	4	19	32	89	32	151	256	713
Honolulu, HI	3	14	24	67	28	132	224	624
Los Angeles, CA	53	251	424	1182	292	1381	2336	6511
Maricopa AZ	4	19	32	89	34	161	272	758
Miami-Dade, FL	13	61	104	290	123	579	980	2731
Multnomah, OR	1	5	8	22	12	57	96	268
New York City, NY	269	1272	2152	5998	5151	24366	41205	114849
Philadelphia, PA	4	19	32	89	67	317	536	1494
Ramsey, MN	5	24	40	111	16	76	128	357
San Diego, CA	16	76	128	357	127	601	1016	2832
San Francisco, CA	28	132	224	624	76	360	608	1695
Suffolk, MA	27	128	216	602	86	407	688	1918
Wayne, MI	8	38	64	178	67	317	536	1494
Whatcom, WA	2	9	16	45	10	47	80	223

Estimated Total Infected in each county at March 15 and March 20 For an 8 day lag in reporting. We compute the 90th and 10th percentiles based on the range of α given in Table 3. Total Infected is computed as $\frac{R_{c,t}}{\alpha}$. Since we are using the same range of α for each county, row entries with the same number of reported cases are identical. Harris County is missing Mar 20, so Mar 19 data is used. Miami-Dade is missing Mar 19 and Mar 20, so the average between Mar 19 and Mar is used instead.

Table 7: Reporting Rate (α) Estimates Including King County

	α_{ice}	α_{US}	$\frac{1-\alpha_{US}}{\alpha_{US}}$
	5 Day Lag		
China and Italy Travel Data	0.0231	0.0206	47.6
Only Chinese Travel Data	0.0231	0.0211	46.5
China, Italy, Spain, Germany, UK	0.0231	0.0211	46.5
	8 Day Lag		
China and Italy Travel Data	0.0231	0.0516	18.4
Only Chinese Travel Data	0.0231	0.0539	17.6
China, Italy, Spain, Germany, UK	0.0231	0.0538	17.6

We report estimated α by OLS without a constant for several specifications of the model. We use T_0 as Feb 23rd and T_1 as Mar 10,13 for each lag respectively. For the versions including European data, European travel to both Iceland and the United States is considered. King County is included in this data.

Table 8: Robustness to Lag

Lag	α
0	0.00558
1	0.00874
2	0.0094
3	0.01
4	0.0123
5	0.0168
6	0.0287
7	0.0299
8	0.0452
9	0.0717
10	0.0758
11	0.124
12	0.211

Table 8 shows estimates of α as the reporting lag period is varied. We use T_0 as Feb 23, and T_1 as March 5 + Lag days. King County is omitted. We include Italy, Spain, Germany, and the United Kingdom as epicenters as well as China.s

Appendix

We derive our model incorporating reporting lags in the appendix and show how we get the estimating equations in Section 3.4.

Recall that we denote true infected, true reported Infected, and true unreported infected in time t and target city i as $I_{i,t}, R_{i,t}, U_{i,t}$ respectively. Those for epicenter c as $I_{c,t}, R_{c,t}, U_{c,t}$. Let k be the lagged report period. At time t city i denote the lagged reported infected $LR_{i,t} = R_{i,t-k}$. For epicenter c, the lagged reported infected is $LR_{c,t} = R_{c,t-k}$.

Define reporting rate at city i as $\alpha = \frac{R_{i,t-k}}{I_{i,t-k}} = \frac{LR_{i,t}}{I_{i,t-k}}$ and at epicenter c as $\alpha_c = \frac{R_{c,t-k}}{I_{c,t-k}} = \frac{LR_{c,t}}{I_{c,t-k}}$. This means that we are considering the reporting rate of lagged reported cases on the lagged total infection.

