Accurately Estimating Total COVID-19 Infections using Information Theory

Jiaming Cui[1], Arash Haddadan[2], A S M Ahsan-Ul Haque[3], Jilles Vreeken[4], Bijaya Adhikari[5], Anil Vullikanti[2,3], and B. Aditya Prakash[1,*]

[1]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, US
[2]Biocomplexity Institute, University of Virginia, Charlottesville, VA 22904, US
[3]Department of Computer Science, University of Virginia, Charlottesville, VA 22904, US
[4]CISPA Helmholtz Center for Information Security, Saarbrücken 66123, Germany
[5]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, US

Abstract

One of the most significant challenges in the early combat against COVID-19 was the difficulty in estimating the true magnitude of infections. Unreported infections drove up disease spread in numerous regions, made it very hard to accurately estimate the infectivity of the pathogen, therewith hampering our ability to react effectively. Despite the use of surveillance-based methods such as serological studies, identifying the true magnitude is still challenging today. This paper proposes an information theoretic approach for accurately estimating the number of total infections. Our approach is built on top of Ordinary Differential Equations (ODE) based models, which are commonly used in epidemiology and for estimating such infections. We show how we can help such models to better compute the number of total infections and identify the parameterization by which we need the fewest bits to describe the observed dynamics of reported infections. Our experiments show that our approach leads to not only substantially better estimates of the number of total infections but also better forecasts of infections than standard model calibration based methods. We additionally show how our learned parameterization helps in modeling more accurate what-if scenarios with non-pharmaceutical interventions. Our results support earlier findings that most COVID-19 infections were unreported and non-pharmaceutical interventions indeed helped to mitigate the spread of the outbreak. Our approach provides a general method for improving epidemic modeling which is applicable broadly.

*To whom correspondence should be addressed. E-mail: badityap@cc.gatech.edu

Introduction

The COVID-19 pandemic has emerged as one of the most formidable public health challenges in recent history. By Nov 1, 2022, there were already more than 98 million reported infections and 1.07 million deaths in the United States alone. Worldwide, the reported infections summed to 636 million with at least 6.61 million deaths [19]. The devastating effects of COVID-19 extends to the economy as well. For example, in the US, the unemployment rate peaked at 15.8 percent in April 2020 [6], and US GDP contracted at a 3.5% annualized rate for 2020 [1]. Similar economic impacts have been observed worldwide.

One of the most significant challenges in the early combat against COVID-19 was estimating the number of total infections. A significant number of COVID-19 infections were unreported, due to various factors such as the lack of testing and asymptomatic infections [13, 11, 57, 55, 39]. The inability in estimating these unreported infections allowed them to drive up disease transmission in many regions. For example, phylogenetic studies revealed that COVID-19 had locally spread in Washington state before early 2020, when active community surveillance was implemented [14]. There were only 23 reported infections in five major U.S. cities by March 1, 2020, but it has been estimated that there were in fact more than 28,000 total infections by then [5]. Similar trends were observed in other countries, such as in Italy, Germany, and the UK [60]. Despite having more advanced surveillance techniques such as serological studies, estimating the total number of infections continues to be a challenge for COVID-19 response even today [8, 30].

An accurate estimation of the number of total infections is a fundamental epidemiological question and critical for pandemic planning and response. Not withstanding its importance, there is not even a commonly agreed upon metric. One proposal is the *case ascertainment rate*, which is defined as the ratio of reported symptomatic infections to the actual number of symptomatic infections [52]. Another popular proposal is the *reported rate* α_{reported}, which is defined as the ratio of reported infections to total infections [46]. This definition includes asymptomatic infections, which are known to contribute substantially to community transmission [58, 41]. In this paper, we focus on this particular measure.

However, estimating the reported rate is challenging, and as a result all current methods have their limitations. One of the most effective current methods to identify the reported rate in a region is through large-scale serological studies [56, 26, 64]. These surveys use blood tests to identify the prevalence of antibodies against SARS-CoV-2 in a large population. The CDC COVID Data Tracker portal [2, 26] summarizes the results of serological studies conducted by commercial laboratories at a national level as well as at 10 specific sites. For example, the estimated reported rate was at most 0.1 in Minneapolis and South Florida as of April 2020. This means that there were at least 10 times more total infections than reported infections. While serological studies can give an accurate estimation, they are expensive and are not sustainable in the long run [4]. Furthermore, it is also challenging to obtain real-time data using such studies since there are unavoidable delays between sample collection and laboratory tests [2, 26]. Additional difficulties include sampling biases that make it necessary to use carefully designed heuristics to account for them [9]. Other methods include exploiting existing surveillance systems of related diseases like influenza, and using them to estimate symptomatic infections [40]. However, this can also be unreliable and requires ad-hoc corrections to account for the similarities between COVID-19 and influenza symptoms.

In the face of these challenges, data scientists and epidemiologists have devoted much time and effort to estimate the reported rate α_{reported} through epidemiological models O_{M}. By now, there exist carefully constructed models that capture the transmission dynamics of COVID-19 well [39, 52, 12, 50, 36, 38, 61, 25, 33, 62, 63, 17, 43]. In general, an epidemiological model O_{M} has a set of parameters Θ that we estimate from *observed data* using a so-called calibration procedure,

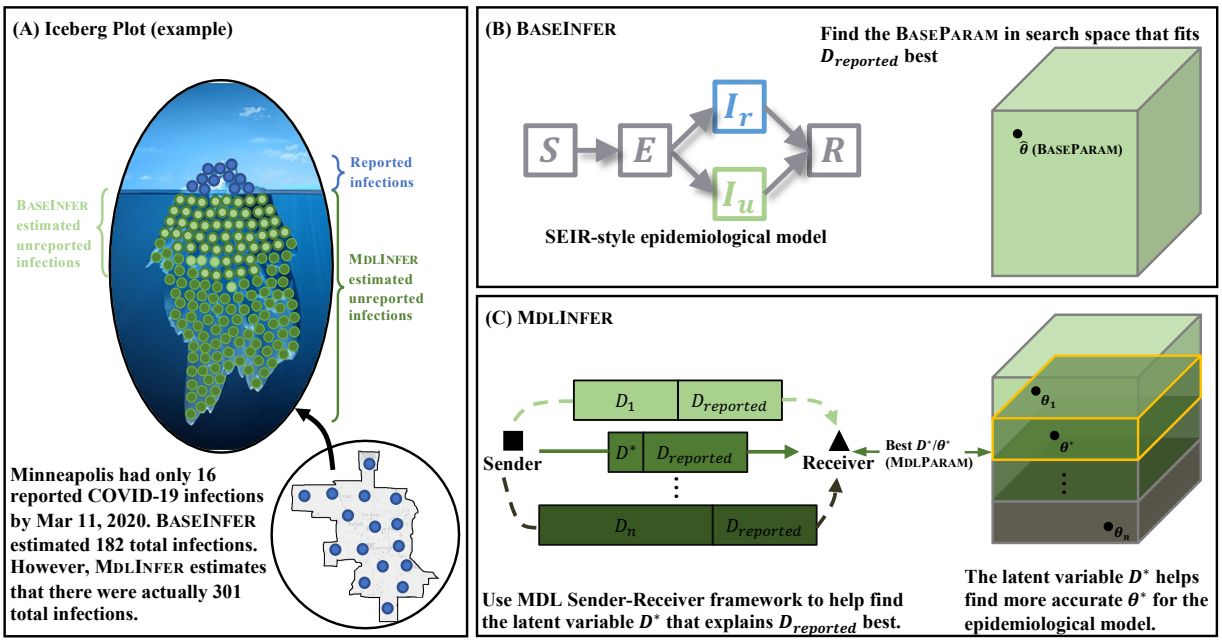

Figure 1: **Overview of our problem and methodology.** (**A**) We visualize the idea of reported rates using the iceberg. The visible portion above water are the reported infections, which is only a fraction of the whole iceberg representing total infections. Light green corresponds to the unreported infections estimated by typical current practice used by researchers (182 in this example). We call it as the basic approach, or BaseInfer. In contrast, dark green corresponds to the more accurate and much larger 301 unreported infections found by our approach MdlInfer. (**B**) The usual practice is to calibrate an epidemiological model to reported data and compute the reported rate from the resultant parameterizaion of the model. Here, an SEIR-style model with explicit compartments for reported-vs-unreported infection is shown in the figure as an example. (**C**) Our new approach MdlInfer instead aims to compute a more accurate reported rate by finding a 'best' parameterization *for the same epidemiological model* (i.e., SEIR-style model in this example) using a principled information theoretic formulation - two-part 'sender-receiver' framework. Assume that a hypothetical Sender S wants to transmit the reported infections as the Data to a Receiver R in the cheapest way possible. Hence S will find/solve for the best D^*, intuitively, the Model that takes the fewest number of bits to encode the Data. Using D^*, we can find the best Θ^* by exploring a smaller search space.

Calibrate. In practice, the data we use for calibration can be the time series of the number of reported infections, which we call D_{reported}. To estimate the number of total infections, these models often explicitly include reported rate α_{reported} as one of their parameters, or include multiple parameters that jointly account for it. There are many calibration procedures commonly used in literature, such as RMSE-based [23] or Bayesian approaches [33, 25].

