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

Jiaming Cui[a], Arash Haddadan[b], A S M Ahsan-Ul Haque[c], Bijaya Adhikari[d], Anil Vullikanti[b,c], and B. Aditya Prakash[a,1]

[a]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332
[b]Biocomplexity Institute, University of Virginia, Charlottesville, VA 22904
[c]Department of Computer Science, University of Virginia, Charlottesville, VA 22904
[d]Department of Computer Science, The University of Iowa, Iowa City, IA 52242

September 14, 2021

Abstract

Estimating the true extent of the outbreak was one of the major challenges in combating COVID-19 outbreak early on. Our inability in doing so, allowed unreported/undetected infections to drive up disease spread in numerous regions in the US and worldwide. Accurately identifying the true magnitude of infections still remains a major challenge, despite the use of surveillance-based methods such as serological studies, due to their costs and biases. In this paper, we propose an information theoretic approach to accurately estimate the unreported infections. Our approach, built on top of an existing ordinary differential equations based epidemiological model, aims to deduce an optimal parameterization of the epidemiological model and the true extent of the outbreak which "best describes" the observed reported infections. Our experiments show that the parameterization learned by our framework leads to a better estimation of unreported infections as well as more accurate forecasts of the *reported* infections compared to the baseline parameterization. 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 a large majority of COVID-19 infections were unreported and non-pharmaceutical interventions indeed helped in mitigating the COVID-19 outbreak.

Keywords: information theory; model selection; epidemiological models; non-pharmaceutical interventions

[1]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. As of May 1, 2021, it had already resulted in more than 32 million reported infections, and half a million deaths just in the United States. Worldwide reported infections total at 161 million and deaths at 3.34 million [26]. The devastating effect of COVID-19 is not limited just to the public health sector, but also extends to the economy as a whole. For example, in the US the unemployment rate peaked to 15.8 percent in April 2020 [5], and U.S. GDP contracted at a 3.5% annualized rate for 2020 [1]. Similar economic impacts have been observed worldwide.

One of the major challenges in combating the COVID-19 pandemic early on was our inability to estimate the true magnitude of unreported infections. As noted by many studies [21, 15, 11, 54, 51, 41], 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. As a result, there were many instances of unreported COVID-19 outbreaks across the globe. In each of the early outbreak "hubs", unreported spread was later identified as the driving force behind the surge in positive infections. For example, phylogenetic studies revealed that COVID-19 had locally spread in Washington state before active community surveillance was implemented in early 2020 [16]. Similarly, there were only 23 reported infections in five major U.S. cities by March 1. However, it has been estimated that there were already more than 28000 total infections in those cities by then [13, 3]. Similar trends were observed in Italy, Germany, and the UK [18]. Even currently, a majority of asymptomatic and mild symptomatic infections remain difficult to identify [21, 35, 12].

More generally, accurate estimation of the true extent of the outbreak is a fundamental epidemiological question, and is also critical for pandemic planning and response. Note that the *true reported rate* (i.e., the ground truth reported rate), which is the ratio of reported infections to total infections, is different from the *case ascertainment rate*, which is the ratio of confirmed symptomatic infections to the true number of symptomatic individuals [50]. The ascertainment rate is also an important parameter, and has been studied extensively. However, the true reported rate can help in better evaluation of the effectiveness of different interventions, which has been a big challenge due to the uncertainty in the number of infections and deaths, e.g., [19, 14, 50]. It can also help in efficient resource allocation and the design and implementation of non-pharmaceutical interventions.

One of the most effective methods to identify the true magnitude of unreported infections in a region is through large-scale serological studies [17, 53, 32, 57]. 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] summarizes the results of serological studies conducted by commercial laboratories at a U.S. national level as well as 10 specific sites. As per the portal, in a sample collected by the CDC in Connecticut between April 26 and May 3 2020, it was estimated that the total infections were at least 6 times higher than the reported infections. Similarly, the estimated ratio of total infections to the reported infections in a sample collected between April 13 and April 25 2020 in Minnesota was at least 10. The same ratio was 12 in New York City between March 23 and Apr 1 2020. While serological studies give an accurate estimation of the unreported infections, they are expensive to carry out and not sustainable in a long run. It is even more challenging to perform real-time serological studies as there are unavoidable delays between sample collection and laboratory tests. The situation is even worse during the early stages of a novel pandemic when proper surveillance channels are not set up. Serological studies can also suffer from sampling biases [9], and heuristics have been designed to account for them.

Other lines of work to estimate unreported infections for COVID-19 try to exploit existing influenza surveillance systems to estimate symptomatic COVID-19 infections [42], due to symptomatic

similarities between the two diseases. However, they also suffer from numerous issues ranging from ad-hoc corrections to account for the symptomatic similarities, low resolution of the influenza surveillance systems (e.g., they fail to account for outbreaks in nursing homes), and inaccuracy stemming from seasonality of influenza when ILI surveillance systems are operating normally.

In the face of these surveillance challenges, data scientists and epidemiologists have devoted much time and effort to estimate the reported rate and forecast future COVID-19 trends through models [41, 50, 14, 49, 38, 40, 56]. Usually the reported rate α_{reported} is one of the parameters of the epidemiological model which needs to be calibrated/estimated. Although these epidemiological models, designed with expert knowledge, help guide policymakers, they can have several drawbacks. First, they often employ ad-hoc modelling assumptions such as equating unreported infections with case ascertainment [50] and/or predefining a set of parameters [22]. Secondly, they often use heuristics for parameter estimation, e.g., limiting the search space to a narrow range [28]. As a result, judging the *quality* of a specific parameterization $\mathbf{p}_M$ of an epidemiological model M is hard. Hence this may lead to multiple divergent parameterizations, which is magnified by noisy reported infection data, especially during the early stages of a pandemic [27, 22]. As we show later in our results, small errors in estimates of reported rates may lead to very different forecasts of future trends as well as non-intervention policy recommendations. Hence a principled approach to distinguish between these nearly equivalent epidemiological model parameterizations is needed in order to estimate the true extent of the outbreak.

To achieve this goal, in this paper, we propose a new information theory based approach. Imagine that we are given the ground truth extent of the outbreak i.e., the total infections D_{gt}. Let us call the parameterization we get by calibrating a given epidemiological model against D_{gt} as $\mathbf{p}_{\text{gt}}$. Note that the ground truth reported rate $\alpha_{\text{reported(gt)}}$ is also one of the parameters in $\mathbf{p}_{\text{gt}}$; so we can easily derive the reported infections as $D_{\text{reported}} = D_{\text{gt}} \times \alpha_{\text{reported(gt)}}$. However in the real-world, we do not know D_{gt} and $\alpha_{\text{reported(gt)}}$. Therefore, we leverage an information theoretic framework to find the total infections D by posing an optimization problem to select that D which minimizes the bits to describe the observed D_{reported}.

To describe D efficiently, we also encode $\mathbf{p}$, some baseline parameterization, and $\mathbf{p'}$, a candidate parameterization. Then the optimization is defined over all possible values of $\textsc{Model} = (D, \mathbf{p'}, \mathbf{p})$ with the goal of describing $\textsc{Data} = D_{\text{reported}}$ efficiently. Following the convention in information theory, we use $L(\textsc{Model})$ to denote the number of bits required to encode the $\textsc{Model}$ (i.e. D, $\mathbf{p'}$ and $\mathbf{p}$). Similarly, $L(\textsc{Data}|\textsc{Model})$ is the number of bits required to encode the $\textsc{Data}$, D_{reported}, given D, $\mathbf{p'}$, and $\mathbf{p}$. Then the overall objective of the optimization problem is to infer the optimal $\textsc{Model}^*$ which minimizes $L(\textsc{Model}) + L(\textsc{Data}|\textsc{Model})$.

The optimization framework described above follows the popular two-part Minimum Description Length (MDL) framework. The two-part MDL (aka sender receiver framework) consists of hypothetical actors S and R. Sender S is in possession of $\textsc{Data}$ and wants to transmit it to receiver R using as fewer bits as possible [31]. Hence, sender S searches for the best possible $\textsc{Model}$, which minimizes the overall cost of encoding and transmitting both the $\textsc{Model}$ and the $\textsc{Data}$ given the $\textsc{Model}$. Note that the two-part MDL (and MDL in general) does not make any assumption on the nature of the $\textsc{Data}$ or the $\textsc{Model}$. As such, it has been widely used for numerous optimization problems ranging from network summarization [37], causality inference [20], 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 [48]. 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 over.