We know that in the epicenter c, we have the following:

$$
\begin{aligned}
I_{c,t} &= I_{c,0}\exp(\beta(t-T_0)) && (19)\\
R_{c,t} &= \alpha_c I_{c,t} && (20)\\
U_{c,t} &= (1-\alpha_c) I_{c,t} && (21)\\
&= (1-\alpha_c) I_{c,0}\exp(\beta(t-T_0)) && (22)
\end{aligned}
$$

When only travel data is available, our assumption 3.1 is

$$
\frac{I^{inc}_{i,t}}{M_{i,t}} = \frac{U_{c,t}}{N_c - \hat{R}_{c,t}} \quad \text{for any time } t \in [T_0, T_1], \text{ region } i \text{ and epicenter } c \qquad (23)
$$

We can then write it as

$$
I^{inc}_{i,t} = \frac{M_{i,t}}{N_c - \hat{R}_{c,t}}(1-\alpha_c) I_{c,0}\exp(\beta(t-T_0)) \qquad (24)
$$

In city i, at time T_1 we observe $\widehat{LR}_{i,T_1}$. We have

$$
\begin{aligned}
I_{i,T_1-k} &= \int_{T_0}^{T_1-k} I_{i,t}^{inc} \exp(\beta(T_1-k-t))dt \\
\widehat{LR}_{i,T_1} &= \alpha I_{i,T_1-k} + \epsilon_{i,T_1} \\
&= \alpha \int_{T_0}^{T_1-k} I_{i,t}^{inc} \exp(\beta(T_1-k-t))dt + \epsilon_{i,T_1} \\
&= \alpha \int_{T_0}^{T_1-k} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} (1-\alpha_c) I_{c,0} \exp(\beta(t-T_0)) \exp(\beta(T_1-k-t))dt + \epsilon_{i,T_1} \\
&= \alpha(1-\alpha_c) I_{c,0} \exp(\beta(T_1-T_0-k)) \int_{T_0}^{T_1-k} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1} \\
&= \alpha(1-\alpha_c) \frac{R_{c,0}}{\alpha_c} \exp(\beta(T_1-T_0-k)) \int_{T_0}^{T_1-k} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1} \\
&= \alpha \frac{1-\alpha_c}{\alpha_c} \exp(\beta(T_1-T_0-k)) LR_{c,k} \int_{T_0}^{T_1-k} \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{i,T_1}
\end{aligned}
$$

When both travel data and randomized testing data are available, we maintain assumption 3.2:

$$
\frac{I_{i,t}^{inc}}{M_{i,t}} = \gamma \frac{U_{c,t}}{N_c - \hat{R}_{c,t}} \quad \text{for any time } t \in [T_0, T_1], \text{ region } i \text{ and epicenter } c \tag{25}
$$

We can write it as

$$
I_{i,t}^{inc} = \gamma \frac{M_{i,t}}{N_c - \hat{R}_{c,t}} (1-\alpha_c) I_{c,0} \exp(\beta(t-T_0)) \tag{26}
$$

In US city i, at time T_1 we observe $\widehat{LR}_{i,T_1}$. Following the same derivation as above, we have

$$
\widehat{LR}_{i,T_1} = \alpha\gamma \frac{1-\alpha_c}{\alpha_c} \exp(\beta(T_1-T_0-k)) LR_{c,k} \int_{T_0}^{T_1-k} \frac{M_{i,t}}{N_c - R_{c,t}} dt + \epsilon_{i,T_1} \tag{27}
$$

Similar derivation shows that for Iceland region j time T_1, we have

$$
\hat{I}_{j,T_1-k} = \gamma \frac{1-\alpha_c}{\alpha_c} \exp(\beta(T_1-T_0-k)) LR_{c,k} \int_{T_0}^{T_1-k} \frac{M_{j,t}}{N_c - \hat{R}_{c,t}} dt + \epsilon_{j,T_1} \tag{28}
$$

References

Alvarez, F., D. Argente, and F. Lippi (2020). A simple planning problem for covid-19 lockdown. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2020-34).