We call the above general methodology the basic approach to estimate the reported rate, or BaseInfer for short. It takes the epidemiological model O_{M}, a calibration procedure Calibrate, and observed data D_{reported} as input. The output of BaseInfer is then a baseline parameterization $\hat{\Theta}$ and, by extension, an estimated reported rate $\hat{\alpha}_{\text{reported}}$. Calibrating a parameterization is generally a complex, high-demensional problem, since $\hat{\Theta}$ consists of multiple interacting parameters. To make matters worse, there exist many possible parameterizations that show similar performance

(e.g. in RMSE, likelihood) yet correspond to vastly different reported rates. BaseInfer cannot select between these competing parameterizations in a principled way: the parameterization $\hat{\Theta}$ it results in may or may not overfit the reported infections and may or may not predict future infections well. One method for selecting them is to take a Bayesian approach. That is, we choose a prior distribution, and then select the best parameterization that maximizes the posterior probability. Choosing such a prior, however, is ad-hoc and does not generalize well across different models O_{M}. As we will see in the experimental evaluation, minor differences in estimates of reported rates can indeed lead to very different forecasts of future trends and therewith intervention policy recommendations.

Instead, we propose a new information theory-based approach named MdlInfer. It takes the same input as BaseInfer, but uses a principled approach to determine the best parameterization Θ^*. It is based on the following central intuition: Suppose an oracle also gives us the time series of the number of *total infections* D in additional to the already known reported number of infections D_{reported}, and we are asked to describe D_{reported} as succinctly as possible. As we know both D and D_{reported}, it is trivial to estimate $\alpha'_{\text{reported}}$. If we know D and $\alpha'_{\text{reported}}$, it is trivial to describe D_{reported}, as it is simply $D \times \alpha'_{\text{reported}}$ plus a little bit of noise. Now to most succinctly describe D, we have to calibrate O_{M} to obtain Θ'. The only things we now have to describe are Θ', $\alpha'_{\text{reported}}$, the (small) errors that O_{M} makes in predicting D, and the (small) errors that we make predicting D_{reported} using D and $\alpha'_{\text{reported}}$. In practice, we are of course not given D, but the key idea of this paper is to estimate D as a latent variable such that we can most succinctly describe (most accurately reconstruct) the dynamics of D_{reported}.

In practice, we need both a way of measuring how well a latent Model (i.e., D and its corresponding $\alpha'_{\text{reported}}$) describes the Data (i.e., reported infections D_{reported}), as well as a way to find the best such Model. To do so, the Minimum Description Length (MDL) principle provides a statistically sound approach. MDL has been widely used for numerous optimization problems ranging from network summarization [34], causality inference [16], and failure detection in critical infrastructures [10]. MDL has also previously been used for some epidemiological problems, mainly in inferring patient-zero and associated infections in cascades over contact networks [49]. However, we are the first to propose an MDL-based approach on top of ODE-based epidemiological models, which are harder to formulate and optimize.

Specifically, we use two-part MDL (aka sender-receiver framework) consisting of hypothetical actors S and R: Sender S has the Data and wants to transmit it to receiver R using as few bits as possible [24]. Hence, sender S searches for the best possible Model, which minimizes the overall cost of encoding and transmitting both the Model and the Data given the Model. Following the convention in information theory, we use $L(\text{Model})$ to denote the number of bits required to encode the Model; and $L(\text{Data}|\text{Model})$ to denote the number of bits required to encode the Data, D_{reported}, given the Model. The overall objective of our optimization problem is to infer an optimal Model*, which minimizes $L(\text{Model}) + L(\text{Data}|\text{Model})$. To put MDL to practice for our problem, we carefully design our MDL cost to minimize the discrepancy in fitting D_{reported}. This cost ensures the generalizability of our learned D^* and $\alpha^*_{\text{reported}}$ - it can avoid overfitting on D_{reported} and predict the future reported infections well. Our later experiments exactly show this. Our approach, MdlInfer, can be applied to any ODE model since two-part MDL does not assume about the nature of the Data or the Model.

We compare MdlInfer and BaseInfer using two different ODE-based epidemiological models: SAPHIRE [25] and SEIR + HD [33] as O_{M}. Following their literature [25, 33], we use Markov Chain Monte Carlo (MCMC) as the calibration procedure Calibrate for SAPHIRE and iterated filtering (IF) for SEIR + HD, both of with are Bayesian approaches[29]. Both these epidemiological models have previously been shown to perform well in fitting reported infections and provided insight

that was beneficial for the COVID-19 response. SAPHIRE focuses on two key features of the outbreak: high covertness and high transmissibility that drove the outbreak of COVID-19 in Wuhan. SEIR + HD investigates how non-pharmaceutical interventions like social distancing will be needed to maintain epidemic control. These models are broadly representative to show that MdlInfer gives consistent performance across multiple epidemiological models with different dynamics. The experiments clearly show that our proposed MDL-based approach MdlInfer performs superior to the state of the art. To illustrate, we give an example in Fig. 1. By March 11, 2020, the Minneapolis Metro Area had only 16 COVID-19 reported infections. BaseInfer estimated 182 total infections, which are colored as light green in the iceberg. On the other hand, our MdlInfer gives an estimate of 301 total infections shown below the sea level, which is closer to the total infections estimated from serological studies [26, 2]. Additionally, MdlInfer also leads to better fits and *future* projections on reported infections. We also demonstrate that MdlInfer can aid policy making by analyzing counter-factual non-pharmaceutical interventions, while inaccurate BaseInfer estimates lead to wrong non-pharmaceutical intervention conclusions.

Results

Next, we present our empirical findings on a large set of experiments in different geographical regions and time periods. We choose 8 regions and periods based on the severity of the outbreak and the availability of serological studies and symptomatic surveillance data. In each region, we divide the timeline into two time periods: (i) observed period, when only the number of reported infections are available, and both BaseInfer and MdlInfer are used to learn the baseline parameterization (BaseParam) $\hat{\Theta}$ and MDL parameterization (MdlParam) Θ^*, and (ii) forecast period, where we evaluate the forecasts generated by the parameterizations learned in the observed period. To handle the time-varying reported rates, we divide the observed period into multiple sub-periods and learn different reported rates for each sub-period separately.

(A) Estimating total infections: MdlInfer estimates total infections more accurately than BaseInfer

Here, we use the point estimates of the total infections calculated from serological studies as the ground truth (black dots shown in Fig. 2). We call it $\text{SeroStudy}_{\text{Tinf}}$. We also plot MdlInfer's estimation of total infections, $\text{MdlParam}_{\text{Tinf}}$, in the same figure (red curve). To compare the performance of MdlInfer and BaseInfer with $\text{SeroStudy}_{\text{Tinf}}$, we use the cumulative value of estimated total infections. Note that values from the serological studies are not directly comparable with the total infections because of the lag between antibodies becoming detectable and infections being reported [2, 26]. In Fig. 2, we have already accounted for this lag following CDC study guidelines [2, 26] (See Methods section for details). The vertical black lines shows a 95% confidence interval for $\text{SeroStudy}_{\text{Tinf}}$. The blue curve represents total infections estimated by BaseInfer, $\text{BaseParam}_{\text{Tinf}}$. As seen in the figure, $\text{MdlParam}_{\text{Tinf}}$ falls within the confidence interval of the estimates given by serological studies. Significantly, in Fig. 2**B** and Fig. 2**F** for South Florida, BaseInfer for SAPHIRE model [25] overestimates the total infections, while for SEIR + HD model underestimates the total infections. However, MdlInfer consistently estimates the total infections correctly. This observation shows that as needed, $\text{MdlParam}_{\text{Tinf}}$ can improve upon the $\text{BaseParam}_{\text{Tinf}}$ in either direction (i.e., by increasing or decreasing the total infections). Note that the $\text{MdlParam}_{\text{Tinf}}$ curves from both models are closer to the $\text{SeroStudy}_{\text{Tinf}}$ even when the $\text{BaseParam}_{\text{Tinf}}$ curves are different. The results of better accuracy in spite of various