We call our MDL-based optimization framework as MdlInfer. We use the ODE model proposed in [24] as the base epidemiological model O_{M} for MdlInfer and this article. It consists of

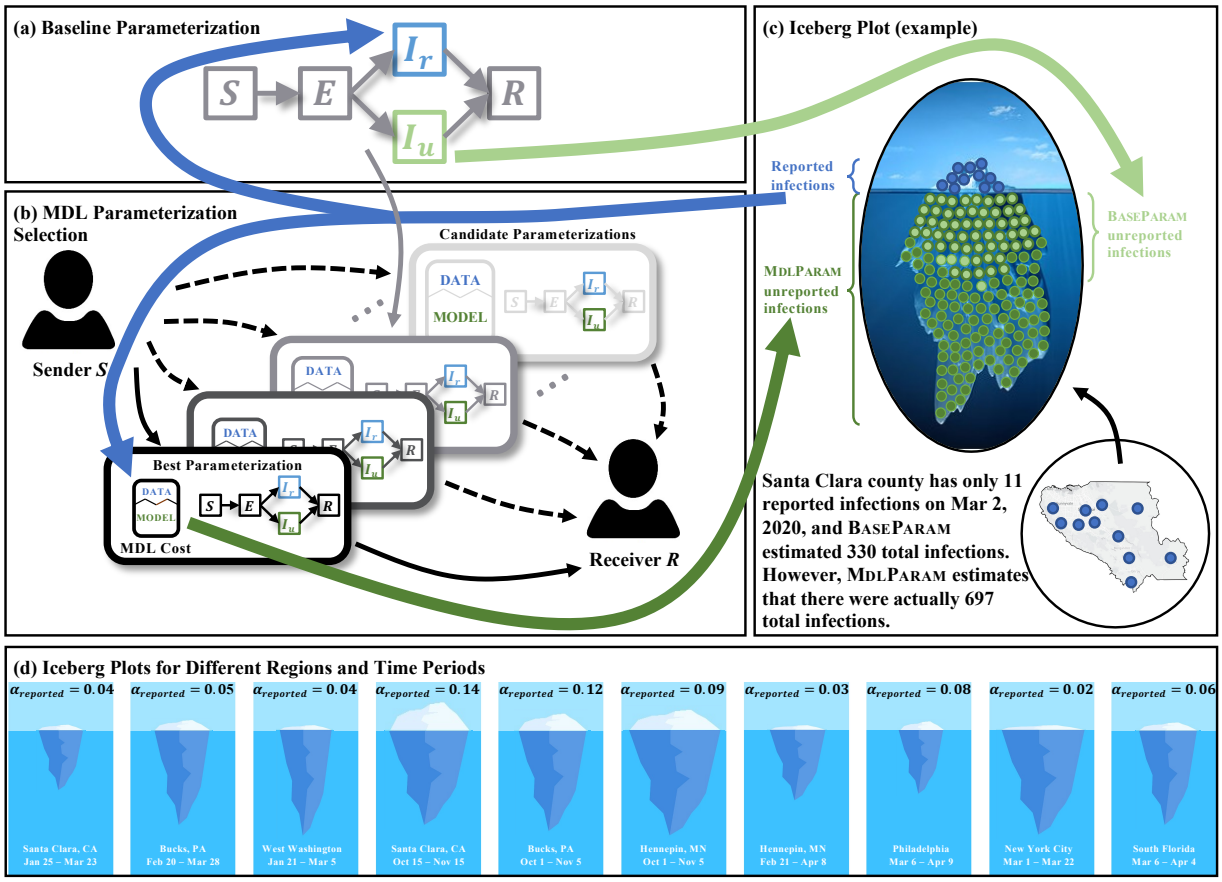


Figure 1: The workflow of MdlInfer. (a) The baseline parameterization underestimates the unreported infections (Shown in light green in the iceberg plot in (c)). (b) Our MDL model selection approach MdlInfer accurately estimates the extent of COVID-19 outbreak by inferring unreported infections given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

compartments representing the Susceptible, Exposed, various levels of Symptomatic infections, Hospitalization, Death and Recovered states. Following past work [36], here we assume that a fraction of symptomatic infections are reported each day and the rest are unreported. The parameters to be inferred include the transmission rate, proportional reduction on transmission post shelter-in-place orders, number of initial infections, and the reported rate (See Materials and Methods section for more detail). We construct a baseline using the standard approach via usual model calibration to estimate the reported rate. More specifically, we calibrate O_{M} using D_{reported} following the proce-

dure in [24], and use this as the baseline parameterization BaseParam represented by **p**. On the other hand, the parameterization O_{M} resulting from our approach is termed MdlParam and is represented by **p'**.

Our experiments show that MdlParam is superior to BaseParam in a variety of tasks. An example is shown in Fig. 1. On March 2, 2020 Santa Clara county has only 11 COVID-19 reported infections. The best version of BaseParam estimated that there were 330 total infections (colored as light green in the iceberg). On the other hand, our MdlParam gave an estimate of 697 total infections shown below the iceberg, which is closer to the total infections estimated from serological studies [17]. In the following Results section, we will show MdlInfer leads to more accurate unreported infections and symptomatic rate estimations than baseline parameterizations. It also leads to better fit and projections on reported infections. We also demonstrate that MdlInfer estimates more accurate reported rate over time and can aid in policy making through analysis of counter-factual non-pharmaceutical interventions.

Results

Here we present our empirical findings on a set of experiments at different geographical regions and different time periods. We chose these regions based on the severity of outbreak in the early stages of the pandemic and the availability of the serological studies and symptomatic surveillance data. In each region, we divide the timeline into two periods: (i) observed period when the number of reported infections are available and the models are trained to learn the parameters, and (ii) future period where we evaluate the forecasts generated by the epidemiological models trained in the observed period. Note that the inferred parameters include the proportional reduction on transmission post shelter-in-place order. Hence one would expect a more accurate parameterization to generalize better regardless of policy changes. The division between the periods is chosen either based on (i) onset of non-intervension policy where we choose the division date based on the onset of interventions such as stay-at-home orders and (ii) random partition where no such information is available.

The days prior to the start of NPI policy are in the observed period and the remaining days are in the future period. Note that the reported infections in the future period are not made available to the models. However, epidemiological model parameters include the proportional reduction in transmission rate post shelter-in-place orders. Hence, one would expect that the better parameterization leads to a better forecast. Next we describe our experiments and findings in detail.

To help understanding, we also listed the notation in the Supplementary Information.

Estimating Total Infections

MdlParam estimates total infections more accurately than BaseParam

We plotted total infections, $\mathrm{MdlParam}_{\mathrm{Tinf}}$, inferred by MdlParam (green curve) against the point estimates of the total infections calculated from serological studies $\mathrm{SeroStudy}_{\mathrm{Tinf}}$ (red dot) in the same figure (See Fig. 2 top row). Note that to compare with $\mathrm{SeroStudy}_{\mathrm{Tinf}}$, we are using the cumulative value of estimated total infections. The vertical red line shows 95% confidence interval on the estimate made by the serological studies. The blue curve represents the total infections, $\mathrm{BaseParam}_{\mathrm{Tinf}}$, estimated by BaseParam.

As seen in the figure, the total infections due to MdlParam fall within the confidence interval of the estimates given by serological studies. Fig. 2 (c) shows the results for the Western Washington region, where the baseline parameterization overestimates the total infections. However, MdlInfer

