Submitted to the Annals of Applied Statistics

CAUSAL INFERENCE IN THE TIME OF COVID-19 *

BY MATTEO BONVINI [1,*], EDWARD H. KENNEDY [1,†] VALERIE VENTURA [1,2,‡] AND LARRY WASSERMAN [1,2,§]

[1]*Department of Statistics and Data Science, Carnegie Mellon University*

[2]*Delphi Group, Carnegie Mellon University*

[*]MBONVINI@STAT.CMU.EDU; [†]EDWARD@STAT.CMU.EDU;
[‡]VVENTURA@STAT.CMU.EDU; [§]LARRY@STAT.CMU.EDU

March 10 2021

In this paper we develop statistical methods for causal inference in epidemics. Our focus is in estimating the effect of social mobility on deaths in the Covid-19 pandemic. We propose a marginal structural model motivated by a modified version of a basic epidemic model. We estimate the counterfactual time series of deaths under interventions on mobility. We conduct several types of sensitivity analyses. We find that the data support the idea that reduced mobility causes reduced deaths, but the conclusion comes with caveats. There is evidence of sensitivity to model misspecification and unmeasured confounding which implies that the size of the causal effect needs to be interpreted with caution. While there is little doubt the the effect is real, our work highlights the challenges in drawing causal inferences from pandemic data.

1. Introduction. During a pandemic, it is reasonable to expect that reduced social mobility will lead to fewer deaths. But how do we quantify this effect? In this paper we combine mechanistic epidemic models with modern causal inference tools to answer this question using state level data on deaths and mobility. Our goal is not to provide definitive estimates for the effects but rather to develop some methods and highlight the challenges in doing causal inference for pandemics. We also show how a generative epidemic model motivates a semiparametric causal model.

Epidemics are usually modeled by using *outcome models* (or generative models), which fully specify the distribution of the outcome (deaths). The most common epidemic model relates exposure, infections, recoveries and deaths by way of a set of differential equations. The simplest version is the SIR model (susceptible, infected, recovered) but there are many flavors of the model. We review the basic model in Section 4.

Instead of an outcome model, we use a marginal structural model (MSM) (Robins, Hernan and Brumback (2000); Robins (2000)). An MSM is a semiparametric model that directly models the effect of mobility on death without specifying an outcome model. Because it is semiparametric, it makes fewer assumptions than an outcome model. However, our MSM is motivated by a modified SIR-type outcome model.

We model deaths in each state separately to reduce confounding due to state differences. After obtaining model parameter estimates for each state, we estimate counterfactual quantities such as: how many deaths would have occurred if mobility had been reduced earlier, or if people had remained more vigilant throughout? Finally, we conduct a thorough sensitivity

[*]Ventura and Wasserman are members of the Delphi Group at CMU delphi.cmu.edu. This project arose from their work with Delphi. We are grateful for their help and support. We thank Rob Tibshirani for suggestions that greatly improved the paper. All the data can be obtained from the Delphi website covidcast.cmu.edu.

Keywords and phrases: Causal Inference, Marginal Structural Model, Covid-19.

analysis to explore the potential impact of model assumptions, model misspecification and unobserved confounding.

We will see that the data provide evidence for an effect of mobility. But the data are very limited. As mentioned above, we use state-specific models with weekly resolution due to concerns about data quality and unmeasured confounding due to geographic differences. The result is that we have about 40 observations per state. With so little data, we are restricted to use fairly simple models. We do find significant causal effects but we conduct sensitivity analyses that show that the effects need to be interpreted cautiously.

Related Work. A number of researchers have considered modeling the effect of interventions on the spread of Covid-19. Notable examples are Unwin et al. (2020), Chang et al. (2020), and IHME (2020). These authors develop very detailed epidemic models of the dynamics of the disease. One advantage of such an approach is that one can then consider the effects of a large array of potential interventions. Further, the models themselves are of great interest for understanding the progression of Covid-19. However, these models are very complex, and they involve a large number of parameters including parameters for various latent variables. Fitting such models and assessing uncertainty is challenging. Some authors take a Bayesian approach with informative priors. Others use heuristics such as reporting intervals based on using various settings of the parameters. To the best of our knowledge, it is not known how to get valid, frequentist confidence intervals in these complex models. This is not meant as a criticism of these papers but rather, this reflects the intrinsic difficulty of dealing with such models.

In contrast, our goal is to make the model as simple as possible and to use standard estimating equation methods so that standard errors can be obtained fairly easily. We do not claim that our approach is superior but we do believe that the model and the resulting confidence intervals are more transparent. Getting precise results from our simple model turns out to be challenging and raises doubts about the accuracy of published studies using highly complex models.