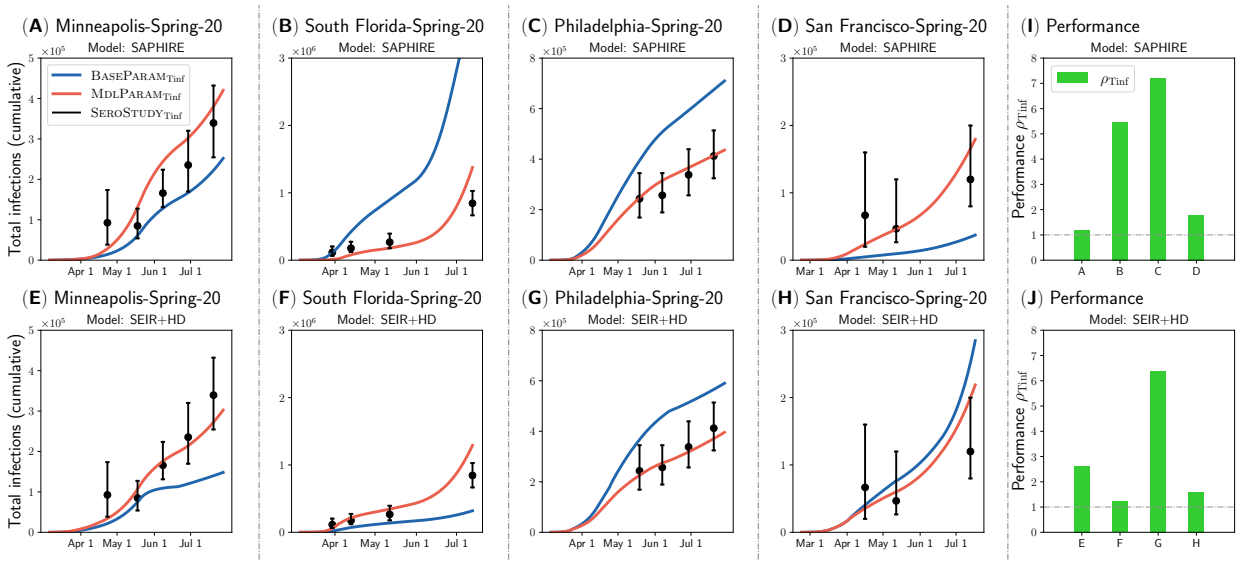

Figure 2: **MdlInfer (red) gives a closer estimation of total infections to serological studies (black) than BaseInfer (blue) on various geographical regions and time periods.** Note that both approaches try to fit the serological studies without being informed with them. (**A**)-(**H**) The red and blue curves represent MdlInfer's estimation of total infections, MdlParam$_{\text{Tinf}}$, and BaseInfer's estimation of total infections, BaseParam$_{\text{Tinf}}$, respectively. The black point estimates and confidence intervals represent the total infections estimated by serological studies [2, 26], SeroStudy$_{\text{Tinf}}$. (**A**)-(**D**) use SAPHIRE model and (**E**)-(**H**) use SEIR + HD model. (**I**)-(**J**) The performance metric, ρ_{Tinf}, comparing MdlParam$_{\text{Tinf}}$ against BaseParam$_{\text{Tinf}}$ in fitting serological studies is shown for each region. (**I**) is for SAPHIRE model in (**A**)-(**D**), and (**J**) is for SEIR + HD model in (**E**)-(**H**). Here, the values of ρ_{Tinf} are 1.20, 5.47, 7.21, and 1.79 in (**I**), and 2.62 ,1.22, 6.39, and 1.58 in (**J**). Note that ρ_{Tinf} larger than 1 means that MdlParam$_{\text{Tinf}}$ is closer to SeroStudy$_{\text{Tinf}}$ than BaseParam$_{\text{Tinf}}$. We show more experiments in the Supplementary Information.

geographical regions and time periods show that MdlInfer is consistently able to estimate total infections more accurately.

To quantify the performance gap between the two approaches, we first compute the root mean squared error (RMSE) between SeroStudy$_{\text{Tinf}}$ and BaseParam$_{\text{Tinf}}$. We also compute the same between SeroStudy$_{\text{Tinf}}$ and MdlParam$_{\text{Tinf}}$. We then compute the ratio, ρ_{Tinf}, of the two RMSE errors as $\frac{\text{Rmse}(\text{BaseParam}_{\text{Tinf}},\text{SeroStudy}_{\text{Tinf}})}{\text{Rmse}(\text{MdlParam}_{\text{Tinf}},\text{SeroStudy}_{\text{Tinf}})}$. Note that the values of ρ_{Tinf} being greater than 1 implies that the MdlParam$_{\text{Tinf}}$ is closer to SeroStudy$_{\text{Tinf}}$ estimates than BaseParam$_{\text{Tinf}}$. In Fig. 2**I** and Fig. 2**J**, we plot ρ_{Tinf}. Overall, the ρ_{Tinf} values are greater than 1 in Fig. 2**I** and Fig. 2**J**, which indicates that MdlInfer performs better than BaseInfer. Note that even when the value of ρ_{Tinf} is 1.20 for Fig. 2**A**, the improvement made by MdlParam$_{\text{Tinf}}$ over BaseParam$_{\text{Tinf}}$ in terms of RMSE is about 12091. Hence, one can conclude that MdlInfer is indeed superior to BaseInfer, when it comes to estimating total infections. We show more experiments in the Supplementary Information.

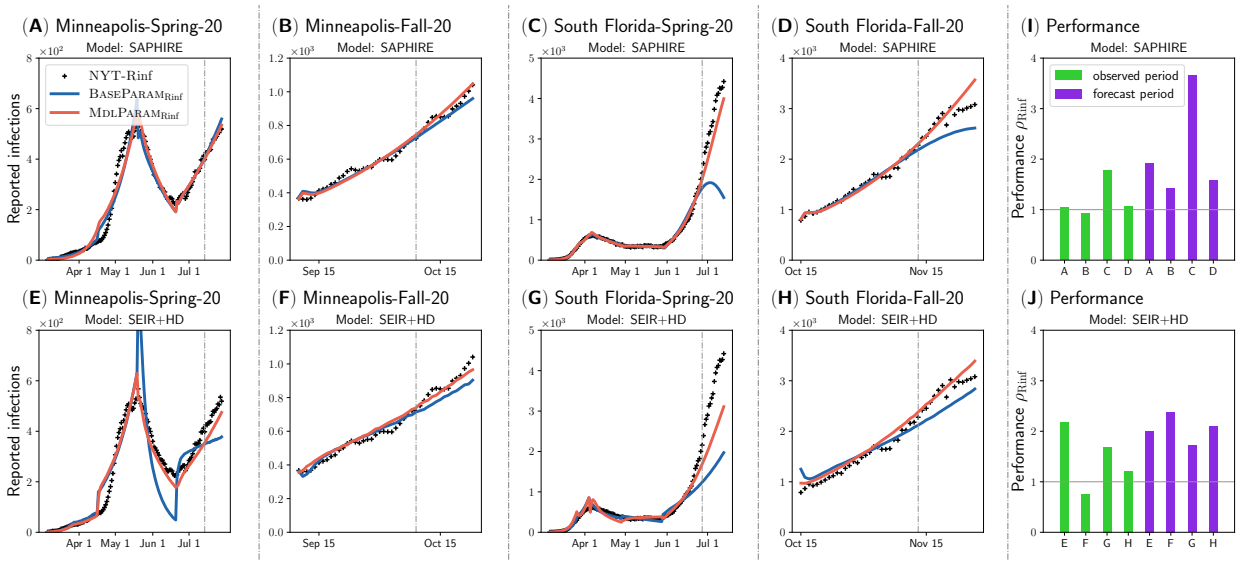

Figure 3: **MdlInfer (red) gives a closer estimation of reported infections (black) than BaseInfer (blue) on various geographical regions and time periods.** We use the reported infections in the observed period as inputs and try to forecast the future reported infections (forecast period). (**A**)-(**H**) The vertical grey dash line divides the observed period (left) and forecast period (right). The red and blue curves represent MdlInfer's estimation of reported infections, $\text{MdlParam}_{\text{Rinf}}$, and BaseInfer's estimation of reported infections, $\text{BaseParam}_{\text{Rinf}}$, respectively. The black plus symbols represent the reported infections collected by the New York Times (NYT-Rinf). (**A**)-(**D**) use SAPHIRE model and (**E**)-(**H**) use SEIR + HD model. (**I**)-(**J**) The performance metric, ρ_{Rinf}, comparing $\text{MdlParam}_{\text{Rinf}}$ against $\text{BaseParam}_{\text{Rinf}}$ in fitting reported infections is shown for each region. (**I**) is for SAPHIRE model in (**A**)-(**D**), and (**J**) is for SEIR + HD model in (**E**)-(**H**). Note that ρ_{Rinf} larger than 1 means that $\text{MdlParam}_{\text{Rinf}}$ is closer to NYT-Rinf than $\text{BaseParam}_{\text{Rinf}}$. We show more experiments in the Supplementary Information.

(B) Estimating reported infections: MdlInfer leads to better fit and projection than BaseInfer at different stages of the COVID-19 epidemic

Here, we first use the observed period to learn the parameterizations. We then *forecast* the future reported infections (i.e., forecast periods), which were *not* accessible to the model while training. The results are summarized in Fig. 3. In Fig. 3**A** to Fig. 3**H**, the vertical grey dash line divides the observed and forecast period. The black plus symbols represent reported infections collected by the New York Times, NYT-Rinf. The red curve represents MdlInfer's estimation of reported infections, $\text{MdlParam}_{\text{Rinf}}$. Similarly, the blue curve represents BaseInfer's estimation of reported infections, $\text{BaseParam}_{\text{Rinf}}$. Note that the curves to the right of the vertical grey line are future predictions. As seen in Fig. 3, $\text{MdlParam}_{\text{Rinf}}$ aligns more closely with NYT-Rinf than $\text{BaseParam}_{\text{Rinf}}$, indicating the superiority of MdlInfer in fitting and forecasting reported infections.