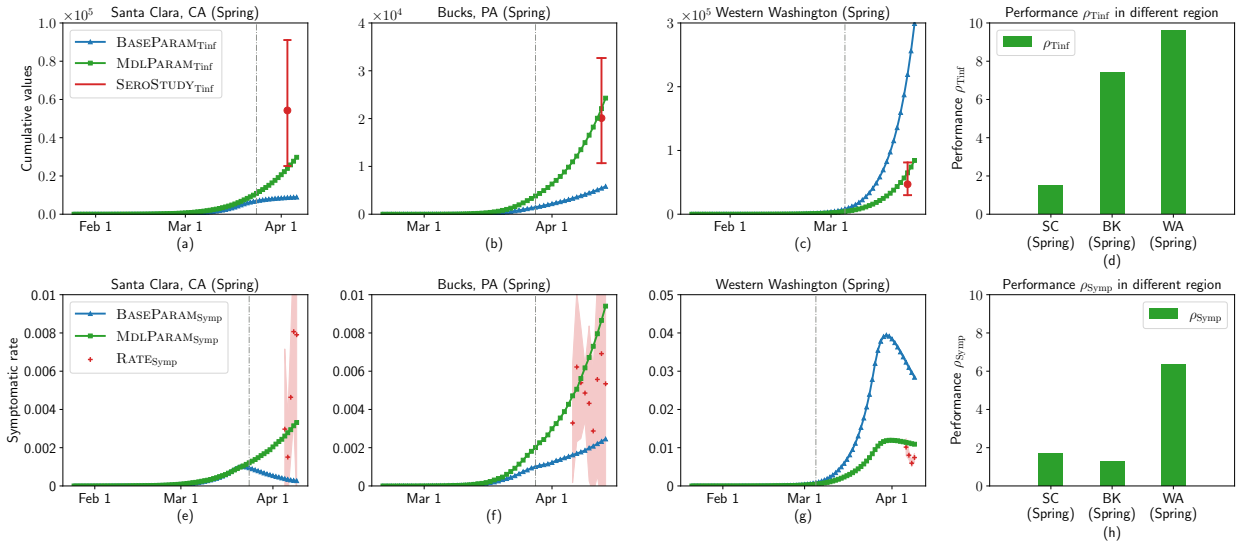


Figure 2: Results of total infections and symptomatic rate estimated by BaseParam and MdlParam. (a)-(c) MdlParam estimates total infections more accurately than BaseParam. The grey dash line divides the observed period (used to train BaseParam as well as MdlParam) and the future period (which was not accessible to the model while training). Blue curve and green curve represent the total infections estimated by baseline parameterization, BaseParam$_{\text{Tinf}}$, and total infections estimated by MdlInfer parameterization, MdlParam$_{\text{Tinf}}$ respectively. The red point estimate SeroStudy$_{\text{Tinf}}$ and confidence interval represent the total infections estimated by serological studies [17, 32]. Note that each plot corresponds to a different geographic region, and the scales are different. (d) The performance metric, ρ_{Tinf}, comparing MdlParam against BaseParam in estimating total infections are shown for the regions in (a)-(c). Here the values of ρ_{Tinf} are 1.50, 7.40, 9.62, implying that MdlParam performs better in estimating total infections than BaseParam. (e)-(g) MdlParam estimates more accurate symptomatic rate than BaseParam. Blue curve and green curve represent the symptomatic rate estimated by baseline parameterization, BaseParam$_{\text{Symp}}$, and symptomatic rate estimated by MdlInfer parameterization, MdlParam$_{\text{Symp}}$ respectively. The red point estimate Rate$_{\text{Symp}}$ and confidence interval represent the COVID-related symptomatic rate from Facebook's symptomatic surveillance data [8]. (h) The performance metric, ρ_{Symp}, comparing MdlParam against BaseParam in estimating symptomatic rate are shown for the regions in (e)-(g). Here the values of ρ_{Symp} are 1.71, 1.29, and 6.36, implying that MdlParam performs better in estimating symptomatic rate than BaseParam.

correctly predicts a lower total infection. This observation, in conjunction with the results in Santa Clara and Bucks county shown that MdlInfer can improve upon the baseline parameterization in either direction (i.e., by increasing and decreasing the total infections) as necessary.

To quantify the performance gap of the two approaches, we define the metric ρ_{Tinf} as $\rho_{\text{Tinf}} = \frac{\text{Rmse}(\text{BaseParam}_{\text{Tinf}},\text{SeroStudy}_{\text{Tinf}})}{\text{Rmse}(\text{MdlParam}_{\text{Tinf}},\text{SeroStudy}_{\text{Tinf}})}$. In Fig. 2 (d), we plotted the ρ_{Tinf}. Note that values of ρ_{Tinf} greater than 1 imply MdlParam estimation of the total infections are closer to the serological studies estimates than the BaseParam's estimation. Overall, the ρ_{Tinf} is usually larger than 1 (1.50, 7.40, and 9.62 in (a)-(c)) which indicates that MdlParam generally performs better in estimating total infections than BaseParam. More results in the Supplementary Information also show similar conclusion.

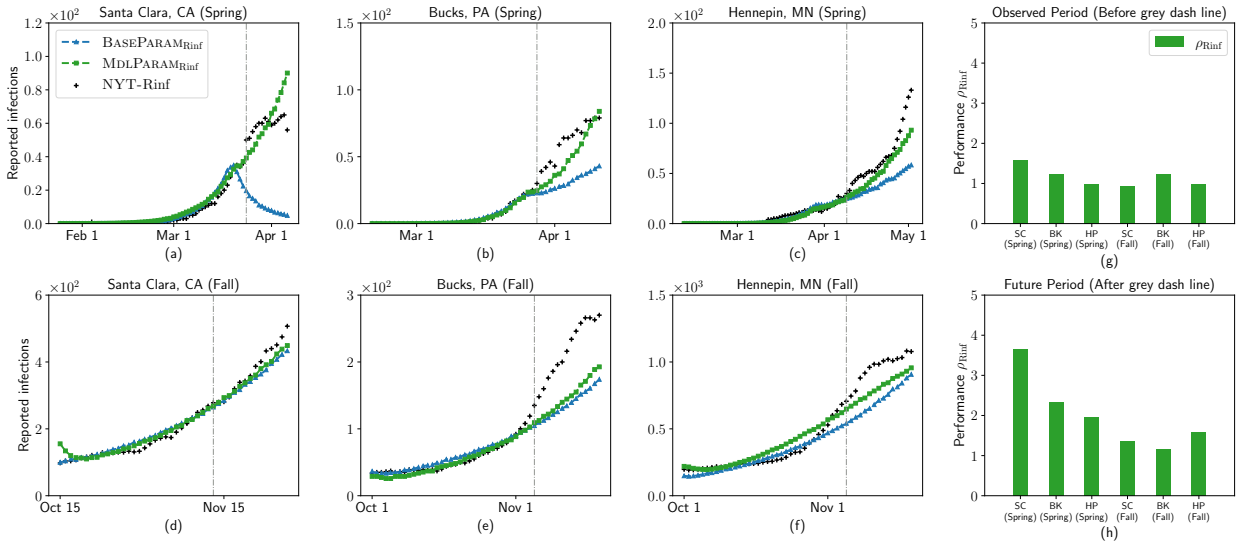


Figure 3: Results of reported infections estimated by BaseParam and MdlParam. (a)-(f) MdlParam leads to better fit and projection than BaseParam. The grey dash line divides the observed period and future period. Black plus symbols, blue curve, and green curve represent the infections reported by the New York Times, NYT-Rinf, reported infections estimated by baseline parameterization, BaseParam$_{\text{Rinf}}$, and reported infections estimated by MdlParam parameterization, MdlParam$_{\text{Rinf}}$ respectively. Note that each plot corresponds to a different geographic region, and the scales are different. The results show that MdlParam$_{\text{Rinf}}$ aligns much closer with NYT-Rinf than BaseParam$_{\text{Rinf}}$. Besides, MdlParam$_{\text{Rinf}}$ projects the future trends better than BaseParam$_{\text{Rinf}}$. (g)-(h) The performance metric, ρ_{Rinf}, comparing MdlParam against BaseParam in estimating reported infections are shown for the regions for both observed period (g), and future period (h). In both observed and future period, reported infections estimated by MdlInfer is better than baseline parameterization overall, and ρ_{Rinf} in future period is even larger than ρ_{Rinf} in observed period implying that MdlParam leads to a better generalization.

Estimating Symptomatic Rate

MdlParam estimates the symptomatic rate more accurately than BaseParam

We validate this observation using Facebook's symptomatic surveillance data [8]. Here we plot the inferred symptomatic rate over time and overlay the estimates and confidence interval from the symptomatic surveillance data (See Fig. 2 bottom row). The green and blue curves are the MdlParam and BaseParam symptomatic rates, MdlParam$_{\text{Symp}}$ and BaseParam$_{\text{Symp}}$ respectively. As seen in the figure, the ground truth symptomatic rate Rate$_{\text{Symp}}$ (represented by red plus symbols) aligns more closely with MdlParam$_{\text{Symp}}$. Using a similar definition as earlier, we define a quantitative performance metric to compare the performance of MdlParam against BaseParam in estimating symptomatic rate, $\rho_{\text{Symp}} = \frac{\text{Rmse}(\text{BaseParam}_{\text{Symp}}, \text{Rate}_{\text{Symp}})}{\text{Rmse}(\text{MdlParam}_{\text{Symp}}, \text{Rate}_{\text{Symp}})}$. In Fig. 2 (h), we plotted the ρ_{Symp}. We notice that the ρ_{Symp} is larger than 1 in all settings (1.71, 1.29, and 6.36 in (e)-(g)), indicating that MdlParam performs better than BaseParam in estimating the symptomatic rate. More results in the Supplementary Information also show similar conclusion.

To summarize, these two sets of experiments together demonstrate that BaseParam is still failing to estimate the true unreported infections. On the other hand, MdlParam infers unreported infections which is closer to the ones estimated by serological studies and symptomatic surveillance

data.

Estimating Reported Infections

MdlParam leads to better fit and projection than BaseParam at different stage of the COVID-19 epidemic

Our approach shows success based on the proximity. Here we train MdlParam on the observed period and calibrate BaseParam using the same. We then use the trained models to forecast the reported infections in future period, which was not accessible to the model while training. The results are summarized in Fig. 3.

In Fig. 3 (a) to (f), the vertical grey dash divides the observed and future period. The black plus symbols represent New York Times reported infections NYT-Rinf. Similarly, the blue curve depicts the number of reported infections, $\text{BaseParam}_{\text{Rinf}}$, as per BaseParam. And finally, the green curve represents the number of reported infections, $\text{MdlParam}_{\text{Rinf}}$, as per MdlParam. Note that the curves to the right of the horizontal grey line are the future predictions. As easily seen in the figure, NYT-Rinf aligns more closely with $\text{MdlParam}_{\text{Rinf}}$ than $\text{BaseParam}_{\text{Rinf}}$, indicating the superiority of our approach in estimating reported infections for both observed and future period.

We again define a performance metric to compare MdlParam against BaseParam in estimating reported infections as $\rho_{\text{Rinf}} = \frac{\text{Rmse}(\text{BaseParam}_{\text{Rinf}},\text{NYT-Rinf})}{\text{Rmse}(\text{MdlParam}_{\text{Rinf}},\text{NYT-Rinf})}$. In Fig. 3 (h), we plotted the ρ_{Rinf} for the future period. In both (g) and (h), we observe that the ρ_{Rinf} is more than 1. This further shows that MdlParam results in a better fit on reported infections than BaseParam. We also notice that MdlParam's gain is minimal when the base epidemiological model is not expressive enough to capture the observed trend (See Fig. 3 (e)). However, such cases are rare.