The papers by Chernozhukov, Kasaha and Schrimpf (2020) and Xiong et al. (2020) are much closer to ours. The authors of Chernozhukov, Kasaha and Schrimpf (2020) use a set of causal linear structural equations to model weekly cases as a function of social behavior (mobility) and social behavior as a function of policies. They model several policies simultaneously and they model all states simultaneously. They do obtain valid frequentist confidence intervals. Xiong et al. (2020) construct a measure of mobility inflow and using daily county level cases they fit a linear structural model to relate cases to mobility inflow. Our approach differs in several ways: we model deaths, we focus only on the effect of mobility, we model one state at a time, and we use a MSM rather than a regression outcome model. By modeling within each state, we have much less data at our disposal, which makes modeling very challenging. On the other hand, the threat of confounding due to state differences is reduced. By using a marginal structural model, our approach is semiparametric and so makes fewer assumptions. We focus on deaths instead of cases because we find the data on cases to be quite unreliable in general; for example, the availability of testing changed over time in various ways within and across states. Also, we place a strong emphasis on sensitivity analysis. These analyses complement each other nicely.

Paper Outline. We describe the data in Section 2. In Section 3 we review some basics of causal inference. In Section 4 we construct the models that we will use and we explain how the models are fit in Section 5. The results are presented in Section 6. Concluding remarks are in Section 7.

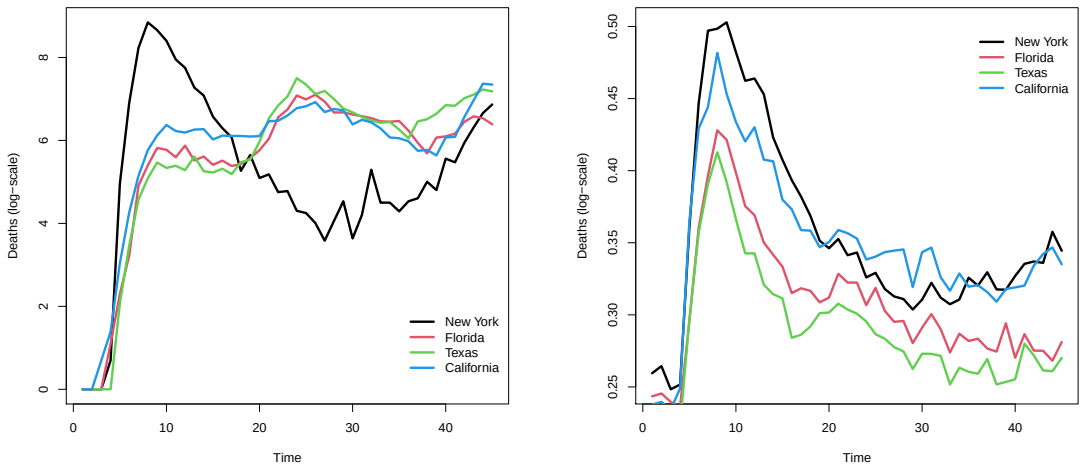

FIG 1. *Left: Plot of log deaths versus time (weeks), from Feb 15 2020 (week 1) to December 25 2020 (week 45), for four populous states. Right: Plot of a anti-mobility measure called "stay at home" versus week.*

2. Data. As mentioned earlier, we model each state separately, at the weekly level. The data are available at the daily county level but the weekly state level data are more reliable. Indeed, the data are subject to many reporting issues. It is not uncommon for a state to fail to report many deaths for a period and then suddenly report a bunch of unreported deaths on a single day. The problems are worse at the county level. Also, there are many small counties with very little data. We find using weekly state level data to be a good compromise between the quantity and quality of the data. We also note that epidemic analyses, such as flu surveillance, are generally done at the weekly level.

The data for each state have the form

$$(A_1, Y_1), \ldots, (A_T, Y_T)$$

where A_t is mobility on week t and Y_t is the number of deaths due to Covid-19 on week t. We obtain our data from CMU's Delphi group (`cmu.covidcast.edu`) which gets the death data from Johns Hopkins (`https://coronavirus.jhu.edu`) and the mobility data from Safegraph (`safegraph.com`). The data are from Feb 15 2020 (week 1) to December 25 2020 (week 45).

Figure 1 shows log deaths $L_t = \log(Y_t + 1)$ and "proportion at home" A_t which is one of the mobility measures, for four states. This is the fraction of mobile devices that did not leave the immediate area of their home. In this case, a higher value means less mobility so we can think of this measure as anti-mobility. This is the variable we will use throughout. In the rest of the paper we standardize mobility by subtracting A_1 from each value of A_t so that mobility starts at zero.