We define a performance metric ρ_{Rinf} as $\frac{\text{Rmse}(\text{BaseParam}_{\text{Rinf}},\text{NYT-Rinf})}{\text{Rmse}(\text{MdlParam}_{\text{Rinf}},\text{NYT-Rinf})}$ to compare $\text{MdlParam}_{\text{Rinf}}$ against $\text{BaseParam}_{\text{Rinf}}$ in a manner similar to ρ_{Tinf}. In Fig. 3**I** and Fig. 3**J**, we plot the ρ_{Rinf} for the observed and forecast period. In both periods, we notice that the ρ_{Rinf} is close to or greater than 1. This further shows that MdlInfer has a better or at least closer fit for reported infections than BaseInfer. Additionally, the ρ_{Rinf} for the forecast period is even greater than ρ_{Rinf}

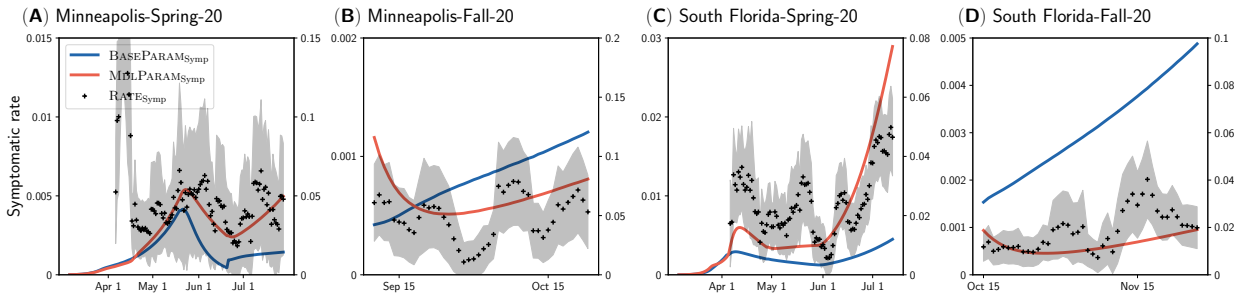

Figure 4: **MdlInfer (red) gives a closer estimation of the trends of symptomatic rate (black) than BaseInfer (blue) on various geographical regions and time periods.** (**A**)-(**D**) The red and blue curves represent MdlInfer's estimation of symptomatic rate, $\text{MdlParam}_{\text{Symp}}$, and BaseInfer's estimation of symptomatic rate, $\text{BaseParam}_{\text{Symp}}$, respectively. They use the y-scale on the left. The black points and the shaded regions are the point estimate with standard error for $\text{Rate}_{\text{Symp}}$ (the COVID-related symptomatic rates derived from the symptomatic surveillance dataset [51, 53]). They use the y-scale on the right. Note that we focus on trends instead of the exact numbers, hence $\text{MdlParam}_{\text{Symp}}/\text{BaseParam}_{\text{Symp}}$, and $\text{Rate}_{\text{Symp}}$ may scale differently. We show more experiments in the Supplementary Information.

for the observed period, which shows that MdlInfer performs even better than BaseInfer while forecasting.

Note that Fig. 3**A**, **C**, **E**, **G** correspond to the early state of the COVID-19 epidemic in spring and summer 2020, and Fig. 3**B**, **D**, **F**, **H** correspond to fall 2020. We can see that MdlInfer performs well in estimating temporal patterns at different stages of the COVID-19 epidemic. We show more experiments in the Supplementary Information.

(C) Estimating symptomatic rate trends: MdlInfer estimates the symptomatic rate trends more accurately than BaseInfer

We validate this observation using Facebook's symptomatic surveillance dataset [51]. We plot MdlInfer's and BaseInfer's estimated symptomatic rate over time and overlay the estimates and standard error from the symptomatic surveillance data in Fig. 4. The red and blue curves are MdlInfer's and BaseInfer's estimation of symptomatic rates, $\text{MdlParam}_{\text{Symp}}$ and $\text{BaseParam}_{\text{Symp}}$ respectively. Note that SAPHIRE model does not contain states corresponding to the symptomatic infections. Therefore, we only focus on SEIR + HD model. We compare the trends of the $\text{MdlParam}_{\text{Symp}}$ and $\text{BaseParam}_{\text{Symp}}$ with the symptomatic surveillance results. We focus on trends rather than actual values because the symptomatic rate numbers could be biased [51] (see Methods section for a detailed discussion) and therefore cannot be compared directly with model outputs like what we have done for serological studies. As seen in Fig. 4, $\text{MdlParam}_{\text{Symp}}$ captures the trends of the surveyed symptomatic rate $\text{Rate}_{\text{Symp}}$ (black plus symbols) better than $\text{BaseParam}_{\text{Symp}}$. We show more experiments in the Supplementary Information.

To summarize, these three sets of experiments in (A), (B) and (C) together demonstrate that BaseInfer fail to accurately estimate the total infections including unreported ones. On the other hand, MdlInfer estimates total infections closer to those estimated by serological studies and better fits reported infections and symptomatic rate trends.

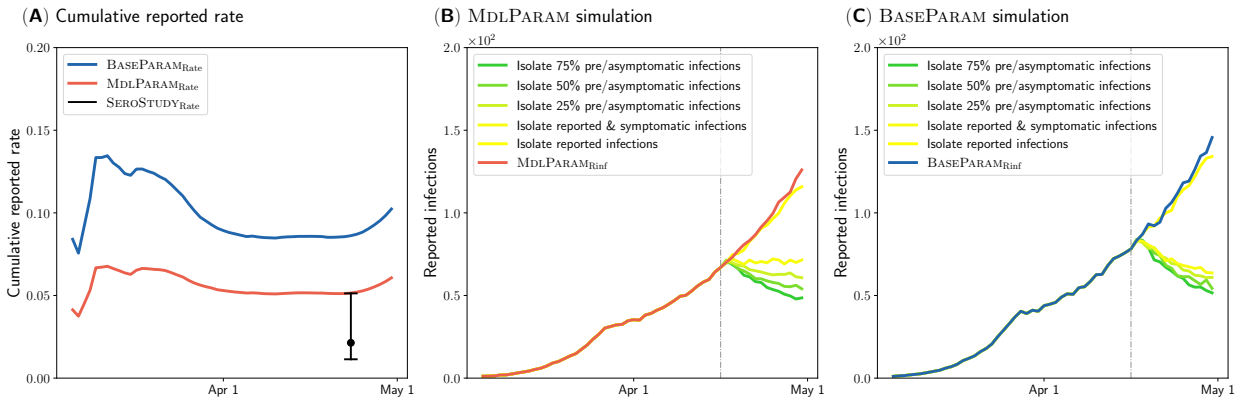

Figure 5: (**A**) MdlInfer estimates cumulative reported rate more accurately than BaseInfer: The blue and red curve represent the cumulative reported rate estimated by BaseInfer, BaseParam$_{\text{Rate}}$, and by MdlInfer, MdlParam$_{\text{Rate}}$, respectively. The black point estimate and its confidence interval represent the cumulative reported rate SeroStudy$_{\text{Rate}}$ estimated by serological studies [2, 26]. Note that both approaches try to fit the SeroStudy$_{\text{Rate}}$ without being informed with them. The results reveal that a large majority of COVID-19 infections were unreported. (**B**) MdlInfer reveals that non-pharmaceutical interventions (NPI) on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic. Here, the red curve and other five curves represent the MdlInfer's estimation of reported infections for no NPI scenario and 5 different NPI scenarios described in the Results section. The vertical grey dash line divides the observed period (left) and forecast period (right). (**C**) Inaccurate estimation by BaseInfer may lead to wrong non-pharmaceutical intervention conclusions. The blue curve and other five curves represent the BaseInfer's estimation of reported infections for no NPI scenario and the same 5 scenarios in (**B**).

Evaluating the effect of non-pharmaceutical Interventions

We have already shown that MdlInfer is able to estimate the number of total infections accurately. In the following three observations, we show that such accurate estimations are important for evaluating the effect of non-pharmaceutical interventions.

(D) MdlInfer reveals that a large majority of COVID-19 infections were unreported

We compute the *cumulative* reported rate MdlParam$_{\text{Rate}}$ measured by the ratio of the cumulative value of reported infections to the total infections estimated by MdlInfer over time and plotted it for Minneapolis-Spring-20 in Fig. 5**A**. The figure shows that the MdlParam$_{\text{Rate}}$ increases in early March, and then gradually decreases. This observation is explained by the community spread-driven COVID-19 outbreaks that were not reported until early March, which fits earlier studies [40].

(E) Non-pharmaceutical interventions on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic

Our simulations show that non-pharmaceutical interventions on asymptomatic and presymptomatic infections are essential to control COVID-19. Here, we plot the simulated reported infections of MdlParam in Fig. 5**B** (red curve). We then repeat the simulation of reported infections for 5 different scenarios: (i) isolate just the reported infections, (ii) isolate just the symptomatic infec-

tions, and isolate symptomatic infections in addition to (iii) 25%, (iv) 50%, and (v) 75% of both asymptomatic and presymptomatic infections. In our setup, we assume that the infectivity reduces by half when a person is isolated. As seen in Fig. 5**B**, when only the reported infections are isolated, there is almost no change in the "future" reported infections. However, when we isolate both the reported and symptomatic infections, the reported infections decreases significantly. Even here, the reported infections are still not in decreasing trend. On the other hand, non-pharmaceutical interventions for some fraction of asymptomatic and presymptomatic infections make reported infections decrease. Thus, we can conclude that non-pharmaceutical interventions on asymptomatic infections are essential in controlling the COVID-19 epidemic.

