Information Theoretic Model Selection for Accurately Estimating Unreported COVID-19 Infections

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

[1]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332
[2]Biocomplexity Institute, University of Virginia, Charlottesville, VA 22904
[3]Department of Computer Science, University of Virginia, Charlottesville, VA 22904
[4]Department of Computer Science, The University of Iowa, Iowa City, IA 52242

Abstract

Estimating the true magnitude of infections was one of the significant challenges in combating the COVID-19 outbreak early on. Our inability in doing so allowed unreported infections to drive up disease spread in numerous regions in the US and worldwide. Even today, identifying the true magnitude (the number of total infections) is still challenging, despite the use of surveillance-based methods such as serological studies, due to their costs and biases. This paper proposes an information theoretic approach to estimate total infections accurately. Our approach is built on top of ordinary differential equations based epidemiological models, which have been used extensively in understanding the dynamics of COVID-19, and aims to estimate the true total infections and a parameterization that "best describes" the observed reported infections. Our experiments show that the parameterization learned by our framework leads to a better estimation of total infections and forecasts of the reported infections compared to a "baseline" parameterization, which is learned via usual model calibration. We also demonstrate that our framework can be leveraged to simulate what-if scenarios with non-pharmaceutical interventions. Our results also support earlier findings that most COVID-19 infections were unreported and non-pharmaceutical interventions indeed helped mitigate the COVID-19 outbreak. Our approach gives a general method to use information theoretic techniques to improve epidemic modeling, which can also be applied to other diseases.

*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 Dec 1, 2021, it had already resulted in more than 48 million reported infections and 0.78 million deaths in the United States alone. Worldwide reported infections total at 262 million and deaths at 5.22 million [18]. The devastating effect of COVID-19 extends for the public health sector and the economy as well. For example, in the US, the unemployment rate peaked at 15.8 percent in April 2020 [6], and U.S. GDP contracted at a 3.5% annualized rate for 2020 [1]. Similar economic impacts have been observed worldwide.

One of the significant challenges in combating the COVID-19 pandemic early on was the inability to estimate the true magnitude of infections. As noted by many studies [13, 11, 55, 53, 39], a significant number of COVID-19 infections were unreported, due to various factors such as the lack of testing and asymptomatic infections. In the early stages, transmissions in previously unreported regions were facilitated by unreported infections before being detected. For example, phylogenetic studies revealed that COVID-19 had locally spread in Washington state before active community surveillance was implemented in early 2020 [14]. Similarly, there were only 23 reported infections in five major U.S. cities by March 1, 2020. However, it has been estimated that there were already more than 28000 total infections in those cities by then [5]. Similar trends were observed in Italy, Germany, and the UK [58], and this continues to be a challenge even today [8, 30].

More generally, accurate estimation of the true magnitude of infections is a fundamental epidemiological question and critical for pandemic planning and response. To capture it, we define the *reported rate* as the ratio of reported infections to total infections in this paper. This contrasts with the more extensively studied notion of *case ascertainment rate*, which is the ratio of reported symptomatic infections to the actual number of symptomatic infections [50] (i.e., this doesn't take asymptomatic infections into account). We focus on the reported rate because asymptomatic infections are also infectious and contribute substantially to the community transmission as many studies have shown [56, 41]. Estimation of the reported rate is challenging, and there are limitations in all current methods, as as discuss next.

One of the most effective methods to identify the reported rate in a region is through large-scale serological studies [54, 26, 62]. 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 (i.e., 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]. It is even more challenging to perform them in real-time as there are unavoidable delays between sample collection and laboratory tests [2, 26]. Serological studies can also suffer from sampling biases, and heuristics need to be designed to account for them [9].

Other lines of work to estimate the reported rate for COVID-19 exploits existing influenza surveillance systems to estimate symptomatic infections [40], due to symptomatic similarities between the two diseases. However, they also suffer from 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 through epidemiological models [39, 50, 12, 48, 36, 38, 60, 25, 33, 61]. They calibrate these models to reported data (say the time series of reported infections D_{reported}) to learn parameterizations and use them to estimate the reported rate. For an epidemiological model, denoted by O_{M}, we refer to a parameterization learned from such model

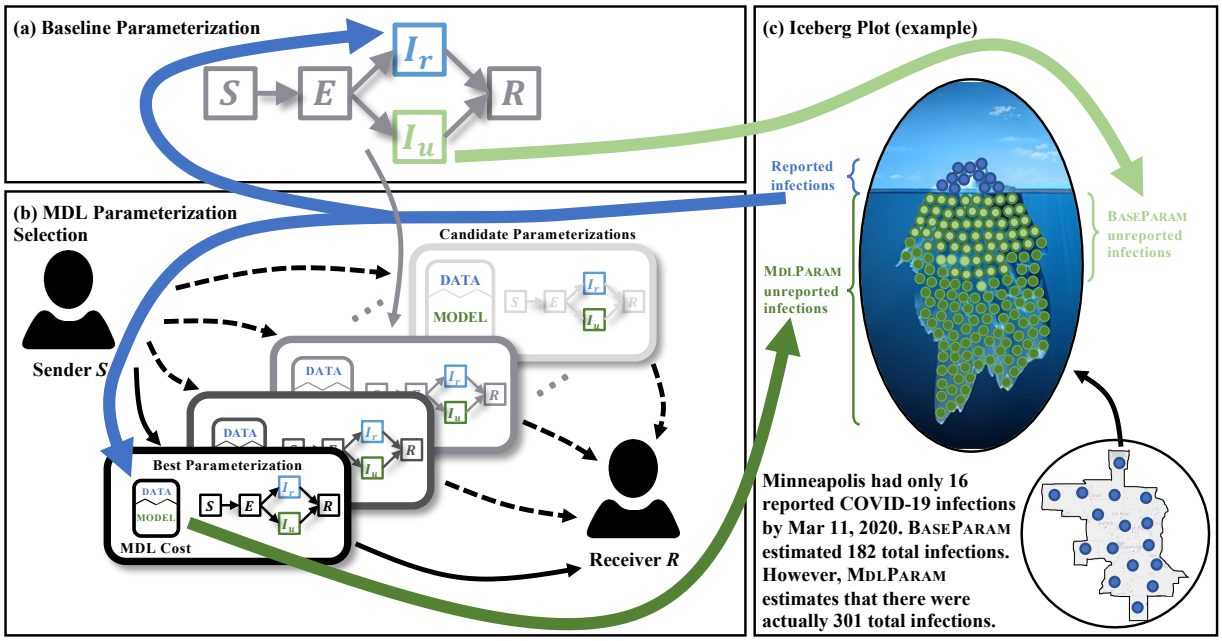


Figure 1: The workflow of MdlInfer. (a) The baseline parameterization underestimates the unreported infections (Shown in light green in the iceberg plot in (c)). (b) Our MDL model selection approach MdlInfer estimates the magnitude of COVID-19 infections accurately by inferring unreported infections given the reported infections. In MdlInfer, the Data is the reported infections D_{reported}, which a hypothetical Sender S wants to transmit to a Receiver R. At a high level, the sender S will send the Data by sending the total infections D, the "baseline parameterization" $\mathbf{p}$, and the "candidate" parameterization $\mathbf{p'}$. (c) MdlParam reveals that a large majority of COVID-19 infections were unreported. We visualize the estimated reported rate using the iceberg. The height of the iceberg over water (in white) and the height below the water (in blue) are proportional to the reported and unreported infections estimated using MdlInfer. The number of COVID-19 reported infections is just the tip of an iceberg, while the whole iceberg corresponds to the total infections estimated by MdlInfer.

calibration as a "baseline parameterization" $\mathbf{p}$, and the estimated reported rate as α_{reported}; later we will consider specific choices of O_{M} from recent work. However, this "model estimated" reported rate is still problematic because of ad-hoc calibrating assumptions such as predefining a set of parameters [17] or heuristics for parameter estimation (e.g., limiting the search space to a narrow range [19]). As we will show later in our results, minor errors in estimates of reported rates may lead to very different forecasts of future trends and intervention policy recommendations. Hence a principled approach is needed in order to estimate the actual reported rate.