Note that Fig. 3 (a) to (c) correspond to the early state of the COVID-19 epidemic in spring 2020, and Fig. 3 (d) to (f) correspond to fall 2020. Here, the value of ρ_{Rinf} is greater than 1 indicating that MdlParam outperforms BaseParam and our MdlInfer performs well in estimating temporal patterns at different stage of the COVID-19 epidemic. More results in the Supplementary Information also show similar conclusion.

Evaluating effect of Non-pharmaceutical Interventions

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

We plotted the reported rate inferred by MdlParam using the icebergs (See Fig. 1 (d)). The height of the iceberg above the water surface (shown in white) and the height below the water(in blue) are proportional to the number of reported and the inferred unreported infections respectively. As seen in the figure, the height above the water is significantly smaller than the height under the water across all regions, implying that the majority of the infections were actually unreported. We also notice that the reported rate for New York City is extremely low when compared with other regions, this could be explained by the fact that New York City has more international transportation and therefore the COVID-19 may have already spread in latent for a longer time than other regions.

We also computed the 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 Fig. 4 (a).

As seen in the figure, the reported rate $\text{MdlParam}_{\text{Rate}}$ sharply increased in early February in Santa Clara county, CA followed by a gradual decay and eventual rise. This observation is explained by the fact that Santa Clara was one of the first counties to observe the imported infections.

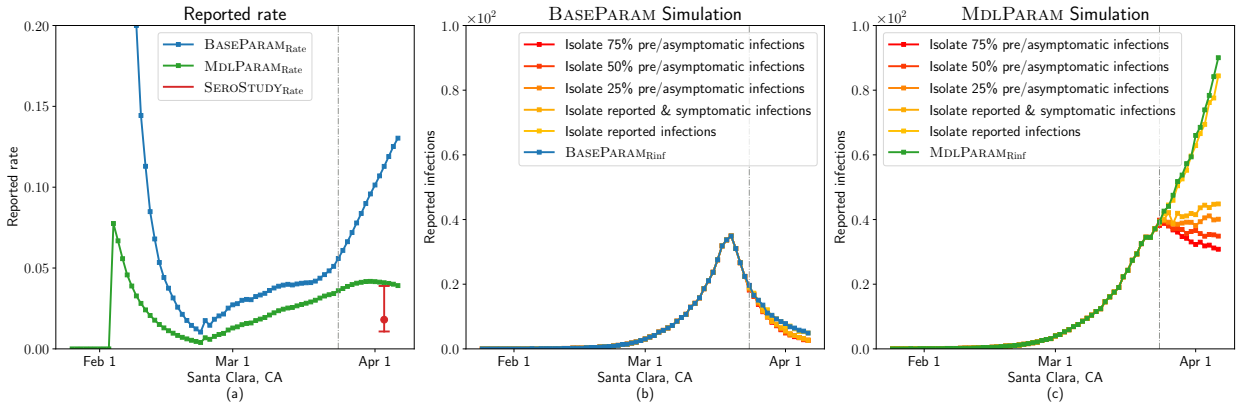


Figure 4: Results of reported rate, and non-pharmaceutical intervention simulations on asymptomatic and presymptomatic infections. (a) The reported rate trend during COVID-19 outbreak. Blue curve and green curve represent the reported rate estimated by baseline parameterization, BaseParam$_{\text{Rate}}$, and reported rate estimated by MdlInfer parameterization, MdlParam$_{\text{Rate}}$ respectively. The red point estimate and confidence interval represent the reported rate SeroStudy$_{\text{Rate}}$ estimated by serological studies [17]. As seen in the figure, the MdlParam$_{\text{Rate}}$ is high in February, possibly due to imported infections, it then gradually decreases. It finally increases in March 2020. Our results in (b) and (c) demonstrate that the accuracy of non-pharmaceutical intervention simulations rely on the accuracy of the inferred unreported infections. (b) The grey dash line divides the observed period and future period. The blue curve represents the reported infections estimated by BaseParam. The other five curves represent the simulated reported infections for 5 scenarios: Isolate the reported infections, symptomatic infections, symptomatic infections and 25%, 50%, 75% asymptomatic and presymptomatic infections, where we reduce the infectiousness of these isolated infections to half in future period. As seen in the figure, non-pharmaceutical interventions on symptomatic infections can be enough to control the COVID-19 epidemic. However, this has been proven to be incorrect by prior studies and real-world observations [43]. (c) Non-pharmaceutical interventions on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic. The green curve represents the reported infections estimated by MdlParam. The other five curves represent the simulated reported infections for the same 5 scenarios as in (b). As seen in the figure, even isolating all symptomatic infections, the reported infections are still increasing. Therefore, only with non-pharmaceutical interventions on some fraction of asymptomatic and presymptomatic infections, the reported infections will decrease.

However, the community spread driven COVID-19 outbreaks were not reported until late February. The observation that MdlParam$_{\text{Rate}}$ is extremely low in early March also fits the earlier study [42].

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 generated by MdlParam in Fig. 4 (c) (green curve). We then repeated the simulation of reported infections for 5 different scenarios: isolate just the reported infections, just the symptomatic infections, and symptomatic infections in addition to 25%, 50%, and 75% of both asymptomatic and presymptomatic infections. In our setup, we assume that the infectiousness 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 on the rise. On the other hand, non-pharmaceutical interventions on some fraction of asymptomatic and presymptomatic infections leads to decrease in reported infections. Thus, we can conclude that non-pharmaceutical interventions on asymptomatic infections are essential in controlling the COVID-19 epidemic.

Accuracy of non-pharmaceutical intervention simulations relies on the good inference of unreported infections

Here we also plotted the simulated reported infections generated by BaseParam in Fig. 4 (b) (blue curve). As seen in the figure, based on BaseParam we can infer that only non-pharmaceutical interventions on symptomatic infections can be enough to control the COVID-19 epidemic. However, this has been proven to be incorrect by prior studies and real-world observations [43]. Therefore, we can conclude that the accuracy of non-pharmaceutical intervention simulation relies on the goodness of the inferred parameterization.

Discussion and Future Work

In this study, we propose MdlInfer, a data-driven model selection approach that automatically estimates unreported infections based on an optimal parameterization of epidemiological models. Our approach leverages the information theoretic Minimum Length Description (MDL) principle, to select unreported infections which best describes the observed outbreak. Our approach addresses several gaps in current practice including long term infeasibility of serological studies [17, 32] and a lack of formal model selection over epidemiological models. MdlInfer employs a principled method that searches over a large space of parameterizations and finally selects the one that describes the outbreak best, while the existing epidemiological models [24, 41, 45] typically rely on the parameterization inferred 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 and the symptomatic rate at various geographical scales more accurately than baseline parameterization from both directions, i.e., it corrects both over- and under-estimates. For example, compared to the baseline parameterization, we correctly estimate 10792 more infections by April 1 in Santa Clara, CA (Fig. 2 (a)), and 214710 fewer infections in Western Washington (Fig. 2 (c)). We also show that MdlInfer leads to better fit of the reported infections in the observed period and more accurate forecasts for the future period than the baseline parameterization. We also 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. Although our results are slightly inconsistent in matching the symptomatic rate, which can be attributed to the noise in the surveillance data, we are still fitting the symptomatic rate better than the baseline parameterization. And our results show consistent performance with respect to the reported infections and serological studies.

The MdlInfer framework is likely to be useful in 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 important for public health. Daily incidence of COVID-19 has decreased since early 2021, according to the CDC COVID Data Tracker portal [6]. However new variants of the SARS-CoV-2 (e.g., the delta and gamma variants) have been spreading rapidly [39, 47, 30]. Testing for these, and large-scale surveillance via laboratory tests may be limited and less systematic than what was

done for COVID-19 in the past few months. 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 base epidemiological model. If the base 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 useful to adapt MdlInfer to give a measure of 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 [25, 52, 34, 55, 23, 46], 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, e.g., there are differences in severity rate or mortality rate, when infected with COVID-19, for different age group [44, 33], which has not been considered in our work.

To summarize, MdlInfer is a robust data-driven method to accurately estimate unreported 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 epidemiology using information theoretic methods.

Materials and Methods

Data

We are using the following publicly available datasets for our study.

1. **New York Times reported infections dataset** The New York Times reported infections dataset, or NYT-Rinf, consists of daily time sequence of reported infections D_{reported} and the mortality $D_{\text{mortality}}$ (cumulative values) for each county in the US starting from January 21, 2020 [7].

2. **Serological studies:** The serological studies [17, 32] consists of the point and 95% confidence interval estimate of the prevalence of antibodies to SARS-CoV-2. Using the population and the prevalence of antibodies, we can compute the estimated total infections $\text{SeroStudy}_{\text{Tinf}}$ and 95% confidence interval in the location.

3. **Symptomatic surveillance data:** The symptomatic surveillance data [8] consists of point estimate $\text{Rate}_{\text{Symp}}$ and confidence interval of the COVID-related symptomatic rate for each county in the US starting from April 6, 2020. The survey asks a series of questions designed to help researchers understand the spread of COVID-19 and its effect on people in the United States.