(F) Accuracy of non-pharmaceutical intervention simulations relies on the good estimation of parameterization

Next, we also plot the simulated reported infections generated by BaseInfer in Fig. 5**C** (blue curve). As seen in the figure, based on BaseInfer, we can infer that only non-pharmaceutical interventions on symptomatic infections are enough to control the COVID-19 epidemic. However, this has been proven to be incorrect by prior studies and real-world observations [41]. Therefore, we can conclude that the accuracy of non-pharmaceutical intervention simulation relies on the quality of the learned parameterization.

Discussion and Future Work

This study proposes MdlInfer, a data-driven model selection approach that automatically estimates the number of total infections using epidemiological models. Our approach leverages the information theoretic Minimum Description Length (MDL) principle to select total infections that "best describe" the observed outbreak. Our approach addresses several gaps in current practice including the long-term infeasibility of serological studies [26], and ad-hoc assumptions in epidemiological models [33, 39, 44, 25].

Overall, our results show that MdlInfer estimates total infections at various geographical locations and different epidemiological models more accurately than BaseInfer from both directions, i.e., it corrects both over- and under-estimates. For example, compared to BaseInfer, we correctly estimate 55719 more infections by April 1 for the SEIR + HD model in Fig. 2**F**, and 87636 fewer infections for the SAPHIRE model in Fig. 2**B** for South Florida-Spring-20. We also show that MdlInfer leads to a better fit of the reported infections in the observed period and more accurate forecasts for the forecast period than BaseInfer. We reveal that a large majority of COVID-19 infections were unreported, where non-pharmaceutical interventions on unreported infections can help to mitigate the COVID-19 outbreak. We also show that MdlInfer estimates more accurate symptomatic rate trends than BaseInfer. Additionally, our results show consistent performance with respect to the reported infections and serological studies on both SAPHIRE and SEIR + HD model. We also show that MdlInfer identifies the ground truth parameters better than BaseInfer (see Supplementary Information section for details). As an aside, BaseInfer may also give uncertainty estimates for their calibrated parameterizations. Our framework MdlInfer can be adapted to generate such estimates as well (see Supplementary Information section for a demonstration).

The MdlInfer framework is likely to be helpful in the surveillance of COVID-19 in the near future, and for future epidemics. Even with the U.S. returning to normalcy, surveillance of the pandemic is still essential for public health. The daily incidence of COVID-19 has decreased from early 2021 to summer 2021, according to the CDC COVID Data Tracker portal [7, 35]. However, new variants of the SARS-CoV-2 (e.g., the Delta and Omicron variants) have been spreading rapidly [37,

48, 21]. Testing for these new variants and large-scale surveillance via laboratory tests may be limited and less systematic than what was done for COVID-19 before. In such settings, using our MdlInfer framework, epidemiologists and policymakers can improve the accuracy of estimates of total infections (without large-scale serological studies), as well as forecasts of their models.

One of the limitations of our work is that the benefits of using MdlInfer depends on the suitability of the epidemiological model. If the epidemiological model is not expressive enough for the observed data, then the gains from MdlInfer may not be significant. As a future work, it may be helpful to adapt MdlInfer to measure the quality of an epidemiological model. We also note that MdlInfer is built on ODE-based epidemiological models; other kinds of epidemic models, e.g., agent-based models [42, 28, 59, 18, 45], are more suitable in some settings. It would be interesting to extend MdlInfer to incorporate such models. Finally, there is significant population or spatial heterogeneity in disease outcomes [15, 31], e.g., differences in severity rate or mortality rate, when infected with COVID-19, for different age groups [22, 27], which has not been considered in our work.

To summarize, MdlInfer is a robust data-driven method to accurately estimate total infections, which will help data scientists, epidemiologists, and policy-makers to further improve existing ODE-based epidemiological models, make accurate forecasts, and combat the ongoing COVID-19 pandemic. More generally, MdlInfer opens up a new line of research in epidemic modeling using information theory.

Materials and Methods

Data

Datasets

We use the following publicly available datasets for our study:

1. **New York Times reported infections [3]:** This dataset (NYT-Rinf) consists of the daily time sequence of reported COVID-19 infections D_{reported} and the mortality $D_{\text{mortality}}$ (cumulative values) for each county in the US starting from January 21, 2020 to current.

2. **Serological studies [26, 2]:** This dataset consists of the point and 95% confidence interval estimates of the prevalence of antibodies to SARS-CoV-2 in 10 US locations every 3–4 weeks from March to July 2020. For each location, CDC works with commercial laboratories to collect the blood specimens in the population and test them for antibodies to SARS-CoV-2. Each specimen collection period ranges from 6 to 14 days. As suggested by prior work [32, 47], these serological studies have high sensitivity to antibodies for 6 months after infections. Hence, using the prevalence and total population in one location, we can compute the estimated total infections $\text{SeroStudy}_{\text{Tinf}}$ for the past 6 months (i.e., from the beginning of the pandemic since January 2020). However, this $\text{SeroStudy}_{\text{Tinf}}$ can not be compared with the epidemiological model estimated total infection numbers directly. The reasons are (i) the antibodies may take 10 to 14 days delay to be detectable after infection [65, 54] and (ii) the 6-14 range period for specimen collection as mentioned before. To account for this, we compare the $\text{SeroStudy}_{\text{Tinf}}$ numbers with the MdlInfer and BaseInfer estimated total infections of 7 days prior to the first day of specimen collection period as suggested by the CDC serological studies work [26].

3. **Symptomatic surveillance [51, 53]:** This dataset consists of point estimate $\text{RATE}_{\text{Symp}}$ and standard error of the COVID-related symptomatic rate for each county in the US starting from April 6, 2020 to date. The survey asks a series of questions on randomly sampled social media (Facebook) users to estimate the percentage of people who have a COVID-like symptoms such as the fever along with cough or shortness of breath or difficulty breathing on a given day. However, there are several caveats such as they could not cover all symptoms of COVID-19 and these symptoms can be also caused by many other conditions, due to which they are not expected to be unbiased estimates for the true symptomatic rate [51]. Besides, as the original symptomatic surveillance data is at a county level, we sum up the numbers to compute the $\text{RATE}_{\text{Symp}}$ and focus on trends instead of the exact numbers.

Our Approach

Two-part sender-receiver framework

In this work, we use two-part sender-receiver framework. The conceptual goal of the framework is to transmit the DATA from the possession of the hypothetical sender S to the hypothetical receiver R. We assume the sender does this by first sending a MODEL and then sending the DATA under this MODEL. In this MDL framework, we want to minimize the number of bits for this process. We do this by identifying the MODEL that encodes the DATA such that the total number of bits needed to encode both the MODEL and the DATA is minimized. Hence our cost function in the total number of bits needed is composed of two parts: (i) model cost $L(\text{MODEL})$: The cost in bits of encoding the MODEL and (ii) data cost $L(\text{DATA}|\text{MODEL})$: The cost in bits of encoding the DATA given the MODEL. Intuitively, the idea is that a good MODEL will lead to a fewer number of bits needed to encode both MODEL and DATA. We formulate the general MDL optimization problem as follows: Given the DATA, $L(\text{MODEL})$, and $L(\text{DATA}|\text{MODEL})$, find MODEL^* such that

$$\text{MODEL}^* = \arg\min_{\text{MODEL}} L(\text{MODEL}) + L(\text{DATA}|\text{MODEL}) \tag{1}$$

In our situation, the DATA is the reported COVID-19 infections D_{reported}: it is the only real-world data given to us. Note that total infections are not directly observed. As described in the introduction section, the MODEL is intuitively $(D, \alpha'_{\text{reported}})$. Here D refers to a candidate total infections time series, and $\alpha'_{\text{reported}}$ is the corresponding reported rate. Specifically, we calibrate O_{M} on (D, D_{reported}) using CALIBRATE to get the "candidate" parameterization Θ', and then compute $\alpha'_{\text{reported}}$ from Θ'. Further, we choose to also add $\hat{\Theta}$ estimated by BASEINFER, making our MODEL to be $(D, \Theta', \hat{\Theta})$. There are alternative MODELs that can be considered, but we choose this MODEL $= (D, \Theta', \hat{\Theta})$ and explain more in the Supplementary Information. Note that as two-part MDL (and MDL in general) does not assume the nature of the DATA or the MODEL, our MDLINFER can be applied to any ODE model. We have also discussed intuitive advantages of the MDLINFER over BASEINFER briefly in the introduction section (see Supplementary Information for more details). Next, we give more details how to formulate our problem of estimating total infections D.

MDL formulation

First, we need to introduce some notations. Given an epidemiological model O_{M} and the paramterization $\hat{\Theta}$ estimated by BASEINFER, we can compute the reported infections. However, this is only an estimate of the reported infections rather than the exact D_{reported}. This is because even though we have already calibrated O_{M} using D_{reported}, the calibration cannot be perfect, and there will be

differences between these estimated reported infections and D_{reported}. Here, we term this estimated reported infections as $D_{\text{reported}}(\hat{\Theta})$. We can also estimate the total infections $D(\hat{\Theta})$ for O_{M} in the same way. Similarly, we have the $D_{\text{reported}}(\Theta^{'})$ and $D(\Theta^{'})$ for $\Theta^{'}$. As described in the introduction section, we can also calculate the reported rate $\hat{\alpha}_{\text{reported}}$ and $\alpha'_{\text{reported}}$ using $\hat{\Theta}$ and $\Theta^{'}$. With these notations, next we will formulate the space of all possible MODELs and give the equation for the cost in bits of encoding MODEL and DATA.