To achieve this goal, we propose a new information theory-based approach. For simplicity, let us assume the reported rate to be a constant (we discuss how we handle time-varying rates in the next section). Our central intuition is: imagine we are given D, a time series consisting of the accurate values of total infections within a given time period. By calibrating the epidemiological model O_{M} against D, the model can learn a "candidate" parameterization $\mathbf{p'}$ and use it to calculate the corresponding reported rate $\alpha'_{\text{reported}}$. Intuitively, this parameterization $\mathbf{p'}$ should be better than the baseline parameterization $\mathbf{p}$ because it is done using the total infections as well not just the reported ones. Using this D and $\alpha'_{\text{reported}}$, we should be able to consequently also "describe"

the reported infections D_{reported} "better" (as D_{reported} should be just $\alpha'_{\text{reported}} \times D$). One way to measure the quality of this description is to compute the number of bits needed to encode it. Hence our problem is one of developing a principled framework to choose among competing descriptions (i.e., perform "model selection") such that we can find the one that takes the fewest number of bits. So we develop an information theoretic framework to estimate the actual total infections - by posing an optimization problem which (roughly) aims to find the D which minimizes the number of bits required to encode the observed D_{reported}.

Following the discussion above, to encode the D_{reported}, we intuitively need the variables D and $\alpha'_{\text{reported}}$. Let us denote $\textsc{Data} = D_{\text{reported}}$ and $\textsc{Model} = (D, \alpha'_{\text{reported}})$. Note that this Model is meant for encoding purposes only and is different from the base epidemiological ODE model O_M. The optimization problem then is defined over all possible values of Model to encode Data efficiently. Following the convention in information theory, we use $L(\textsc{Model})$ to denote the number of bits required to encode the Model; and $L(\textsc{Data}|\textsc{Model})$ 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 $\textsc{Model}^*$, which minimizes $L(\textsc{Model})+L(\textsc{Data}|\textsc{Model})$. To be more specific, we in fact set our $\textsc{Model} = (D, \mathbf{p'}, \mathbf{p})$ in our complete formulation (see Methods section for details).

The optimization framework described above follows the popular two-part Minimum Description Length (MDL) framework which follows the general maxim of 'induction by compression'. The two-part MDL (aka sender-receiver framework) consists of hypothetical actors S and R. Sender S has 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. Note that the two-part MDL (and MDL in general) does not assume the nature of the Data or the Model, which exactly fits our objective. It 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 [47]. However, our technique is the first to propose an MDL-based approach on top of ODE-based epidemiological models, which are harder to formulate and optimize.

We call our MDL-based optimization framework MdlInfer. Our experiments showcase our method using two different ODE-based epidemiological models as O_{M}: SAPHIRE model [25] and SEIR+HD model [33]. They both have previously performed well in fitting reported infections and highlighted beneficial insights for the COVID-19 response. The SAPHIRE model focuses on two key features of the outbreak: high covertness and high transmissibility that drove the outbreak of COVID-19 in Wuhan. The SEIR+HD model investigates how non-pharmaceutical interventions like social distancing will be needed to maintain epidemic control. These models are representative enough to show that MdlInfer gives consistent performance with multiple epidemiological models with different dynamics. As we will show later, we do not estimate an accurate reported rate and total infections in both models without our framework. We use BaseParam to represent the baseline parameterization $\mathbf{p}$ obtained via standard model calibration. Similarly, we use MdlParam to represent the optimal parameterization identified by MdlInfer. Our experiments show that MdlParam is clearly superior to BaseParam in estimating total infections. An example is shown in Figure 1. By March 11, 2020, the Minneapolis Metro Area had only 16 COVID-19 reported infections. The best version of BaseParam estimated 182 total infections (colored as light green in the iceberg). On the other hand, our MdlParam gave 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, MdlParam also leads to better a fit and *future* projections on reported infections. We also demonstrate that MdlParam can aid policy making by analyzing

counter-factual non-pharmaceutical interventions, while inaccurate BaseParam estimates lead to wrong non-pharmaceutical intervention conclusions.

Results

Here, we present our empirical findings on a set of experiments at different geographical regions and time periods. We choose these regions and periods based on the severity of outbreak and the availability of the serological studies and symptomatic surveillance dataset. As mentioned above, we use the SAPHIRE model [25] and SEIR+HD model [33] as the base epidemiological models O_{M}. We show consistently good performance for MdlInfer on both models.

In each region, we divide the timeline into two time periods: (i) observed period, when only the number of reported infections are available, and the models are calibrated to learn the parameterizations, and (ii) future period, where we evaluate the forecasts generated by the epidemiological models and calibrated 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 rates for each sub period separately.

(A) Estimating Total Infections: MdlParam estimates total infections more accurately than BaseParam

Here, we use the point estimates of the total infections calculated from serological studies as the ground truth (black dots shown in Figure 2). We call it $\textsc{SeroStudy}_{\mathrm{Tinf}}$. We also plot MdlParam's estimation of total infections, $\textsc{MdlParam}_{\mathrm{Tinf}}$, in the same figure (red curve). To compare the performance of our approach and the baselines with $\textsc{SeroStudy}_{\mathrm{Tinf}}$, we use the cumulative value of estimated total infections. However, serological studies numbers are not directly comparable with the total infections because of the lag between antibodies becoming detectable and infections being reported [2, 26]. In Figure 2, we have already accounted for this lag following the CDC serological studies [2, 26] (See Methods section for details). The vertical black line shows a 95% confidence interval for $\textsc{SeroStudy}_{\mathrm{Tinf}}$. The blue curve represents BaseParam's estimation of total infections, $\textsc{BaseParam}_{\mathrm{Tinf}}$. As seen in the figure, $\textsc{MdlParam}_{\mathrm{Tinf}}$ falls within the confidence interval of the estimates given by serological studies. Significantly, in Figure 2 (c) and (d) for South Florida, the BaseParam for the SAPHIRE model [25] overestimates the total infections, while the SEIR+HD model underestimates the total infections. However, MdlParam consistently predicts the total infections correctly. This observation shows that as necessary, $\textsc{MdlParam}_{\mathrm{Tinf}}$ can improve upon the $\textsc{BaseParam}_{\mathrm{Tinf}}$ in either direction (i.e., by increasing or decreasing the total infections). Note that the $\textsc{MdlParam}_{\mathrm{Tinf}}$ curve from both the models are similar even when the $\textsc{BaseParam}_{\mathrm{Tinf}}$ curves are different. This shows 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 $\textsc{SeroStudy}_{\mathrm{Tinf}}$ and $\textsc{BaseParam}_{\mathrm{Tinf}}$. We then compute the same between $\textsc{SeroStudy}_{\mathrm{Tinf}}$ and $\textsc{MdlParam}_{\mathrm{Tinf}}$. We then compute the ratio, ρ_{Tinf}, of the two RMSE errors as $\frac{\textsc{Rmse}(\textsc{BaseParam}_{\mathrm{Tinf}},\textsc{SeroStudy}_{\mathrm{Tinf}})}{\textsc{Rmse}(\textsc{MdlParam}_{\mathrm{Tinf}},\textsc{SeroStudy}_{\mathrm{Tinf}})}$. Note that the values of ρ_{Tinf} being greater than 1 imply that the $\textsc{MdlParam}_{\mathrm{Tinf}}$ is closer to $\textsc{SeroStudy}_{\mathrm{Tinf}}$ estimates than $\textsc{BaseParam}_{\mathrm{Tinf}}$. In Figure 2 (e), we plot ρ_{Tinf}. Overall, the ρ_{Tinf} values are greater in Figure 2 (a) to (d) (1.20, 2.62, 5.46, and 1.22), which indicates that MdlParam performs better overall. Note that even when the value of ρ_{Tinf} is 1.20, the improvement made by $\textsc{MdlParam}_{\mathrm{Tinf}}$ over $\textsc{BaseParam}_{\mathrm{Tinf}}$ in terms of RMSE is about **12091**. Hence, one can conclude that $\textsc{MdlParam}_{\mathrm{Tinf}}$ is indeed superior to $\textsc{BaseParam}_{\mathrm{Tinf}}$, when it comes to estimating total number of infections.