Our Approach

Minimum Length Description

We formulate the problem of inferring COVID-19 unreported infections using the Minimum Length Description (MDL) principle [10, 31]. Here we use the two-part sender-receiver framework. The goal of the framework is to transmit the Data from the possession of the sender S to the receiver

R using a MODEL. We do this by identifying the MODEL describes the DATA such that the total number of bits needed to represent both the MODEL and the DATA is minimized. We term this as a cost function composed by two parts:

1. Model cost $L(\text{MODEL})$: The cost of describing the MODEL.
2. Data cost $L(\text{DATA}|\text{MODEL})$: The cost of describing DATA given the MODEL.

MDL Formulation

Before formalizing the model space, $L(\text{MODEL})$ and $L(\text{DATA}|\text{MODEL})$, we will first define some notations and concepts.

In this work, we term the COVID-19 reported infections as D_{reported}, candidate total infections as D (which corresponds to unreported infections $D_{\text{unreported}}$), and mortality as $D_{\text{mortality}}$. We also term the epidemiological model [24] as O_{M}.

For O_{M}, we define the calibration on D_{reported} and $D_{\text{mortality}}$ as the procedure to find the parameterization $\mathbf{p}$ that minimizes the log likelihood function between $D_{\text{reported}}(\mathbf{p})$ and D_{reported}, and between the $D_{\text{mortality}}(\mathbf{p})$ and $D_{\text{mortality}}$. Here, the $D_{\text{reported}}(\mathbf{p})$ is the O_M output reported infections based on $\mathbf{p}$, and the $D_{\text{mortality}}(\mathbf{p})$ is the O_M output mortality based on $\mathbf{p}$. In this work, $\mathbf{p}$ is the baseline parameteriztion. We write this procedure as follows:

$$\mathbf{p} = \text{CALIBRATE}(O_{\text{M}}, D_{\text{reported}}, D_{\text{mortality}}) \tag{1}$$

Note that the procedure can be extended without $D_{\text{mortality}}$ easily.

From $\mathbf{p}$, we can generate the O_M output unreported infections $D_{\text{unreported}}(\mathbf{p})$, and total infections $D(\mathbf{p}) = D_{\text{reported}}(\mathbf{p}) + D_{\text{unreported}}(\mathbf{p})$. We can also calculate the baseline reported rate $\mathbf{p}[\alpha_{\text{reported}}]$:

$$\mathbf{p}[\alpha_{\text{reported}}] = \frac{\sum_t D_{\text{reported}}(\mathbf{p})}{\sum_t D(\mathbf{p})} \tag{2}$$

Similarly, we can also calibrate O_{M} on $D_{\text{unreported}} = D - D_{\text{reported}}$, D_{reported}, and $D_{\text{mortality}}$ to find the parameterization $\mathbf{p'}$. We write this procedure as follows:

$$\mathbf{p'} = \text{CALIBRATE}(O_{\text{M}}, D_{\text{unreported}}, D_{\text{reported}}, D_{\text{mortality}}) \tag{3}$$

From $\mathbf{p'}$, we can generate the O_M output unreported infections $D_{\text{unreported}}(\mathbf{p'})$, reported infections $D_{\text{reported}}(\mathbf{p'})$, total infections $D(\mathbf{p'}) = D_{\text{reported}}(\mathbf{p'}) + D_{\text{unreported}}(\mathbf{p'})$. We can also calculate the reported rate $\mathbf{p'}[\alpha_{\text{reported}}]$:

$$\mathbf{p'}[\alpha_{\text{reported}}] = \frac{\sum_t D_{\text{reported}}(\mathbf{p'})}{\sum_t D(\mathbf{p'})} \tag{4}$$

With the notations and concepts defined above, we will next formalize the model space, $L(\text{MODEL})$ and $L(\text{DATA}|\text{MODEL})$.

Model Space

In this work, the DATA to describe is D_{reported}, and therefore the most natural MODEL would have been $\text{MODEL} = (\mathbf{p})$ as it directly corrects from the BASEPARAM reported infections $D_{\text{reported}}(\mathbf{p})$ to describe D_{reported}. However, this has the disadvantage that the model space could be fragile that slightly different $\mathbf{p}$ could lead to vastly different costs. To account for this, we propose our MODEL as $\text{MODEL} = (D, \mathbf{p'}, \mathbf{p})$, which consists of three components. We use D to reparameterize the $\mathbf{p'}$, use $\mathbf{p}$ to send the $\mathbf{p'}$, and use the $\mathbf{p'}$ and D to correct from $\mathbf{p'}[\alpha_{\text{reported}}] \times D$ to send D_{reported}.

Model Cost

With the model space proposed above, the sender S will send the $\textsc{Model} = (D, \mathbf{p'}, \mathbf{p})$ to the receiver R in three parts:

1. First send the $\mathbf{p}$.
2. Next send the $\mathbf{p'}$ given $\mathbf{p}$.
3. Then send D given $\mathbf{p'}$ and $\mathbf{p}$.

Therefore the MDL Model Cost, $L(D, \mathbf{p'}, \mathbf{p})$ will have three components:

$$L(D, \mathbf{p'}, \mathbf{p}) = \textsc{Cost}(\mathbf{p}) + \textsc{Cost}(\mathbf{p'}|\mathbf{p}) + \textsc{Cost}(D|\mathbf{p'}, \mathbf{p}) \tag{5}$$

Specifically, we will send the $\mathbf{p}$ directly, send the $\mathbf{p'}$ given $\mathbf{p}$ by sending $\mathbf{p'} - \mathbf{p}$, and send D given $\mathbf{p'}$ and $\mathbf{p}$ by sending $\mathbf{p'}[\alpha_{\text{reported}}] \times D - D_{\text{reported}}(\mathbf{p})$. We write this down as below:

$$\begin{aligned} L(D, \mathbf{p'}, \mathbf{p}) = \textsc{Cost}(\mathbf{p}) &+ \textsc{Cost}(\mathbf{p'} - \mathbf{p}|\mathbf{p}) \\ &+ \textsc{Cost}(\mathbf{p'}[\alpha_{\text{reported}}] \times D - D_{\text{reported}}(\mathbf{p})|\mathbf{p'}, \mathbf{p}) \end{aligned} \tag{6}$$

The details of the encoding method to express the cost for each of the three components can be found in the Supplementary Information.

Data Cost

Next we need to send (describe) the $\textsc{Data}$, reported infections D_{reported} in terms of the $\textsc{Model}$. Given the $\textsc{Model} = (D, \mathbf{p'}, \mathbf{p})$, we describe D_{reported} by describing $\frac{D - D_{\text{reported}}}{1 - \mathbf{p'}[\alpha_{\text{reported}}]} - D(\mathbf{p'})$. We write down the MDL Data Cost as below:

$$L(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p}) = \textsc{Cost}(\frac{D - D_{\text{reported}}}{1 - \mathbf{p'}[\alpha_{\text{reported}}]} - D(\mathbf{p'})|D, \mathbf{p'}, \mathbf{p}) \tag{7}$$

Total MDL Cost

Putting both the MDL Model Cost and MDL Data Cost together, the Total MDL Cost is

$$\begin{aligned} L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p}) &= L(D, \mathbf{p'}, \mathbf{p}) + L(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p}) \\ &= \textsc{Cost}(\mathbf{p}) + \textsc{Cost}(\mathbf{p'}|\mathbf{p}) + \textsc{Cost}(D|\mathbf{p'}, \mathbf{p}) \\ &+ \textsc{Cost}(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p}) \\ &= \textsc{Cost}(\mathbf{p}) + \textsc{Cost}(\mathbf{p'} - \mathbf{p}|\mathbf{p}) \\ &+ \textsc{Cost}(\mathbf{p'}[\alpha_{\text{reported}}] \times D - D_{\text{reported}}(\mathbf{p})|\mathbf{p'}, \mathbf{p}) \\ &+ \textsc{Cost}(\frac{D - D_{\text{reported}}}{1 - \mathbf{p'}[\alpha_{\text{reported}}]} - D(\mathbf{p'})|D, \mathbf{p'}, \mathbf{p}) \end{aligned} \tag{8}$$

Problem Statement

With $L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p})$ formulated above, we can state the problem of as one of searching for the best total infections D^*: Given the time sequence D_{reported}, epidemiological model O_{M}, find D^* that minimizes the MDL total cost i.e.

$$D^* = \arg\min_{D} L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p}) \tag{9}$$

Algorithm

Next, we will present our algorithm to find D^*. Note that directly searching D^* naively is intractable. An alternate method is to first find a good reported rate $\alpha^*_{\text{reported}}$ fast since we can constrain $D = \frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$ and reduce the search space. Then with $\alpha^*_{\text{reported}}$ from step 1, we can search for the optimal D^*. Hence we propose a two-step search algorithm to find D^*. Key steps involved in our algorithm are as follows:

1. First, we will do a linear search to find a good reported rate $\alpha^*_{\text{reported}}$.
2. Given the $\alpha^*_{\text{reported}}$ found above, we will 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 $\alpha^*_{\text{reported}}$. We formulate this as below

$$\alpha^*_{\text{reported}} = \arg\min_{\alpha_{\text{reported}}} L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p}) \tag{10}$$

where the $D = \frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$ in the linear search. This helps to reduce the search space. Hence, the $\mathbf{p'}$ is

$$\mathbf{p'} = \textsc{Calibrate}(O_{\text{M}}, \frac{D_{\text{reported}}}{\alpha_{\text{reported}}} - D_{\text{reported}}, D_{\text{reported}}, D_{\text{mortality}}) \tag{11}$$

For stability and robustness, as D_{reported} can be noisy, we use $D(\mathbf{p'})$ instead of $\frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$ as D in the MDL Total Cost in this step.