MODEL space

We have MODEL $= (D, \Theta^{'}, \hat{\Theta})$ as described above. Hence our MODEL space will be all possible daily sequences for D and all possible parameterizations for $\Theta^{'}$ and $\hat{\Theta}$. The MDL framework will search in this space to find the MODEL*.

Model cost

With MODEL $= (D, \Theta^{'}, \hat{\Theta})$, we conceptualize the model cost by imagining that the sender S will send the MODEL $= (D, \Theta^{'}, \hat{\Theta})$ to the receiver R in three parts: (i) first send the $\hat{\Theta}$ by encoding $\hat{\Theta}$ directly (ii) next send the $\Theta^{'}$ given $\hat{\Theta}$ by encoding $\Theta^{'} - \hat{\Theta}$ and (iii) then send D given $\Theta^{'}$ and $\hat{\Theta}$ by encoding $\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\hat{\Theta})$. Intuitively, both $\alpha'_{\text{reported}} \times D$ and $D_{\text{reported}}(\hat{\Theta})$ should be close to D_{reported}, and the receiver could recover the D using $\hat{\Theta}$, $\alpha'_{\text{reported}}$, and $D_{\text{reported}}(\hat{\Theta})$ as they have already been sent. We term the model cost as $L(D, \Theta^{'}, \hat{\Theta})$ with three components: COST$(\hat{\Theta})$, COST$(\Theta^{'}|\hat{\Theta})$, and COST$(D|\Theta^{'}, \hat{\Theta})$. Hence,

$$L(D, \Theta^{'}, \hat{\Theta}) = \text{COST}(\hat{\Theta}) + \text{COST}(\Theta^{'} - \hat{\Theta}|\hat{\Theta}) + \text{COST}(\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\hat{\Theta})|\Theta^{'}, \hat{\Theta}) \tag{2}$$

For Equation 2, the COST$(\cdot)$ function gives the total number of bits we need to spend in encoding each term. The details of the encoding method can be found in the Supplementary Information.

Data cost

We need to send the DATA $= D_{\text{reported}}$ next given the MODEL. Given MODEL $= (D, \Theta^{'}, \hat{\Theta})$, we send DATA by encoding $\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\Theta^{'})$. Intuitively, $D - D_{\text{reported}}$ corresponds to the unreported infections, and $1 - \alpha'_{\text{reported}}$ is the unreported rate. Therefore, $\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}}$ should be close to the total infections D and $D(\Theta^{'})$. The receiver could also recover the D_{reported} using D, $\alpha'_{\text{reported}}$, and $D(\Theta^{'})$ as they have already been sent. We term data cost as $L(D_{\text{reported}}|D, \Theta^{'}, \hat{\Theta})$ and formulate it as Equation 3.

$$L(D_{\text{reported}}|D, \Theta^{'}, \hat{\Theta}) = \text{COST}(\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\Theta^{'})|D, \Theta^{'}, \hat{\Theta}) \tag{3}$$

Total cost

With $L(D, \Theta^{'}, \hat{\Theta})$ as in Equation 2 and $L(D_{\text{reported}}|D, \Theta^{'}, \hat{\Theta})$ as in Equation 3 above, the total cost $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$ is:

$$
\begin{aligned}
L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta}) &= L(D, \Theta^{'}, \hat{\Theta}) + L(D_{\text{reported}}|D, \Theta^{'}, \hat{\Theta}) \\
&= \textsc{Cost}(\hat{\Theta}) + \textsc{Cost}(\Theta^{'} - \hat{\Theta}|\hat{\Theta}) \\
&+ \textsc{Cost}(\alpha^{\prime}_{\text{reported}} \times D - D_{\text{reported}}(\hat{\Theta})|\Theta^{'}, \hat{\Theta}) \\
&+ \textsc{Cost}(\frac{D - D_{\text{reported}}}{1 - \alpha^{\prime}_{\text{reported}}} - D(\Theta^{'})|D, \Theta^{'}, \hat{\Theta})
\end{aligned} \tag{4}
$$

Problem statement

Note that our main objective is to estimate the total infections D. With $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$ formulated in Equation 4, we can state the problem as: Given the time sequence D_{reported}, epidemiological model O_{M}, and a calibration procedure CALIBRATE, find D^* that minimizes the MDL total cost i.e.

$$
D^* = \arg\min_{D} \; L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta}) \tag{5}
$$

Algorithm

Next, we will present our algorithm to solve the problem in Equation 5. Note that directly searching D^* naively is intractable since D^* is a daily sequence not a scalar. Instead, we propose first finding a "good enough" reported rate $\alpha^*_{\text{reported}}$ quickly with the constraint $D = \frac{D_{\text{reported}}}{\alpha^*_{\text{reported}}}$ to reduce the search space. Then with this $\alpha^*_{\text{reported}}$, we can search for the optimal D^* in Equation 5. Hence we propose a two-step algorithm: (i) do a linear search to find a good reported rate $\alpha^*_{\text{reported}}$ (ii) given the $\alpha^*_{\text{reported}}$ found above, use an optimization method to find the D^* that minimizes $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$ with $\alpha^*_{\text{reported}}$ constraints.

Step 1: Find the $\alpha^*_{\text{reported}}$

In step 1, we do a linear search on α_{reported} to find the $\alpha^*_{\text{reported}}$. As stated before, we use $\frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$ as D in $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$ to help reduce the search space. Here, we formulate step 1 algorithm as Equation 6.

$$
\alpha^*_{\text{reported}} = \arg\min_{\alpha_{\text{reported}}} L(D_{\text{reported}}, \frac{D_{\text{reported}}}{\alpha_{\text{reported}}}, \Theta^{'}, \hat{\Theta}) \tag{6}
$$

Step 2: Find the D^* given $\alpha^*_{\text{reported}}$

With the $\alpha^*_{\text{reported}}$ found in step 1, we next find the D^* that minimizes the $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$. Note that we have already found a good $\alpha^*_{\text{reported}}$, we can constrain the D^* to ensure that the sum of D^* equals to the sum of $\frac{D_{\text{reported}}}{\alpha^*_{\text{reported}}}$. We use the Nelder-Mead method [20] to solve this constrained optimization problem for D^*. Here, we formulate step 2 algorithm as Equation 7.

$$
D^* = \arg\min_{D} L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta}) \tag{7}
$$

We describe the two-step algorithm in more detail in the Supplementary Information.

BaseInfer and MdlInfer formulation

Here, we also give the mathematical formulations for BaseInfer and MdlInfer. As described in the introduction section, given an epidemiological model O_{M}, a typical approach is to calibrate the O_{M} to D_{reported} using the calibration procedure Calibrate. We call this methodology as BaseInfer(O_{M}, Calibrate, D_{reported}). As in Equation 8, the output of BaseInfer is the baseline parameterization (BaseParam) $\hat{\Theta}$.

$$\begin{aligned}\hat{\Theta} &= \textsc{BaseInfer}(O_{\mathrm{M}}, \textsc{Calibrate}, \mathrm{D}_{\text{reported}}) \\ &= \textsc{Calibrate}(O_{\mathrm{M}}, \mathrm{D}_{\text{reported}})\end{aligned} \tag{8}$$

As for the MdlInfer, it also takes the same input (O_{M}, Calibrate, D_{reported}) as BaseInfer. Assume we are given the total infections D, we calibrate the O_{M} on (D, D_{reported}) to get a "candidate" paramterization $\Theta^{'}$ in Equation 9.

$$\Theta^{'} = \textsc{Calibrate}(O_{\mathrm{M}}, (D, D_{\text{reported}})) \tag{9}$$

However, we are not given the D. Hence, we use the MDL framework to find such D^* as in Equation 7. With such D^*, we could finally calibrate the O_{M} on $(D^*, D_{\text{reported}})$ and gets another parameterization Θ^*. As in Equation 10, we call Θ^* as the MDL parameterization, or MdlParam.

$$\begin{aligned}\Theta^* &= \textsc{MdlInfer}(O_{\mathrm{M}}, \textsc{Calibrate}, \mathrm{D}_{\text{reported}}) \\ &= \textsc{Calibrate}(O_{\mathrm{M}}, (D^*, D_{\text{reported}}))\end{aligned} \tag{10}$$

where $D^* = \arg\min_D L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$. Intuitively, if $\hat{\Theta}$ estimated by BaseInfer is perfect, MdlInfer will also give the same Θ^* as $\hat{\Theta}$.