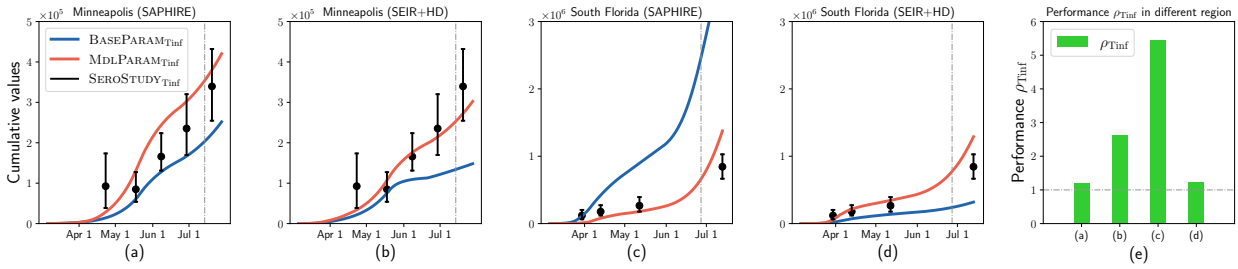


Figure 2: BaseParam's and MdlParam's estimation of total infections. (a)-(d) MdlParam's estimation of total infections is more accurate than BaseParam's. The grey dash line divides the observed and future periods (which was not accessible to the model while calibrating). The blue and red curves represent BaseParam's estimation of total infections, BaseParam$_{\text{Tinf}}$, and MdlParam's estimation of total infections, MdlParam$_{\text{Tinf}}$, respectively. The black point estimates and confidence intervals represent the total infections estimated by serological studies [2, 26], or SeroStudy$_{\text{Tinf}}$. Note that the plot corresponds to different geographic regions for SAPHIRE and SEIR+HD epidemiological models, and hence the scales differ. (e) The performance metric, ρ_{Tinf}, comparing MdlParam$_{\text{Tinf}}$ against BaseParam$_{\text{Tinf}}$ is shown for the regions in (a)-(d). Here, the values of ρ_{Tinf} are 1.20, 2.62, 5.46, and 1.22, implying that MdlParam performs better in the estimation of total infections than BaseParam for different geographic regions and epidemiological models.

(B) Estimating Reported Infections: MdlParam leads to better fit and projection than BaseParam at different stages of the COVID-19 epidemic

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

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 MdlParam$_{\text{Rinf}}$ against BaseParam$_{\text{Rinf}}$ in a manner similar to above. In Figure 3 (e) and (j), we plot the ρ_{Rinf} for the observed and the future period. In both (e) and (j), we notice that the ρ_{Rinf} is close to or greater than 1. This further shows that MdlParam$_{\text{Rinf}}$ has a better or at least closer fit for reported infections than BaseParam$_{\text{Rinf}}$. Additionally, the ρ_{Rinf} for the future period is even greater than ρ_{Rinf} for the observed period, which shows that MdlParam$_{\text{Rinf}}$ performs even better in forecast than BaseParam$_{\text{Rinf}}$.

Note that Figure 3 (a) to (d) correspond to the early state of the COVID-19 epidemic in spring and summer 2020, and Figure 3 (f) to (i) correspond to fall 2020. Here, the value of ρ_{Rinf} is greater than 1 indicating that MdlParam$_{\text{Rinf}}$ outperforms BaseParam$_{\text{Rinf}}$ and MdlInfer performs well in estimating temporal patterns at different stages of the COVID-19 epidemic.

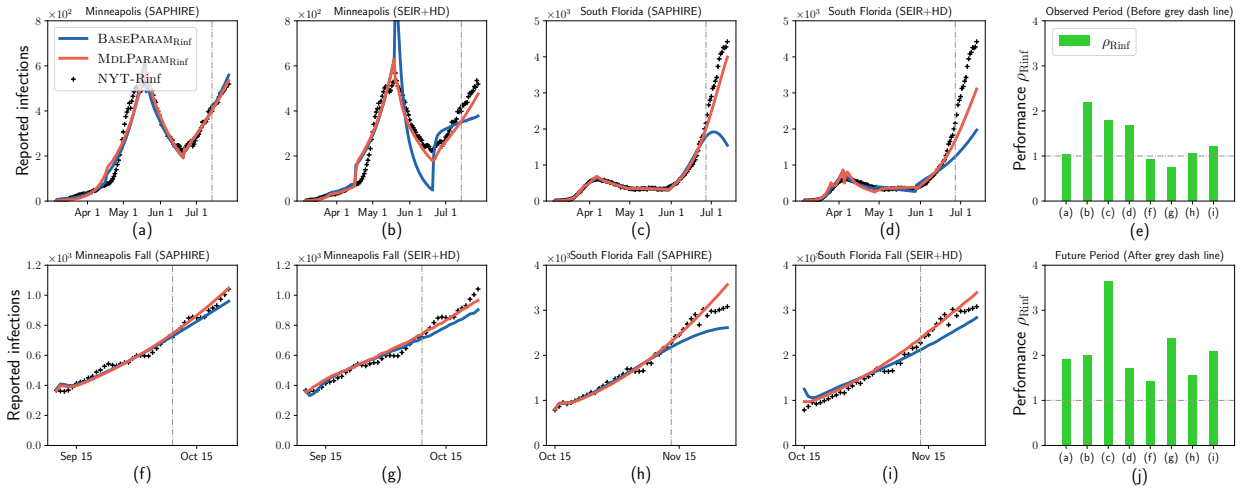


Figure 3: BaseParam's and MdlParam's estimation of reported infections. (a)-(d) and (f)-(i) MdlParam's estimation of reported infections has better fit and projection than BaseParam's. The grey dash line divides the observed period and future period. Black plus symbols represent the infections reported by the New York Times (NYT-Rinf). The blue curve represents BaseParam's estimation of reported infections BaseParam$_{\text{Rinf}}$. Similarly, red curve represents MdlParam's estimation of reported infections MdlParam$_{\text{Rinf}}$. Each plot corresponds to a different geographic region for the SAPHIRE model or SEIR+HD model with different scales. The results show that MdlParam$_{\text{Rinf}}$ aligns closer with NYT-Rinf than BaseParam$_{\text{Rinf}}$. Besides, MdlParam$_{\text{Rinf}}$ has a better projection for the future period than BaseParam$_{\text{Rinf}}$. (e) and (j) The performance metric, ρ_{Rinf}, comparing MdlParam$_{\text{Rinf}}$ against BaseParam$_{\text{Rinf}}$, is shown for the regions for both the observed period (e), and the future period (j). MdlParam$_{\text{Rinf}}$ is close to or better than BaseParam$_{\text{Rinf}}$ in both the observed and future period.