3. Causal Inference. In this section, we briefly review basic ideas from causal inference. Consider weekly mobility and death data $(A_1, Y_1), \ldots, (A_T, Y_T)$ in one state. Define $\overline{A}_t = (A_1, \ldots, A_t)$ and $\overline{Y}_t = (Y_1, \ldots, Y_t)$ for $t \geq 1$.

Now consider the causal question: what would Y_t be if we set $\overline{A}_t$ equal to some value $\overline{a}_t = (a_1, \ldots, a_t)$? Let $Y_t^{\overline{a}_t}$ denote this counterfactual quantity. It is important to distinguish the observed data $(\overline{A}_T, \overline{Y}_T)$ from the collection of unobserved counterfactual random variables

$$\left\{ Y^{\overline{a}_T} : \overline{a}_T \in \mathbb{R}^T \right\},$$

which is an infinite collection of random vectors, one for each possible mobility trajectory $\overline{a}_T$. We make the usual consistency assumption that $\overline{Y}_T = \overline{Y}_T^{\overline{A}_T}$. To make sure this is clear, consider a simple case where a subject gets either treatment $A = 1$ or control $A = 0$. In this case, the random variables are (A, Y, Y^0, Y^1) and the consistency assumption is that the observed outcome Y satisfies $Y = Y^1$ if $A = 1$ and $Y = Y^0$ if $A = 0$.

Causal inference when the treatment varies over time is subtle. It may be tempting to simply regress Y_T on the past and get the regression coefficient for mobility. This strategy has serious problems because $\overline{Y}_{T-1}$ are both confounding and mediating variables. Indeed, previous deaths can affect both future mobility and future deaths, while also being affected by previous mobility. More precisely, a large number of deaths implies a large number of infections which can cause future infections which then cause future deaths, and a large number of deaths might scare people into staying home. So we must adjust for past deaths. A common principle in epidemiology is to adjust for pre-treatment variables but not for post-treatment variables. But Y_s comes after A_{s-1} and before A_{s+1} making it both a pre-treatment and post-treatment variable. So how do we properly define the causal effect?

The solution is to use Robins' g-formula. Assuming for the moment that there are no other confounding variables except past deaths, Robins (1986) proved that the mean of $Y_t^{\overline{a}_t}$ is given by the g-formula:

$$\psi(\overline{a}_t) \equiv \mathbb{E}[Y_t^{\overline{a}_t}] = \int \cdots \int \mathbb{E}[Y_t | \overline{A}_t = \overline{a}_t, \overline{Y}_{t-1} = \overline{y}_{t-1}] \prod_{s=1}^{t-1} p(y_s | \overline{y}_{s-1}, \overline{a}_s)\, dy_s. \tag{1}$$

Some authors denote $\mathbb{E}[Y_t^{\overline{a}_t}]$ by $\mathbb{E}[Y_t | \mathrm{do}(\overline{a}_t)]$. When there are other confounders X_t besides past deaths, the formula becomes

$$\psi(\overline{a}_t) \equiv \int \cdots \int \mathbb{E}[Y_t | \overline{A}_t = \overline{a}_t, \overline{Y}_{t-1} = \overline{y}_{t-1}, \overline{X}_{t-1} = \overline{x}_{t-1}] \prod_{s=1}^{t-1} p(y_s, x_s | \overline{y}_{s-1}, \overline{a}_s, \overline{x}_{s-1})\, dy_s\, dx_s.$$

Intuitively, the g-formula can be obtained as follows. The density of $(\overline{y}_t, \overline{a}_t)$ can be written as

$$p(\overline{y}_t, \overline{a}_t) = \prod_{s=1}^{t} p(y_s | \overline{y}_{s-1}, \overline{a}_s) p(a_s | \overline{a}_{s-1}, \overline{y}_{s-1}). \tag{2}$$

Now replace $p(a_s | \overline{a}_{s-1}, \overline{y}_{s-1})$ with a point mass at a_s (i.e. the A's are fixed, no longer random) and then find of the mean of Y_t from this new distribution. It will be useful later in the paper to bear in mind that $\psi(\overline{a}_t) \equiv \psi(\overline{a}_t, p)$ is a functional of the joint density p from (2).

The next question is: how do we estimate $\psi(\overline{a}_t)$? A natural idea is to plug-in estimates of all the unknown quantities in the g-formula which leads to

$$\widehat{\psi}(\overline{a}_t) \equiv \int \cdots \int \widehat{\mathbb{E}}[Y_t | \overline{A}_t = \overline{a}_t, \overline{Y}_{t-1} = \overline{y}_{t-1}] \prod_{s=1}^{t-1} \widehat{p}(y_s | \overline{y}_{s-1}, \overline{a}_s)\, dy_s. \tag{3}$$

As discussed in Robins, Hernan and Brumback (2000); Robins (2000, 1989) there are a number of problems with this approach, called g-computation. If we plug-in nonparametric estimates, we quickly face the curse of dimensionality. If we use parametric estimates, we encounter the null-paradox (Robins and Wasserman (1997)): there may be no setting of the parameters which can represent the case where there is no treatment effect, i.e., there is no setting of the parameters which makes $\psi(\overline{a}_t)$ a constant function of $\overline{a}_t$.