Epidemiological models

Next, we describe the two epidemiological models we use in our experiments: SEIR + HD and SAPHIRE model. SEIR + HD [33] consists of 10 states: Susceptible S, exposed E, pre-symptomatic I_P, severe symptomatic I_S, mild symptomatic I_M, asymptomatic I_A, hospitalized (eventual death) H_D, hospitalized (eventual recover) H_R, recovered R, and dead D. The parameters to be calibrated are the transmission rate β_0 (the transmission rate in the absence of interventions), σ (the proportional reduction on β_0 under shelter-in-place), and E_0 (number of initial infections). The other parameters are fixed and given. They assume the importations only happen at the beginning of the pandemic (captured by E_0), and the total population N remains constant. We also extend SEIR + HD model to infer two more parameters: α (proportion of asymptomatic infections) and α_1 (proportion of new symptomatic infections that are reported). We compute the new reported infections and unreported infections as follows:

1. New reported infections = $\alpha_1 \times (N_{I_P I_S} + N_{I_P I_M})$: Here $N_{I_P I_S} + N_{I_P I_M}$ is the number of new symptomatic infections everyday. $N_{I_P I_S}$ is the number of patients switching their state from I_P to I_S (and similarly for $N_{I_P I_M}$). We assume α_1 proportion of new symptomatic infections every day are reported.

2. New unreported infections = $(1 - \alpha_1) \times (N_{I_P I_S} + N_{I_P I_M}) + N_{E I_A}$.

SAPHIRE [25] consists of 7 states: Susceptible S, exposed E, pre-symptomatic P, ascertained infectious I, unascertained infectious A, hospitalized H, and recovered R. The parameters to be calibrated are the transmission rate β and reported rate r while keeping other parameters fixed as given values. We also compute the new reported infections and unreported infections as follows:

1. New reported infections $= \frac{rP}{D_p}$: Here $\frac{P}{D_p}$ is the number of new infections from pre-symptomatic every day. D_p is the parameter for the presymptomatic infectious period and is fixed. r is the reported rate estimated by the epidemiological model.

2. New unreported infections $= \frac{(1-r)P}{D_p}$.

Estimating infections using BaseParam and MdlParam

Here, we describe how we get the estimations in the results section using BaseParam and MdlParam. Here we use the BaseParam from BaseInfer as the example (this can also be repeated for MdlParam for MdlInfer). Using the epidemiological model O_{M}, we can calculate the BaseParam's estimation of total infections $\mathrm{BaseParam}_{\mathrm{Tinf}}$ as the cumulative values of $D(\hat{\Theta})$ from pandemic's beginning. $D_{\mathrm{reported}}(\hat{\Theta})$ can be directly used as the BaseParam's estimation of reported infections. For the cumulative reported rate $\mathrm{BaseParam}_{\mathrm{Rate}}$, we calculate it as the cumulative values of NYT-Rinf divided by $D(\hat{\Theta})$. For the symptomatic rate, SEIR + HD model [33] could estimate the number of symptomatic rate $\mathrm{BaseParam}_{\mathrm{Symp}}$ by dividing the number of infections in state I_S and I_M by the population number. However, SAPHIRE model [25] does not contain states that correspond to the symptomatic cases, so we cannot estimate the symptomatic rate using this model.

Acknowledgements

This paper was supported in part by the NSF (Expeditions CCF-1918770 and CCF-1918656, CAREER IIS-2028586, RAPID IIS-2027862, Medium IIS-1955883, Medium IIS-2106961, IIS-1931628, IIS-1955797, IIS-2027848, PIPP CCF-2200269), NIH 2R01GM109718, CDC MInD program U01CK000589, ORNL and funds/computing resources from Georgia Tech and GTRI. B. A. was in part supported by the CDC MInD-Healthcare U01CK000531-Supplement. A.V.'s work is also supported in part by grants from the UVA Global Infectious Diseases Institute (GIDI). J.V. is institutionally funded by CISPA.

References

[1] 4q gdp: Economy expands at a 4.0% annualized rate. https://finance.yahoo.com/news/4q-gdp-2020-us-economy-coronavirus-pandemic-180133456.html.

[2] Commercial laboratory seroprevalence survey data. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html.

[3] Coronavirus in the u.s.:latest map and case count. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html.

[4] Covid-19 test price information, https://www.questdiagnostics.com/business-solutions/health-plans/covid-19/pricing.

[5] Hidden outbreaks spread through u.s. cities far earlier than americans knew, estimates say. https://www.nytimes.com/2020/04/23/us/coronavirus-early-outbreaks-cities.html.

[6] Interim economic projections for 2020 and 2021. https://www.cbo.gov/publication/56351.

[7] United states covid-19 cases, deaths, and laboratory testing (naats) by state, territory, and jurisdiction. https://covid.cdc.gov/covid-data-tracker/#cases.

[8] Us covid cases likely more than double official count, experts say. https://www.cidrap.umn.edu/news-perspective/2021/07/us-covid-cases-likely-more-double-official-count-experts-say, 2021.

[9] Accorsi, E. K., Qiu, X., Rumpler, E., Kennedy-Shaffer, L., Kahn, R., Joshi, K., Goldstein, E., Stensrud, M. J., Niehus, R., Cevik, M., et al. How to detect and reduce potential sources of biases in studies of sars-cov-2 and covid-19. *European Journal of Epidemiology* (2021), 1–18.

[10] Adhikari, B., Rangudu, P., Prakash, B. A., and Vullikanti, A. Near-optimal mapping of network states using probes. In *Proceedings of the 2018 SIAM International Conference on Data Mining* (2018), SIAM, pp. 108–116.

[11] Aguilar, J. B., Faust, J. S., Westafer, L. M., and Gutierrez, J. B. Investigating the impact of asymptomatic carriers on covid-19 transmission. *MedRxiv* (2020).

[12] Angelopoulos, A. N., Pathak, R., Varma, R., and Jordan, M. I. On identifying and mitigating bias in the estimation of the covid-19 case fatality rate. *arXiv preprint arXiv:2003.08592* (2020).

[13] Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.-Y., Chen, L., and Wang, M. Presumed asymptomatic carrier transmission of covid-19. *JAMA 323*, 14 (2020), 1406–1407.

[14] Bedford, T., Greninger, A. L., Roychoudhury, P., Starita, L. M., Famulare, M., Huang, M.-L., Nalla, A., Pepper, G., Reinhardt, A., Xie, H., et al. Cryptic transmission of sars-cov-2 in washington state. *Science 370*, 6516 (2020), 571–575.

[15] Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S. A., Zhang, T., et al. Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen, china: a retrospective cohort study. *The Lancet infectious diseases 20*, 8 (2020), 911–919.

[16] Budhathoki, K., and Vreeken, J. Origo: causal inference by compression. *Knowledge and Information Systems 56*, 2 (2018), 285–307.

[17] Cao, Q., and Heydari, B. Micro-level social structures and the success of covid-19 national policies. *Nature Computational Science 2*, 9 (2022), 595–604.

[18] Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., and Leskovec, J. Mobility network models of covid-19 explain inequities and inform reopening. *Nature 589*, 7840 (2021), 82–87.

[19] Dong, E., Du, H., and Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases 20*, 5 (2020), 533–534.

[20] Gao, F., and Han, L. Implementing the nelder-mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications 51*, 1 (2012), 259–277.

[21] Geers, D., Shamier, M. C., Bogers, S., den Hartog, G., Gommers, L., Nieuwkoop, N. N., Schmitz, K. S., Rijsbergen, L. C., van Osch, J. A., Dijkhuizen, E., et al. Sars-cov-2 variants of concern partially escape humoral but not t-cell responses in covid-19 convalescent donors and vaccinees. *Science Immunology 6*, 59 (2021).

[22] Goldstein, J. R., and Lee, R. D. Demographic perspectives on the mortality of covid-19 and other epidemics. *Proceedings of the National Academy of Sciences 117*, 36 (2020), 22035–22041.

[23] Gopalakrishnan, V., Pethe, S., Kefayati, S., Srinivasan, R., Hake, P., Deshpande, A., Liu, X., Hoang, E., Davila, M., Bianco, S., et al. Globally local: Hyper-local modeling for accurate forecast of covid-19. *Epidemics 37* (2021), 100510.

[24] Grünwald, P. D. *The minimum description length principle*. MIT press, 2007.

[25] Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., and Wang, C. Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature 584*, 7821 (2020), 420–424.

[26] Havers, F. P., Reed, C., Lim, T., Montgomery, J. M., Klena, J. D., Hall, A. J., Fry, A. M., Cannon, D. L., Chiang, C.-F., Gibbons, A., et al. Seroprevalence of antibodies to sars-cov-2 in 10 sites in the united states, march 23-may 12, 2020. *JAMA internal medicine 180*, 12 (2020), 1576–1586.

[27] Ho, F. K., Petermann-Rocha, F., Gray, S. R., Jani, B. D., Katikireddi, S. V., Niedzwiedz, C. L., Foster, H., Hastie, C. E., Mackay, D. F., Gill, J. M., et al. Is older age associated with covid-19 mortality in the absence of other risk factors? general population cohort study of 470,034 participants. *PloS one 15*, 11 (2020), e0241824.

[28] Hoertel, N., Blachier, M., Blanco, C., Olfson, M., Massetti, M., Rico, M. S., Limosin, F., and Leleu, H. A stochastic agent-based model of the sars-cov-2 epidemic in france. *Nature medicine 26*, 9 (2020), 1417–1421.

[29] Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences 112*, 3 (2015), 719–724.