(C) Estimating Symptomatic Rate: MdlParam estimates the symptomatic rate more accurately than BaseParam

We validate this observation using Facebook's symptomatic surveillance dataset [49]. Here we plot the estimated symptomatic rate over time and overlay the estimates and standard error from the symptomatic surveillance data (See Figure 4). The red and blue curves are the MdlParam's and BaseParam's estimation of symptomatic rates, MdlParam$_{\text{Symp}}$ and BaseParam$_{\text{Symp}}$ respectively. Note that the SAPHIRE model does not contain states corresponding to the symptomatic infections. Therefore, we only focus on the SEIR+HD model. We compare the trends of the MdlParam$_{\text{Symp}}$ and BaseParam$_{\text{Symp}}$ with the symptomatic surveillance results. We focus on trends rather than actual values because the symptomatic rate numbers could be biased [49] (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 the figure, the MdlParam$_{\text{Symp}}$ captures the trends of the surveyed symptomatic rate Rate$_{\text{Symp}}$ (represented by black plus symbols) much better than BaseParam$_{\text{Symp}}$.

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

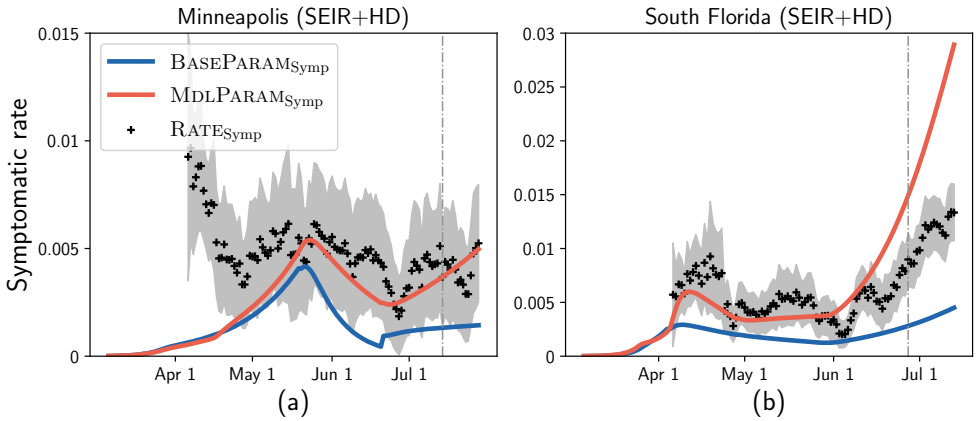


Figure 4: MdlParam's estimation of symptomatic rate is more accurate than BaseParam's. The blue and red curves represent the BaseParam's estimation of symptomatic rate ($\text{BaseParam}_{\text{Symp}}$), and MdlParam's estimation of symptomatic rate ($\text{MdlParam}_{\text{Symp}}$) respectively. The 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 [49, 51]).

Evaluating the effect of Non-pharmaceutical Interventions

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

We computed the cumulative reported rate $\text{MdlParam}_{\text{Rate}}$ measured by the ratio of the cumulative value of reported infections to the total infections estimated by MdlParam over time and plotted it in Figure 5 (a). The figure shows that the $\text{MdlParam}_{\text{Rate}}$ 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 the earlier study [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 plotted the simulated reported infections of MdlParam in Figure 5 (c) (red curve). We then repeated the simulation of reported infections for 5 different scenarios: (i) isolate just the reported infections, (ii) just the symptomatic infections, 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 the figure, 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 decreasing. On the other hand, non-pharmaceutical interventions for some fraction of asymptomatic and presymptomatic infections decrease reported infections. Thus, we can conclude that non-pharmaceutical interventions on asymptomatic infections are essential in controlling the COVID-19 epidemic.

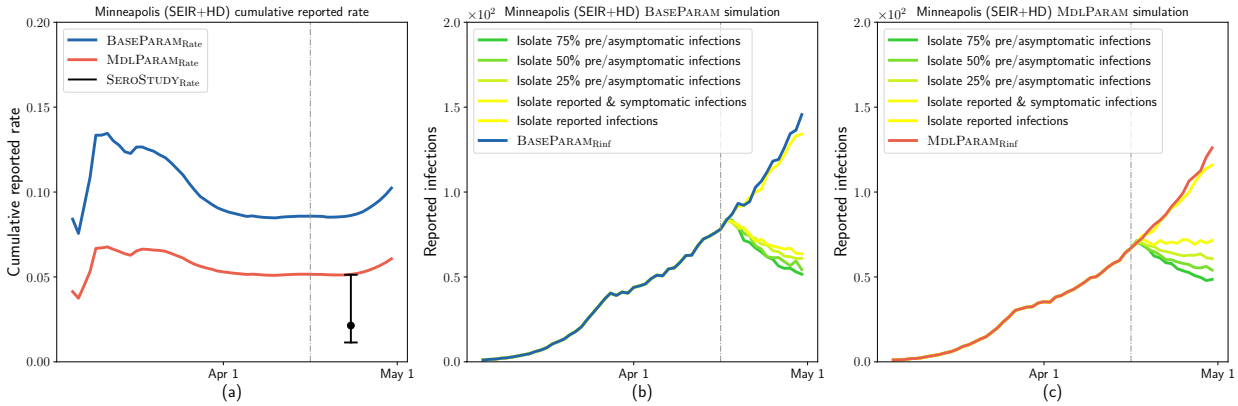


Figure 5: Cumulative reported rate and non-pharmaceutical intervention simulation results. (a) The cumulative reported rate trend during COVID-19 outbreak. The blue and red curve represent the cumulative reported rate of BaseParam, $\text{BaseParam}_{\text{Rate}}$, and of MdlParam, $\text{MdlParam}_{\text{Rate}}$, respectively. The black point estimate and its confidence interval represent the reported rate $\text{SeroStudy}_{\text{Rate}}$ estimated by serological studies [2, 26]. (b) The grey dash line divides the observed period and future period. The blue curve represents the BaseParam's estimation of reported infections, $\text{BaseParam}_{\text{Rinf}}$. The other five curves represent the simulated reported infections for the 5 scenarios described in the Results section. (c) Non-pharmaceutical interventions on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic. The red curve represents the MdlParam's estimation of reported infections, $\text{MdlParam}_{\text{Rinf}}$. The other five curves represent the simulated reported infections for the same 5 scenarios as in (b).