Andrei, M. (March 26 2020). Icelands testing suggests 50 *ZME Science*.

Berger, D., K. Herkenhoff, and S. Mongey (2020). An seir infectious disease model with testing and conditional quarantine. Working paper, Federal Reserve Bank of Minneapolis.

Bogoch, I. I., A. Watts, A. Thomas-Bachli, C. Huber, M. U. Kraemer, and K. Khan (2020). Pneumonia of unknown etiology in wuhan, china: Potential for international spread via commercial air travel. *Journal of Travel Medicine*.

Eichenbaum, M. S., S. Rebelo, and M. Trabandt (2020). The macroeconomics of epidemics. Technical report, National Bureau of Economic Research.

Flaxman, S., S. Mishra, and A. e. a. Gandy (March 30 2020). Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries. *Imperial College London*.

Imai, N., I. Dorigatti, A. Cori, S. Riley, and N. M. Ferguson (January 17 2020). Report 1: Estimating the potential total number of novel coronavirus cases in wuhan city, china. *Imperial College London*.

Kaplan, S. and K. Thomas (April 6 2020). Delays and shortages exacerbate coronavirus testing gaps in the u.s. *The New York Times*.

Korolev, I. (2020). Identification and estimation of the seird epidemic model for covid-19. Working paper.

Kucharski, A. J., T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, et al. (2020). Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The lancet infectious diseases*.

Lai, S., I. I. Bogoch, N. Ruktanonchai, A. G. Watts, Y. Li, J. Yu, X. Lv, W. Yang, Y. Hongjie, K. Khan, et al. (2020). Assessing spread risk of wuhan novel coronavirus within and beyond china, january-april 2020: a travel network-based modelling study.

Lauer, S. A., K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler (2020, 03). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*.

Li, Q., X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, et al. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.

Li, R., S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*.

Liu, T., J. Hu, M. Kang, L. Lin, H. Zhong, J. Xiao, G. He, T. Song, Q. Huang, Z. Rong, et al. (2020). Transmission dynamics of 2019 novel coronavirus (2019-ncov).

Liu, Z., P. Magal, O. Seydi, and G. Webb (2020a). Predicting the cumulative number of cases for the covid-19 epidemic in china from early data. *arXiv preprint arXiv:2002.12298*.

Liu, Z., P. Magal, O. Seydi, and G. Webb (2020b). Understanding unreported cases in the covid-19 epidemic outbreak in wuhan, china, and the importance of major public health interventions. *Biology 9*(3), 50.

Nishiura, H., T. Kobayashi, T. Miyama, A. Suzuki, S. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, et al. (2020). Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19). *medRxiv*.

Nishiura, H., T. Kobayashi, Y. Yang, K. Hayashi, T. Miyama, R. Kinoshita, N. M. Linton, S.-m. Jung, B. Yuan, A. Suzuki, et al. (2020). The rate of underascertainment of novel coronavirus (2019-ncov) infection: Estimation using japanese passengers data on evacuation flights.

Read, J. M., J. R. Bridgen, D. A. Cummings, A. Ho, and C. P. Jewell (2020). Novel coronavirus 2019-ncov: early estimation of epidemiological parameters and epidemic predictions. *MedRxiv*.

Shen, M., Z. Peng, Y. Xiao, and L. Zhang (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in china. *bioRxiv*.

Stock, J. (2020). Data gaps and the policy response to the novel coronavirus. Working paper.

Stock, J., K. Aspelund, M. Droste, and C. Walker (April 6 2020). Estimates of the undetected rate among the sars-cov-2 infected using testing data from iceland. Working paper.

Thompson, S., Y. Serkez, and L. Kelley (March 23 2020). How has your state reacted to social distancing? *The New York Times*.

Wu, J. T., K. Leung, and G. M. Leung (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet 395*(10225), 689–697.

Zhao, S., Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, et al. (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases 92*, 214–217.