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

With the $\alpha^*_{\text{reported}}$ found in step 1, we will next find the D^* that minimizes the MDL cost.

$$D^* = \arg\min_{D} L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p}) \tag{12}$$

Since we have already found a good $\alpha^*_{\text{reported}}$, we will constrain the D^* as below

$$\sum_t D^* = \frac{\sum_t D_{\text{reported}}}{\alpha^*_{\text{reported}}} \tag{13}$$

We then use the Nelder-Mead method [29] solve this constrained optimization problem for D^*, and we initialize the search from $\frac{D_{\text{reported}}}{\alpha^*_{\text{reported}}}$. We describe the two-step algorithm in more detail in the Supplementary Information.

Calibration

The epidemiological model [24] 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 process described in [24] 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. During inference, the model calibrates on mortality.

We extend the calibration process described above to include the reported infections and unreported infections. Note that we do not make any structural changes to the existing model. Similar to [36], in our calibration process we compute the newly reported infections and unreported infection as follows:

1. New reported $= \alpha_1 \times (dI_P I_S + dI_P I_M)$: Here $dI_P I_S + dI_P I_M$ is the number of new symptomatic infections everyday. We assume α_1 proportion of new symptomatic infections everyday are reported.

2. New unreported $= (1 - \alpha_1) \times (dI_P I_S + dI_P I_M) + dEI_A$.

We also extend the calibration to infer two more parameters: α (proportion of infections that are asymptomatic) and α_1 (proportion of new symptomatic infections that are reported). The extended epidemiological model O_M now calibrates on β_0, σ, E_0, α, and α_1. Hence, our vector $\mathbf{p}$ of model parameters is defined as $\mathbf{p} = [\mathbf{p}[\beta_0], \mathbf{p}[\sigma], \mathbf{p}[E_0], \mathbf{p}[\alpha], \mathbf{p}[\alpha_1]]$

Calibration Results

From the parameterization vector $\mathbf{p}$ and O_{M}, we can infer the reported infections $D_{\text{reported}}(\mathbf{p})$, unreported infections $D_{\text{unreported}}(\mathbf{p})$, and total infections $D(\mathbf{p}) = D_{\text{reported}}(\mathbf{p}) + D_{\text{unreported}}(\mathbf{p})$. Here, we calculate the BaseParam total infections (cumulative values) as

$$\textsc{BaseParam}_{\text{Tinf}} = \sum_t D(\mathbf{p}) \tag{14}$$

Note that BaseParam also estimates the number of symptomatic people, which can be calculated from the number of infections in state I_S (severe symptomatic) and I_M (mild symptomatic). With the number of symptomatic people, we can also calculate the COVID-19 related symptomatic rate among the population N:

$$\textsc{BaseParam}_{\text{Symp}} = \frac{I_S(\mathbf{p}) + I_M(\mathbf{p})}{N} \tag{15}$$

Here, we also calculate the BaseParam estimated reported infections

$$\textsc{BaseParam}_{\text{Rinf}} = D_{\text{reported}}(\mathbf{p}) \tag{16}$$

and the reported rate

$$\textsc{BaseParam}_{\text{Rate}} = \frac{\sum_t \text{NYT-Rinf}}{\sum_t D(\mathbf{p})} \tag{17}$$

From **p'**, we can also calculate similar measures for MdlParam.

Data Availability

Code and data have been deposited in GitHub [4].

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.

[2] Commercial Laboratory Seroprevalence Survey Data.

[3] Hidden Outbreaks Spread Through U.S. Cities Far Earlier Than Americans Knew, Estimates Say.

[4] https://github.com/adityalab/mdl-ode-missing.

[5] Interim Economic Projections for 2020 and 2021.

[6] United States COVID-19 Cases, Deaths, and Laboratory Testing (NAATs) by State, Territory, and Jurisdiction.

[7] Coronavirus in the U.S.:Latest Map and Case Count, 2020.

[8] Delphi's COVID-19 Surveys, 2020.

[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] Alene, M., Yismaw, L., Assemie, M. A., Ketema, D. B., Mengist, B., Kassie, B., and Birhan, T. Y. Magnitude of asymptomatic covid-19 cases throughout the course of infection: A systematic review and meta-analysis. *PloS one 16*, 3 (2021), e0249090.

[13] Aleta, A., Martín-Corral, D., Bakker, M. A., y Piontti, A. P., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N. E., Halloran, M. E., Longini, I. M., et al. Quantifying the importance and location of sars-cov-2 transmission events in large metropolitan areas. *medRxiv* (2020).

[14] 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. *Harvard Data Science Review* (7 2020). https://hdsr.mitpress.mit.edu/pub/y9vc2u36.

[15] 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.

[16] 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.

[17] Bendavid, E., Mulaney, B., Sood, N., Shah, S., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra-Walker, R., Tedrow, J., Bogan, A., et al. Covid-19 antibody seroprevalence in santa clara county, california. *International journal of epidemiology 50*, 2 (2021), 410–419.

[18] Böhning, D., Rocchetti, I., Maruotti, A., and Holling, H. Estimating the undetected infections in the covid-19 outbreak by harnessing capture–recapture methods. *International Journal of Infectious Diseases 97* (2020), 197–201.

[19] Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A. B., Leech, G., Altman, G., Mikulik, V., Norman, A. J., Monrad, J. T., Besiroglu, T., Ge, H., Hartwick, M. A., Teh, Y. W., Chindelevitch, L., Gal, Y., and Kulveit, J. Inferring the effectiveness of government interventions against covid-19. *Science 371*, 6531 (2021).

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

[21] Byambasuren, O., Cardona, M., Bell, K., Clark, J., McLaws, M.-L., and Glasziou, P. Estimating the extent of asymptomatic covid-19 and its potential for community transmission: systematic review and meta-analysis. *Official Journal of the Association of Medical Microbiology and Infectious Disease Canada 5*, 4 (2020), 223–234.

[22] 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.

[23] Chang, S. Y., Wilson, M. L., Lewis, B., Mehrab, Z., Dudakiya, K. K., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., et al. Supporting covid-19 policy response with large-scale mobility-based modeling. *medRxiv* (2021).

[24] Childs, M. L., Kain, M. P., Kirk, D., Harris, M., Couper, L., Nova, N., Delwel, I., Ritchie, J., and Mordecai, E. A. The impact of long-term non-pharmaceutical interventions on covid-19 epidemic dynamics and control. *medRxiv* (2020).

[25] Cuevas, E. An agent-based model to evaluate the covid-19 transmission risks in facilities. *Computers in biology and medicine 121* (2020), 103827.

[26] 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.

[27] Drake, J. M., Handel, A., Marty, E., B.O'Dea, E., and Tredennick, A. T. Transmission model for sars-cov-2 in us states.

[28] 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.

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

[30] 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).

[31] Grunwald, P. A tutorial introduction to the minimum description length principle. *arXiv preprint math/0406077* (2004).

[32] 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.

[33] 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.

[34] 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.

[35] Johansson, M. A., Quandelacy, T. M., Kada, S., Prasad, P. V., Steele, M., Brooks, J. T., Slayton, R. B., Biggerstaff, M., and Butler, J. C. Sars-cov-2 transmission from people without covid-19 symptoms. *JAMA network open 4*, 1 (2021), e2035057–e2035057.

[36] 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.

[37] 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.

[38] 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.

[39] Kustin, T., Harel, N., Finkel, U., Perchik, S., Harari, S., Tahor, M., Caspi, I., Levy, R., Leshchinsky, M., Dror, S. K., et al. Evidence for increased breakthrough rates of sars-cov-2 variants of concern in bnt162b2-mrna-vaccinated individuals. *Nature Medicine* (2021), 1–6.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Mueller, A. L., McNamara, M. S., and Sinclair, D. A. Why does covid-19 disproportionately affect older people? *Aging (Albany NY) 12*, 10 (2020), 9959.

[45] 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.

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

[47] 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* (2021), 1–7.

[48] 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.

[49] Press, W. H., and Levin, R. C. Modeling, post covid-19, 2020.

[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] Shaman, J. An estimation of undetected covid cases in france, 2020.

[52] Silva, P. C., Batista, P. V., Lima, H. S., Alves, M. A., Guimarães, F. G., and Silva, R. C. Covid-abs: An agent-based model of covid-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals 139* (2020), 110088.