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

Here, we also plot the simulated reported infections generated of BaseParam in Figure 5 (b) (blue curve). As seen in the figure, based on BaseParam, 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 total infections based on the optimal parameterization of epidemiological models. Our approach leverages the information theoretic Minimum Length Description (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 of epidemiological models. MdlInfer employs a principled method that selects the total infections which "describes" the reported infections best. At the same time, the existing epidemiological models [33, 39, 43, 25] typically rely on the baseline parameterization learned by ad-hoc heuristics. The MdlInfer framework can also be adapted to work on a set of calibrated parameterizations to generate uncertainty estimates.

Overall, our results show that MdlInfer estimates total infections at various geographical

locations and different epidemiological models more accurately than the baseline parameterization from both directions, i.e., it corrects both over- and under-estimates. For example, compared to the baseline parameterization, we correctly estimate 55719 more infections by April 1 for the SEIR+HD model (Figure. 2 (d)), and 87636 fewer infections for the SAPHIRE model (Fig. 2 (c)) for South Florida. We also show that MdlInfer leads to a better fit of the reported infections in the observed period and more accurate forecasts for the future period than the baseline parameterization. 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 a more accurate symptomatic rate than the baseline parameterization. Additionally, our results show consistent performance with respect to the reported infections and serological studies on both the SAPHIRE and SEIR+HD models.

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, 46, 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 future work, it may be helpful to adapt MdlInfer to measure the quality of the base 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, 57, 17, 44], are more suitable in some settings. It would be interesting to extend MdlInfer to incorporate such models. Finally, there is significant population heterogeneity in disease outcomes [15, 31], e.g., differences in severity rate or mortality rate, when infected with COVID-19, for different age group [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 theoretic frameworks.

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, 45], 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 $\textsc{SeroStudy}_{\text{Tinf}}$ for the past 6 months (i.e., from the beginning of the pandemic since January 2020). However, this $\textsc{SeroStudy}_{\text{Tinf}}$ 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 [63, 52] and (ii) the 6-14 range period for specimen collection as mentioned before. To account for this, we compare the $\textsc{SeroStudy}_{\text{Tinf}}$ numbers with the model estimated total infections of 7 days prior to the first day of specimen collection period (as also suggested by the CDC serological studies work [26]).

3. **Symptomatic surveillance [49, 51]:** This dataset consists of point estimate $\textsc{Rate}_{\text{Symp}}$ (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 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 [49]. Hence we focus on trends instead of the exact numbers. As the original symptomatic surveillance data is at a county level, we weighted the numbers using the total population of each county to compute the $\textsc{Rate}_{\text{Symp}}$.

Our Approach

Minimum Length Description Principle

We formulate the problem of estimating total COVID-19 infections using the Minimum Length Description (MDL) principle [10, 24]. We use the 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 the 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 (the total number of bits needed) is composed of two parts: (i) model cost $L(\textsc{Model})$: The cost in bits of encoding the Model and (ii) data cost $L(\textsc{Data}|\textsc{Model})$: 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, $L(\textsc{Model})$, and $L(\textsc{Data}|\textsc{Model})$, find $\textsc{Model}^*$ such that

$$\textsc{Model}^* = \arg\min_{\textsc{Model}} L(\textsc{Model}) + L(\textsc{Data}|\textsc{Model}) \tag{1}$$

In our situation, the Data is the reported COVID-19 infections D_{reported}: it is the only real-world data given to us (indeed 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 from D. To compute

$\alpha'_{\text{reported}}$, we need the candidate parameterization **p'** (after O_M is calibrated using D). Further, we choose to also add **p**, making our MODEL to be $(D, \mathbf{p'}, \mathbf{p})$. There are alternative MODELs that can be considered, but we choose this MODEL $= (D, \mathbf{p'}, \mathbf{p})$ and explain more in the Supplementary Information. Next, we will introduce how we formulate our problem of estimating total infections using the MDL principle.

MDL Formulation

First, we need to introduce some notations used for MDL Formulation. By running the epidemiological model with **p**, O_{M} will output reported infections. However, this output is an estimate of the reported infections by O_{M} rather than the exact D_{reported}. This is because even though the model may have been calibrated using D_{reported}, the calibration cannot be perfect, and there will be differences between these estimated reported infections and the D_{reported}. Therefore, we term this estimated reported infections as $D_{\text{reported}}(\mathbf{p})$. Similarly, we have the $D_{\text{reported}}(\mathbf{p'})$ and estimated total infections $D(\mathbf{p'})$ for **p'**. As described in the introduction section, we can also calculate the estimated reported rate α_{reported} and $\alpha'_{\text{reported}}$ using **p** and **p'**. With these notations, next we will formulate the space of all possible MODEL and give the equation for the cost in bits of encoding MODEL and DATA.

The Space of MODEL

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

Model Cost

With MODEL $= (D, \mathbf{p'}, \mathbf{p})$, we conceptualize the model cost by imagining that the sender S will send the MODEL $= (D, \mathbf{p'}, \mathbf{p})$ to the receiver R in three parts: (i) first send the **p** (sent by encoding **p** directly) (ii) next send the **p'** given **p** (sent by encoding $\mathbf{p'} - \mathbf{p}$) and (iii) then send D given **p'** and **p** (sent by encoding $\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\mathbf{p})$). We term the model cost as $L(D, \mathbf{p'}, \mathbf{p})$, with three components: COST$(\mathbf{p})$, COST$(\mathbf{p'}|\mathbf{p})$, and COST$(D|\mathbf{p'}, \mathbf{p})$. Hence,

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

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 $= (D, \mathbf{p'}, \mathbf{p})$, we send DATA by encoding $\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\mathbf{p'})$. Since D, **p** and **p'** have already been sent, the receiver R could recover back to D_{reported} from $\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\mathbf{p'})$. We term data cost as $L(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p})$ and formulate it as Equation 3.

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

Total Cost

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

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

Problem Statement

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

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

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 $L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p})$ 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 the D in $L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p})$ to help reduce the search space. Here, we formulate step 1 algorithm as Equation 6.

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

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 $L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p})$. 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] solve this constrained optimization problem for D^*. Here, we formulate step 2 algorithm as Equation 7.

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

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

Calibrate Procedure

Our idea is to use the baseline parameterization **p** and candidate parameterization **p'** for the MDL formulation. To learn the parameterization **p** and **p'**, we need a calibrate procedure to learn them. Typically, there are many calibration procedures in literature [29, 59], and any calibrate procedure is acceptable. Here, we assume the existence of the calibrate procedure for O_{M} and term it as CALIBRATE. The CALIBRATE takes two inputs: O_{M} and the dataset. We further assume the dataset contains at least D_{reported}. This CALIBRATE uses loss function to learn parameterizations, usually root mean squared error (RMSE) [23] or log likelihood [33, 25] loss functions. Here, the baseline parameterization **p** is learned as Equation 8

$$\mathbf{p} = \textsc{Calibrate}(O_{\mathrm{M}}, \{D_{\text{reported}}, \text{others}\}) \tag{8}$$

We are also trying to find a better candidate parameterization **p'** by calibrating on D (and D_{reported}). Our idea is to calibrate **p'** on D_{reported} and the daily sequence of unreported infections ($D_{\text{unreported}}$). We use the same CALIBRATE procedure but just add another term on $D_{\text{unreported}}$ in the loss function. The candidate parameterization **p'** is learned as Equation 9.

$$\mathbf{p'} = \textsc{Calibrate}(O_{\mathrm{M}}, \{D_{\text{reported}}, D_{\text{unreported}}, \text{others}\}) \tag{9}$$

SEIR+HD model and SAPHIRE model

Next, we describe how we use the two epidemiological models: SEIR+HD model and SAPHIRE model. As described in the main article, we use BASEPARAM and MDLPARAM to represent baseline parameterization and optimal parameterization identified by MDLINFER respectively. Note that as our MDLINFER is a principled and general framework, other epidemiological models or calibration procedures are also acceptable.