[30] Irons, N. J., and Raftery, A. E. Estimating sars-cov-2 infections from deaths, confirmed cases, tests, and random surveys. *Proceedings of the National Academy of Sciences 118*, 31 (2021).

[31] Jay, J., Bor, J., Nsoesie, E. O., Lipson, S. K., Jones, D. K., Galea, S., and Raifman, J. Neighbourhood income and physical distancing during the covid-19 pandemic in the united states. *Nature human behaviour 4*, 12 (2020), 1294–1302.

[32] Jones, J. M., Stone, M., Sulaeman, H., Fink, R. V., Dave, H., Levy, M. E., Di Germanio, C., Green, V., Notari, E., Saa, P., et al. Estimated us infection-and vaccine-induced sars-cov-2 seroprevalence based on blood donations, july 2020-may 2021. *JAMA 326*, 14 (2021), 1400–1409.

[33] Kain, M. P., Childs, M. L., Becker, A. D., and Mordecai, E. A. Chopping the tail: How preventing superspreading can help to maintain covid-19 control. *Epidemics 34* (2021), 100430.

[34] Koutra, D., Kang, U., Vreeken, J., and Faloutsos, C. Vog: Summarizing and understanding large graphs. In *Proceedings of the 2014 SIAM international conference on data mining* (2014), SIAM, pp. 91–99.

[35] Kraemer, M. U., Scarpino, S. V., Marivate, V., Gutierrez, B., Xu, B., Lee, G., Hawkins, J. B., Rivers, C., Pigott, D. M., Katz, R., et al. Data curation during a pandemic and lessons learned from covid-19. *Nature Computational Science 1*, 1 (2021), 9–10.

[36] Kraemer, M. U., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science 368*, 6490 (2020), 493–497.

[37] Kustin, T., Harel, N., Finkel, U., Perchik, S., Harari, S., Tahor, M., Caspi, I., Levy, R., Leshchinsky, M., Ken Dror, S., et al. Evidence for increased breakthrough rates of sars-cov-2 variants of concern in bnt162b2-mrna-vaccinated individuals. *Nature medicine 27*, 8 (2021), 1379–1384.

[38] Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., Wesolowski, A., Santillana, M., Zhang, C., Du, X., et al. Effect of non-pharmaceutical interventions to contain covid-19 in china. *Nature 585*, 7825 (2020), 410–413.

[39] Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science 368*, 6490 (2020), 489–493.

[40] Lu, F. S., Nguyen, A. T., Link, N. B., Molina, M., Davis, J. T., Chinazzi, M., Xiong, X., Vespignani, A., Lipsitch, M., and Santillana, M. Estimating the cumulative incidence of covid-19 in the united states using influenza surveillance, virologic testing, and mortality data: Four complementary approaches. *PLOS Computational Biology 17*, 6 (2021), e1008994.

[41] Moghadas, S. M., Fitzpatrick, M. C., Sah, P., Pandey, A., Shoukat, A., Singer, B. H., and Galvani, A. P. The implications of silent transmission for the control of covid-19 outbreaks. *Proceedings of the National Academy of Sciences 117*, 30 (2020), 17513–17515.

[42] Nande, A., Sheen, J., Walters, E. L., Klein, B., Chinazzi, M., Gheorghe, A. H., Adlam, B., Shinnick, J., Tejeda, M. F., Scarpino, S. V., et al. The effect of eviction moratoria on the transmission of sars-cov-2. *Nature communications 12*, 1 (2021), 1–13.

[43] Padmanabhan, P., Desikan, R., and Dixit, N. M. Modeling how antibody responses may determine the efficacy of covid-19 vaccines. *Nature Computational Science 2*, 2 (2022), 123–131.

[44] Pei, S., Kandula, S., and Shaman, J. Differential effects of intervention timing on covid-19 spread in the united states. *Science advances 6*, 49 (2020), eabd6370.

[45] Pei, S., Teng, X., Lewis, P., and Shaman, J. Optimizing respiratory virus surveillance networks using uncertainty propagation. *Nature communications 12*, 1 (2021), 1–10.

[46] Pei, S., Yamana, T. K., Kandula, S., Galanti, M., and Shaman, J. Burden and characteristics of covid-19 in the united states during 2020. *Nature 598*, 7880 (2021), 338–341.

[47] Peluso, M. J., Takahashi, S., Hakim, J., Kelly, J. D., Torres, L., Iyer, N. S., Turcios, K., Janson, O., Munter, S. E., Thanh, C., et al. Sars-cov-2 antibody magnitude and detectability are driven by disease severity, timing, and assay. *Science advances 7*, 31 (2021), eabh3409.

[48] Planas, D., Veyer, D., Baidaliuk, A., Staropoli, I., Guivel-Benhassine, F., Rajah, M. M., Planchais, C., Porrot, F., Robillard, N., Puech, J., et al. Reduced sensitivity of sars-cov-2 variant delta to antibody neutralization. *Nature 596*, 7871 (2021), 276–280.

[49] Prakash, B. A., Vreeken, J., and Faloutsos, C. Spotting culprits in epidemics: How many and which ones? In *2012 IEEE 12th International Conference on Data Mining* (2012), IEEE, pp. 11–20.

[50] Press, W. H., and Levin, R. C. Modeling, post covid-19. *Science 370*, 6520 (2020), 1015–1015.

[51] Reinhart, A., Brooks, L., Jahja, M., Rumack, A., Tang, J., Agrawal, S., Al Saeed, W., Arnold, T., Basu, A., Bien, J., et al. An open repository of real-time covid-19 indicators. *Proceedings of the National Academy of Sciences 118*, 51 (2021).

[52] Russell, T. W., Golding, N., Hellewell, J., Abbott, S., Wright, L., Pearson, C. A., van Zandvoort, K., Jarvis, C. I., Gibbs, H., Liu, Y., et al. Reconstructing the early global dynamics of under-ascertained covid-19 cases and infections. *BMC medicine 18*, 1 (2020), 1–9.

[53] Salomon, J. A., Reinhart, A., Bilinski, A., Chua, E. J., La Motte-Kerr, W., Rönn, M. M., Reitsma, M. B., Morris, K. A., LaRocca, S., Farag, T. H., et al. The us covid-19 trends and impact survey: Continuous real-time measurement of covid-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences 118*, 51 (2021).

[54] Sethuraman, N., Jeremiah, S. S., and Ryo, A. Interpreting diagnostic tests for sars-cov-2. *JAMA 323*, 22 (2020), 2249–2251.

[55] Shaman, J. An estimation of undetected covid cases in france. *Nature 590* (2020), 38–39.

[56] Sood, N., Simon, P., Ebner, P., Eichner, D., Reynolds, J., Bendavid, E., and Bhattacharya, J. Seroprevalence of sars-cov-2–specific antibodies among adults in los angeles county, california, on april 10-11, 2020. *JAMA 323*, 23 (2020), 2425–2427.

[57] Stockmaier, S., Stroeymeyt, N., Shattuck, E. C., Hawley, D. M., Meyers, L. A., and Bolnick, D. I. Infectious diseases and social distancing in nature. *Science 371*, 6533 (2021).

[58] Subramanian, R., He, Q., and Pascual, M. Quantifying asymptomatic infection and transmission of covid-19 in new york city using observed cases, serology, and testing capacity. *Proceedings of the National Academy of Sciences 118*, 9 (2021).

[59] Tian, Y., Sridhar, A., Yağan, O., and Poor, H. V. Analysis of the impact of mask-wearing in viral spread: Implications for covid-19. In *2021 American Control Conference (ACC)* (2021), IEEE, pp. 3132–3137.

[60] Tiwari, S., Vyasarayani, C., and Chatterjee, A. Data suggest covid-19 affected numbers greatly exceeded detected numbers, in four european countries, as per a delayed seiqr model. *Scientific reports 11*, 1 (2021), 1–12.

[61] Wells, C. R., Sah, P., Moghadas, S. M., Pandey, A., Shoukat, A., Wang, Y., Wang, Z., Meyers, L. A., Singer, B. H., and Galvani, A. P. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences 117*, 13 (2020), 7504–7509.

[62] Wilder, B., Charpignon, M., Killian, J. A., Ou, H.-C., Mate, A., Jabbari, S., Perrault, A., Desai, A. N., Tambe, M., and Majumder, M. S. Modeling between-population variation in covid-19 dynamics in hubei, lombardy, and new york city. *Proceedings of the National Academy of Sciences 117*, 41 (2020), 25904–25910.

[63] Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Reingold, A., et al. Substantial underestimation of sars-cov-2 infection in the united states. *Nature communications 11*, 1 (2020), 1–10.

[64] Zhang, W., Govindavari, J. P., Davis, B. D., Chen, S. S., Kim, J. T., Song, J., Lopategui, J., Plummer, J. T., and Vail, E. Analysis of genomic characteristics and transmission routes of patients with confirmed sars-cov-2 in southern california during the early stage of the us covid-19 pandemic. *JAMA network open 3*, 10 (2020), e2024191.

[65] Zhao, J., Yuan, Q., Wang, H., Liu, W., Liao, X., Su, Y., Wang, X., Yuan, J., Li, T., Li, J., et al. Antibody responses to sars-cov-2 in patients with novel coronavirus disease 2019. *Clinical infectious diseases 71*, 16 (2020), 2027–2034.