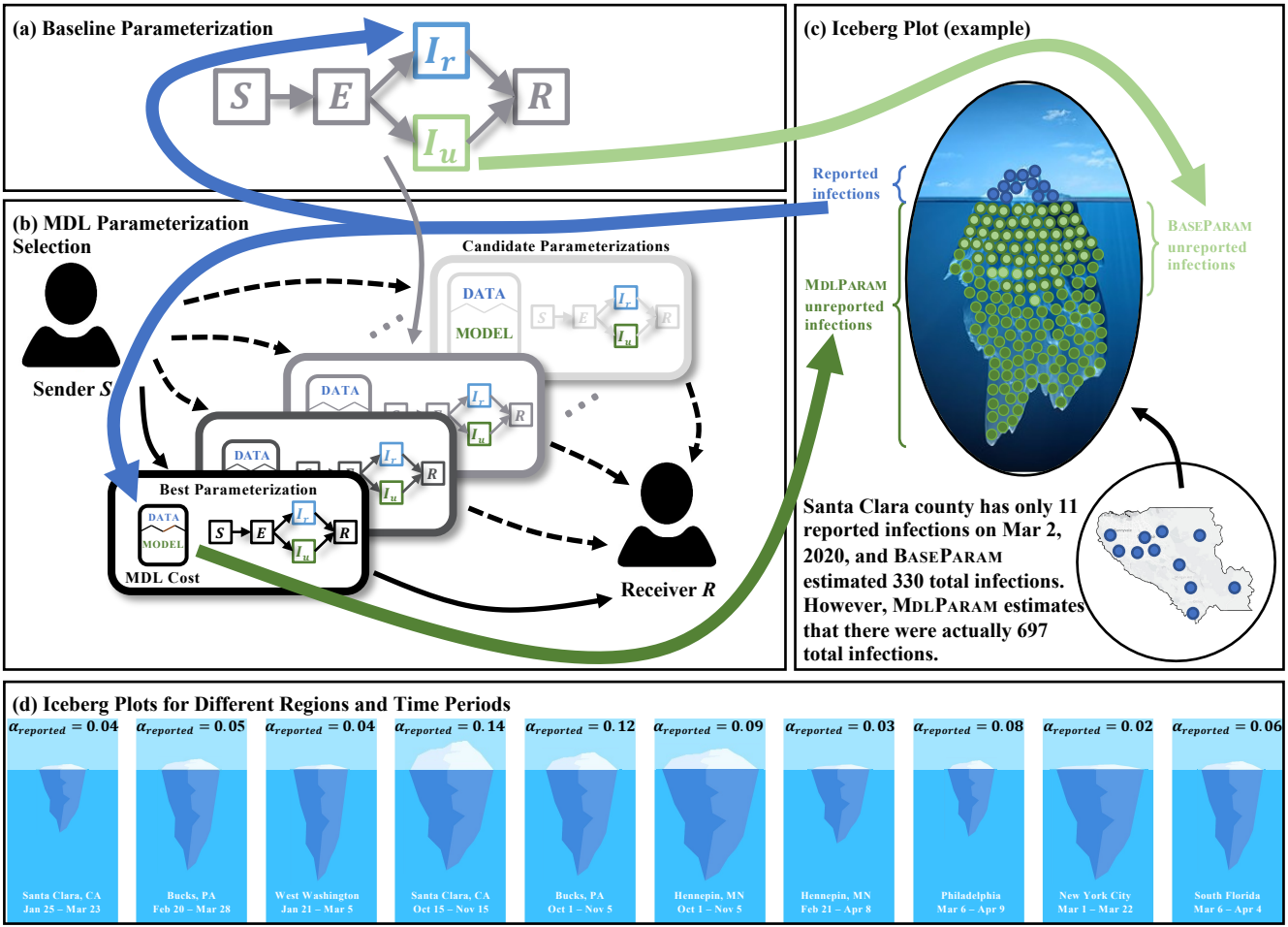
[53] Sood, N., Simon, P., Ebner, P., Eichner, D., Reynolds, J., Bendavid, E., and Bhattacharya, J. Seroprevalence of sars-cov-2–specific antibodies among adults in los angeles county, california, on april 10-11, 2020. *Jama* (2020).

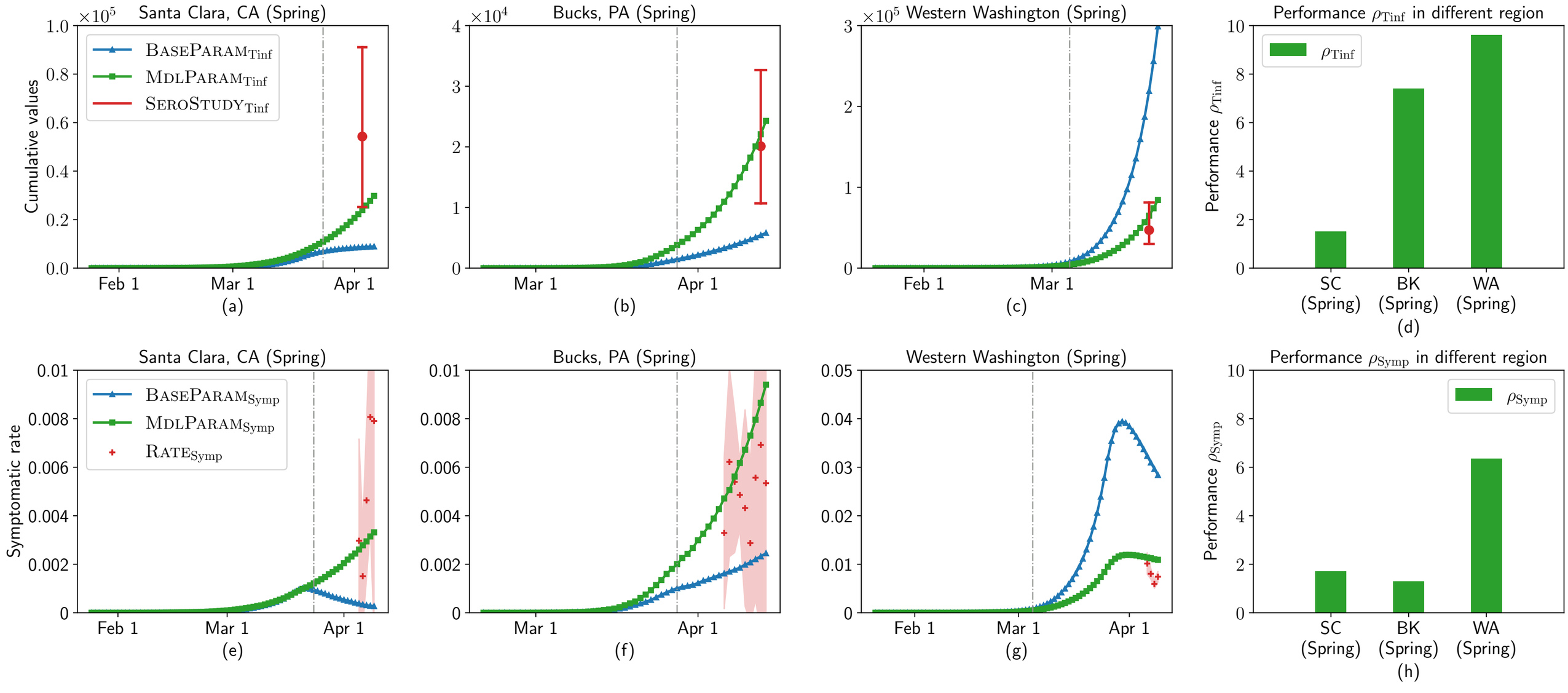
[54] Stockmaier, S., Stroeymeyt, N., Shattuck, E. C., Hawley, D. M., Meyers, L. A., and Bolnick, D. I. Infectious diseases and social distancing in nature. *Science 371*, 6533 (2021).