SEIR+HD model

The SEIR+HD model [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 calibration procedure described in Kain et al [33] only infers 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. They assume the importations only happen at the beginning of the pandemic (captured by E_0), and the total population N remains constant. Note that our work can be also extended to time-varying importations as well easily. We also extend the calibration procedure to infer two more parameters: α (proportion of asymptomatic infections) and α_1 (proportion of new symptomatic infections that are reported). We compute the newly reported infections and unreported infections as follows:

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

2. New unreported infections $= (1 - \alpha_1) \times (dI_PI_S + dI_PI_M) + dEI_A$.

SAPHIRE model

The SAPHIRE model [25] consists of 7 states: Susceptible S, exposed E, pre-symptomatic P, ascertained infectious I, unascertained infectious A, hospitalized D, and recovered R. The calibration procedure [25] only infers the transmission rate β and reported rate r while keeping other parameters fixed. We also compute the newly reported infections and unreported infections as follows:

1. New reported infections $= \frac{\alpha P}{D_p}$: 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. α is the reported rate estimated by the epidemiological model.

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

BaseParam and MdlParam

Finally, we describe how we get the results from BaseParam and MdlParam. Here, we use the baseline parameterization **p** and BaseParam as examples (this can also be repeated for **p'** and MdlParam). Using both **p** and the epidemiological model O_{M}, we can calculate the BaseParam's estimation of total infections $\text{BaseParam}_{\text{Tinf}}$ as the cumulative values of $D(\mathbf{p})$ from pandemic's beginning. $D_{\text{reported}}(\mathbf{p})$ can be directly used as the BaseParam's estimation of reported infections. For the cumulative reported rate $\text{BaseParam}_{\text{Rate}}$, we calculate it as the cumulative values of NYT-Rinf divided by $D(\mathbf{p})$. For the symptomatic rate, the SEIR+HD model [33] could estimate the number of symptomatic rate $\text{BaseParam}_{\text{Symp}}$ 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 partially supported 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), 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).

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] 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.

[18] 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.

[19] Edeling, W., Arabnejad, H., Sinclair, R., Suleimenova, D., Gopalakrishnan, K., Bosak, B., Groen, D., Mahmood, I., Crommelin, D., and Coveney, P. V. The impact of uncertainty on predictions of the covidsim epidemiological code. *Nature Computational Science 1*, 2 (2021), 128–135.

[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., Bretó, C., and King, A. A. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences 103*, 49 (2006), 18438–18443.

[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] 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.

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

[45] 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.

[46] 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.

[47] 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.

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

[49] 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).

[50] 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.

[51] 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).

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

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

[54] 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.

[55] 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).

[56] 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).

[57] 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.

[58] 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.

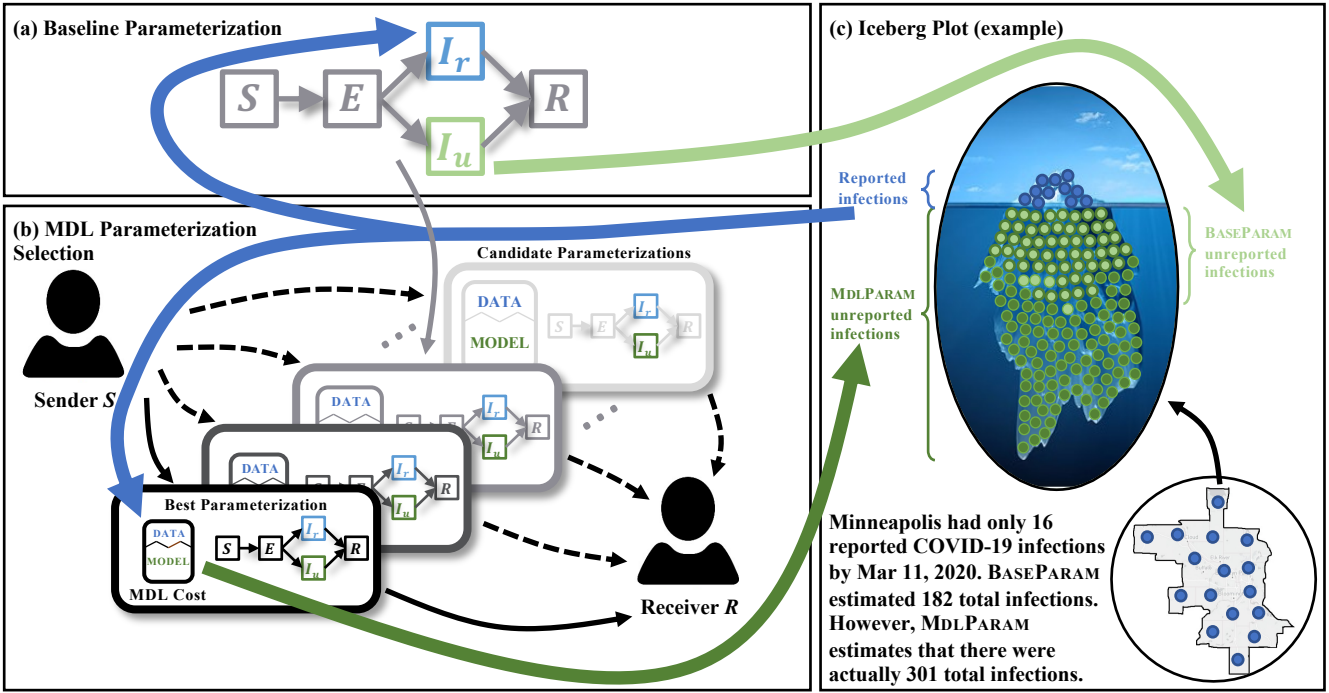
[59] Toth, D. J., Tanner, W. D., Khader, K., and Gundlapalli, A. V. Estimates of the risk of large or long-lasting outbreaks of middle east respiratory syndrome after importations outside the arabian peninsula. *Epidemics 16* (2016), 27–32.

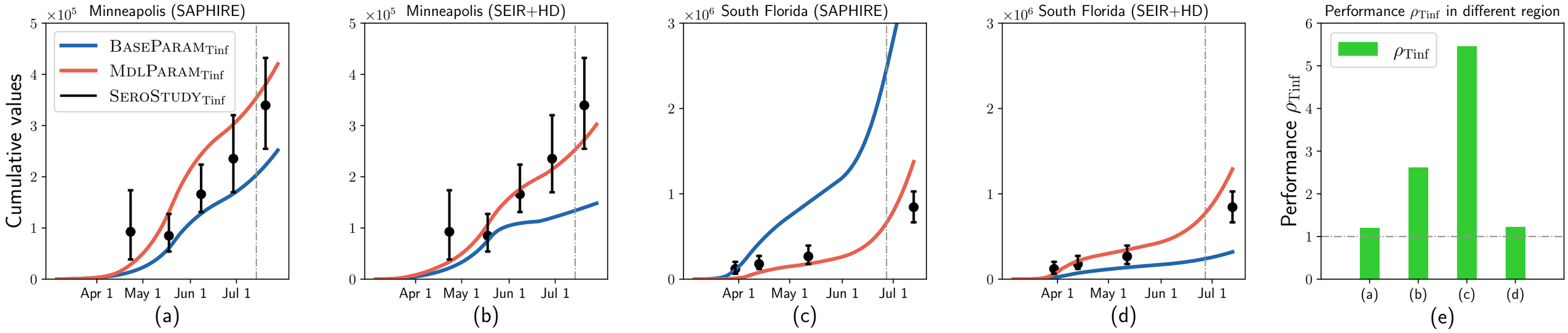
[60] Wells, C. R., Sah, P., Moghadas, S. M., Pandey, A., Shoukat, A., Wang, Y., Wang, Z., Meyers, L. A., Singer, B. H., and Galvani, A. P. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences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

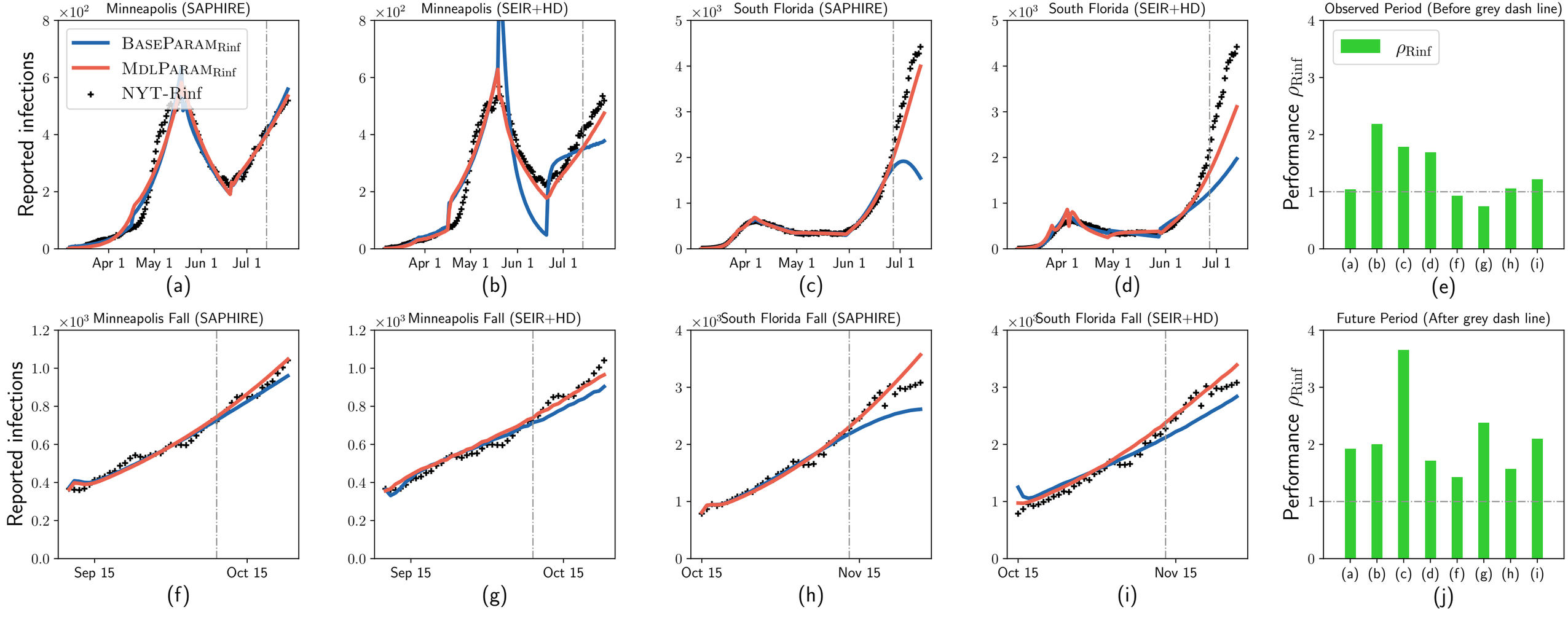
[61] Wilder, B., Charpignon, M., Killian, J. A., Ou, H.-C., Mate, A., Jabbari, S., Perrault, A., Desai, A. N., Tambe, M., and Majumder, M. S. Modeling between-population variation in covid-19 dynamics in hubei, lombardy, and new york city. *Proceedings of the National Academy of Sciences 117*, 41 (2020),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

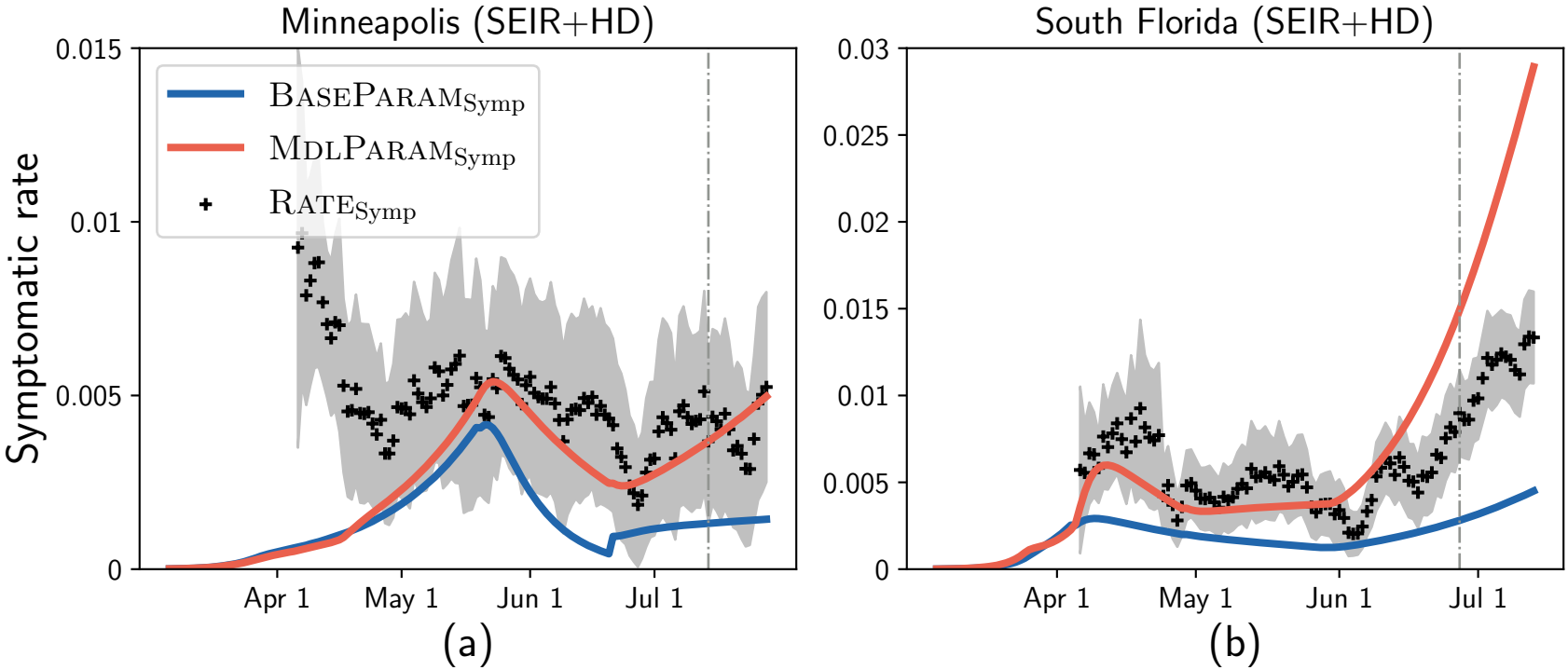
[62] Zhang, W., Govindavari, J. P., Davis, B. D., Chen, S. S., Kim, J. T., Song, J., Lopategui, J., Plummer, J. T., and Vail, E. Analysis of genomic characteristics and transmission routes of patients with confirmed sars-cov-2 in southern california during the early stage of the us covid-19 pandemic.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