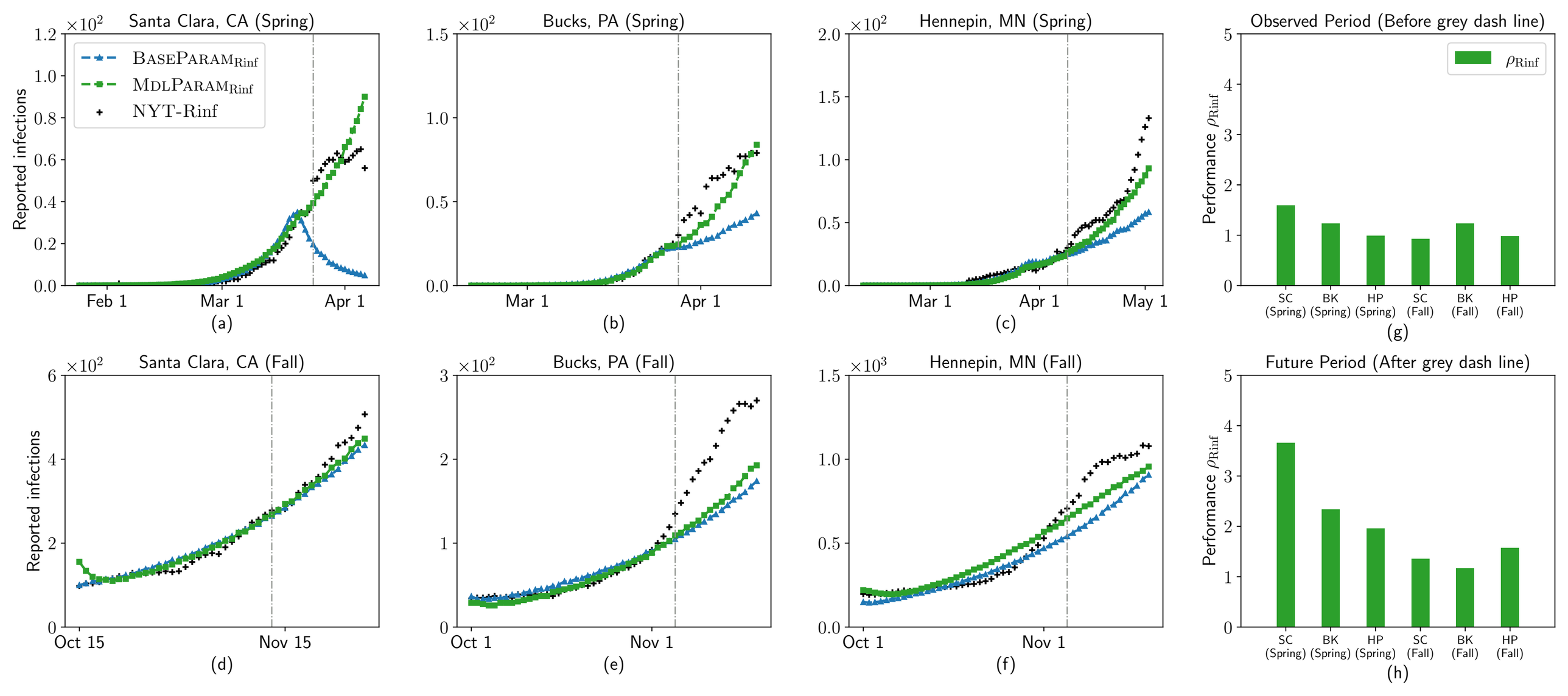
[55] Tian, Y., Sridhar, A., Yağan, O., and Poor, H. V. Analysis of the impact of mask-wearing in viral spread: Implications for covid-19. In *2021 American Control Conference (ACC)* (2021), IEEE, pp. 3132–3137.

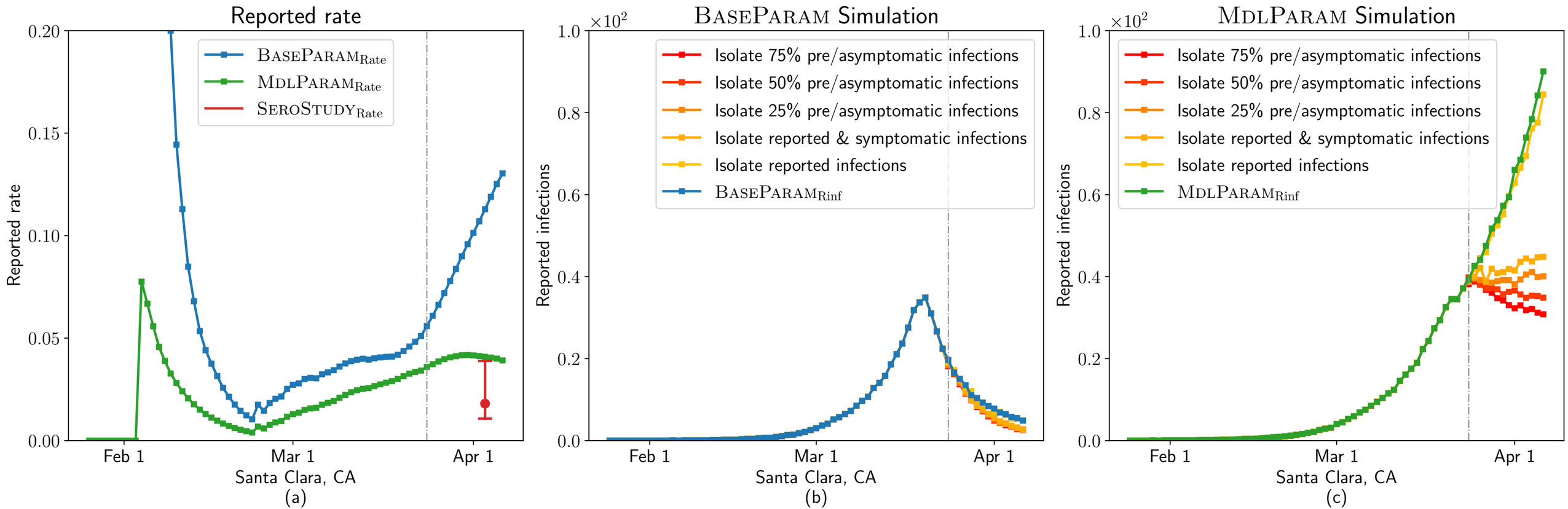
[56] Wells, C. R., Sah, P., Moghadas, S. M., Pandey, A., Shoukat, A., Wang, Y., Wang, Z., Meyers, L. A., Singer, B. H., and Galvani, A. P. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences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

[57] 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–e2024191.


(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
Santa Clara county has only 11 reported infections on Mar 2, 2020, and BASEPARAM estimated 330 total infections. However, MDLPARAM estimates that there were actually 697 total infections.
(d) Iceberg Plots for Different Regions and Time Periods
$\alpha_{reported} = 0.04$
Santa Clara, CA Jan 25 – Mar 23
$\alpha_{reported} = 0.05$
Bucks, PA Feb 20 – Mar 28
$\alpha_{reported} = 0.04$
West Washington Jan 21 – Mar 5
$\alpha_{reported} = 0.14$
Santa Clara, CA Oct 15 – Nov 15
$\alpha_{reported} = 0.12$
Bucks, PA Oct 1 – Nov 5
$\alpha_{reported} = 0.09$
Hennepin, MN Oct 1 – Nov 5
$\alpha_{reported} = 0.03$
Hennepin, MN Feb 21 – Apr 8
$\alpha_{reported} = 0.08$
Philadelphia Mar 6 – Apr 9
$\alpha_{reported} = 0.02$
New York City Mar 1 – Mar 22
$\alpha_{reported} = 0.06$
South Florida Mar 6 – Apr 4


Santa Clara, CA (Spring)
Bucks, PA (Spring)
Western Washington (Spring)
Performance ρ_{Tinf} in different region
Cumulative values
$\times 10^5$
$\times 10^4$
$\times 10^5$
BaseParam$_{\mathrm{Tinf}}$
MdlParam$_{\mathrm{Tinf}}$
SeroStudy$_{\mathrm{Tinf}}$
ρ_{Tinf}
Performance ρ_{Tinf}
Feb 1
Mar 1
Apr 1
SC (Spring)
BK (Spring)
WA (Spring)
(a)
(b)
(c)
(d)
Santa Clara, CA (Spring)
Bucks, PA (Spring)
Western Washington (Spring)
Performance ρ_{Symp} in different region
Symptomatic rate
BaseParam$_{\mathrm{Symp}}$
MdlParam$_{\mathrm{Symp}}$
Rate$_{\mathrm{Symp}}$
ρ_{Symp}
Performance ρ_{Symp}
(e)
(f)
(g)
(h)


Santa Clara, CA (Spring)
Bucks, PA (Spring)
Hennepin, MN (Spring)
Observed Period (Before grey dash line)
Santa Clara, CA (Fall)
Bucks, PA (Fall)
Hennepin, MN (Fall)
Future Period (After grey dash line)
BaseParamRinf
MdlParamRinf
NYT-Rinf
Reported infections
Performance ρ_{Rinf}
ρ_{Rinf}
Feb 1
Mar 1
Apr 1
May 1
Oct 15
Nov 15
Oct 1
Nov 1
SC (Spring)
BK (Spring)
HP (Spring)
SC (Fall)
BK (Fall)
HP (Fall)
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)


Reported rate
BaseParam$_{Rate}$
MdlParam$_{Rate}$
SeroStudy$_{Rate}$
Reported rate
0.20
0.15
0.10
0.05
0.00
Feb 1
Mar 1
Apr 1
Santa Clara, CA
(a)
BaseParam Simulation
$\times 10^2$
1.0
0.8
0.6
0.4
0.2
0.0
Isolate 75% pre/asymptomatic infections
Isolate 50% pre/asymptomatic infections
Isolate 25% pre/asymptomatic infections
Isolate reported & symptomatic infections
Isolate reported infections
BaseParam$_{Rinf}$
Reported infections
Feb 1
Mar 1
Apr 1
Santa Clara, CA
(b)
MdlParam Simulation
$\times 10^2$
1.0
0.8
0.6
0.4
0.2
0.0
Isolate 75% pre/asymptomatic infections
Isolate 50% pre/asymptomatic infections
Isolate 25% pre/asymptomatic infections
Isolate reported & symptomatic infections
Isolate reported infections
MdlParam$_{Rinf}$
Reported infections
Feb 1
Mar 1
Apr 1
Santa Clara, CA
(c)