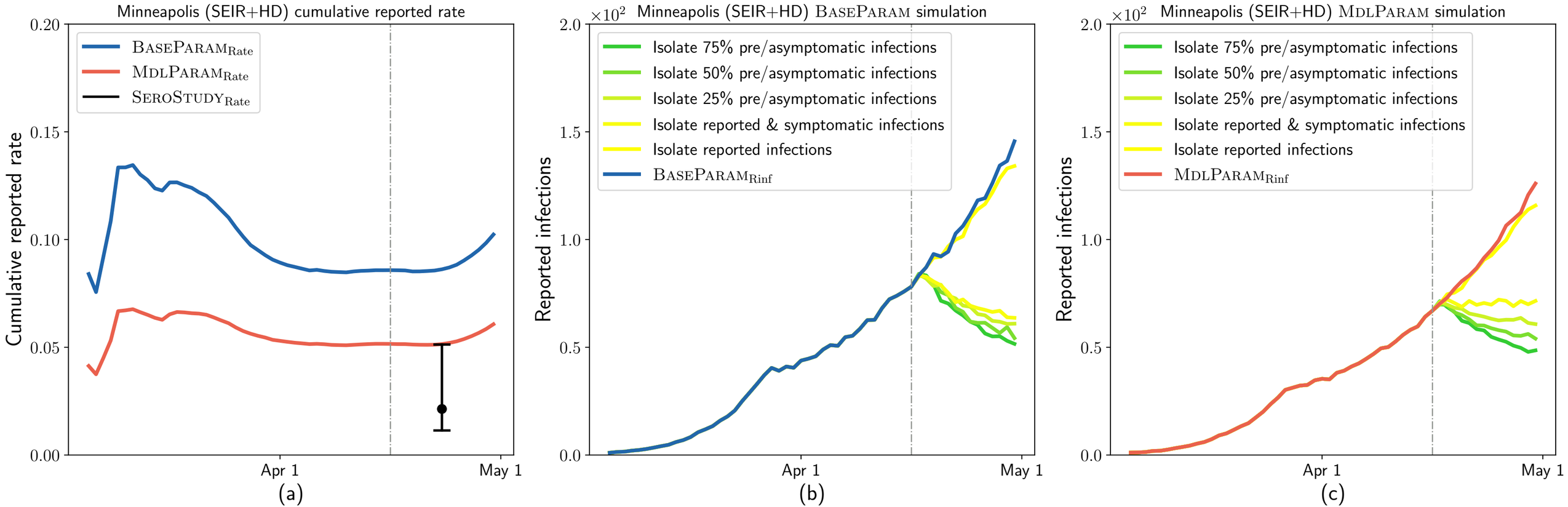
[63] Zhao, J., Yuan, Q., Wang, H., Liu, W., Liao, X., Su, Y., Wang, X., Yuan, J., Li, T., Li, J., et al. Antibody responses to sars-cov-2 in patients with novel coronavirus disease 2019. *Clinical infectious diseases 71*, 16 (2020), 2027–2034.


(a) Baseline Parameterization
S
E
I_r
I_u
R
(b) MDL Parameterization Selection
Candidate Parameterizations
DATA
MODEL
Sender S
Best Parameterization
MDL Cost
Receiver R
(c) Iceberg Plot (example)
Reported infections
MDLPARAM unreported infections
BASEPARAM unreported infections
Minneapolis had only 16 reported COVID-19 infections by Mar 11, 2020. BASEPARAM estimated 182 total infections. However, MDLPARAM estimates that there were actually 301 total infections.


Minneapolis (SAPHIRE)
Minneapolis (SEIR+HD)
South Florida (SAPHIRE)
South Florida (SEIR+HD)
Performance ρ_{Tinf} in different region
Cumulative values
Performance ρ_{Tinf}
$\times 10^5$
$\times 10^6$
BaseParam$_{\mathrm{Tinf}}$
MdlParam$_{\mathrm{Tinf}}$
SeroStudy$_{\mathrm{Tinf}}$
ρ_{Tinf}
Apr 1
May 1
Jun 1
Jul 1
(a)
(b)
(c)
(d)
(e)


Minneapolis (SAPHIRE)
Minneapolis (SEIR+HD)
South Florida (SAPHIRE)
South Florida (SEIR+HD)
Observed Period (Before grey dash line)
Minneapolis Fall (SAPHIRE)
Minneapolis Fall (SEIR+HD)
South Florida Fall (SAPHIRE)
South Florida Fall (SEIR+HD)
Future Period (After grey dash line)
Reported infections
Performance ρ_{Rinf}
BaseParam$_{\text{Rinf}}$
MdlParam$_{\text{Rinf}}$
NYT-Rinf
ρ_{Rinf}
Apr 1
May 1
Jun 1
Jul 1
Sep 15
Oct 15
Nov 15
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)


Minneapolis (SEIR+HD)
South Florida (SEIR+HD)
Symptomatic rate
BaseParamSymp
MdlParamSymp
RateSymp
0.015
0.01
0.005
0
0.03
0.025
0.02
0.015
0.01
0.005
0
Apr 1
May 1
Jun 1
Jul 1
(a)
(b)


Minneapolis (SEIR+HD) cumulative reported rate
BaseParam$_{\text{Rate}}$
MdlParam$_{\text{Rate}}$
SeroStudy$_{\text{Rate}}$
Cumulative reported rate
0.20
0.15
0.10
0.05
0.00
Apr 1
May 1
(a)
$\times 10^2$
Minneapolis (SEIR+HD) BaseParam simulation
Isolate 75% pre/asymptomatic infections
Isolate 50% pre/asymptomatic infections
Isolate 25% pre/asymptomatic infections
Isolate reported & symptomatic infections
Isolate reported infections
BaseParam$_{\text{Rinf}}$
Reported infections
2.0
1.5
1.0
0.5
0.0
Apr 1
May 1
(b)
$\times 10^2$
Minneapolis (SEIR+HD) MdlParam simulation
Isolate 75% pre/asymptomatic infections
Isolate 50% pre/asymptomatic infections
Isolate 25% pre/asymptomatic infections
Isolate reported & symptomatic infections
Isolate reported infections
MdlParam$_{\text{Rinf}}$
Reported infections
2.0
1.5
1.0
0.5
0.0
Apr 1
May 1
(c)