An alternative approach to estimating $\psi(\overline{a}_t)$ (Robins, Hernan and Brumback (2000)) is to directly specify a parametric functional form $g(\overline{a}_t, \beta)$ for $\psi(\overline{a}_t)$. Such a model is called a

marginal structural model (MSM). Robins, Hernan and Brumback (2000) showed that β can be estimated by solving the following inverse-probability-weighted estimating equation:

$$\sum_t W_t^* \; h^*(\overline{A}_t)(Y_t - g(\overline{A}_t, \widehat{\beta})) = 0, \tag{4}$$

where the function h^* will be described shortly and the weights W_t^* are defined by

$$W_t^* = \prod_{s=1}^{t} \frac{1}{\pi(A_s|\overline{A}_{s-1}, \overline{Y}_{s-1})}, \tag{5}$$

where $\pi(a_t|\cdot)$ is the conditional density of mobility, assumed to be positive. For notational simplicity, we dropped the potential dependence on t in the functions $g(\overline{a}_t, \beta) = g_t(\overline{a}_t, \beta)$ and $h^*(\overline{a}_t) = h_t^*(\overline{a}_t)$, however they do differ across timepoints.

If the MSM is correctly specified, the function $h^*(\overline{a}_t)$ is arbitrary; each choice of h yields a consistent estimator. However, the choice of h^* affects the efficiency of the estimator. We follow the common practice of choosing a stabilized h^* of the form

$$h^*(\overline{a}_t) = h(\overline{a}_t) \prod_{s=1}^{t} \pi(A_s \mid \overline{A}_{s-1})$$

for some h; a common choice is $h(\overline{a}_t) = \partial g(\overline{a}_t, \beta)/\partial\beta$. We will then find it convenient to rewrite (4) as

$$\sum_t W_t \; h(\overline{A}_t)(Y_t - g(\overline{A}_t, \widehat{\beta})) = 0 \tag{6}$$

where

$$W_t = \prod_{s=1}^{t} \frac{\pi(A_s|\overline{A}_{s-1})}{\pi(A_s|\overline{A}_{s-1}, \overline{Y}_{s-1})}. \tag{7}$$

We refer to W_t as the stabilized weights.

REMARK. *There is a difference between the standard MSM setup and the one we are considering that warrants mentioning. Typically one assumes access to n different time series $(Z_1, ..., Z_n)$, with each series $Z = \{(A_1, Y_1), ..., (A_T, Y_T)\} = (\overline{A}_T, \overline{Y}_T)$ observed for n different independent units (e.g., states). There, one could have a different estimating equation at each time, for example,*

$$\sum_i W_{ti} \; h_t(\overline{A}_{ti})(Y_{ti} - g_t(\overline{A}_{ti}, \widehat{\beta})) = 0$$

where the i subscript denotes weights, treatments, outcomes, etc. for series i. If there are common parameters across timepoints, then these estimating equations could be combined, for example by summing over time, or using a generalized method of moments approach, etc. However, we model states individually, and so do not assume different states are independent. This leaves us with one observation per state at each time, which we then combine across time (but only within state) to obtain estimating equation (6). This represents the trade-off between independence versus modeling assumptions (e.g., Markov assumptions in the weights, or linearity in $g(\cdot)$): the less we require of one, the more we require of the other.

An MSM is a semiparametric model in the sense that it leaves the data generating process unspecified, subject to the restriction that the functional $\psi(\overline{a}_t)$ has a specific form. Specifically, let us write $\psi(\overline{a}_t)$ as $\psi(\overline{a}_t, p)$ to make it clear that $\psi(\overline{a}_t, p)$ depends on the joint density of the data $p(\overline{a}_T, \overline{y}_T)$ from (2). The model we are using is then

$$\mathcal{P} = \Big\{ p(\overline{a}_T, \overline{y}_T) : \text{ there exists } \beta \text{ such that } \psi(\overline{a}_t, p) = g(\overline{a}_t, \beta) \text{ for all } t \Big\}. \tag{8}$$

The model g is typically chosen to be interpretable. For example, suppose that $g(\overline{a}_t, \beta) = \beta_0 + \beta_1 \sum_s a_s$. Then the effect of the parameter settings is simple (i.e., mean outcomes only depend linearly on the amount of cumulative treatment), and the null (of no treatment effect) simply corresponds to $\beta_1 = 0$. It is important to keep in mind that this is not a model for the entire data generating process, just for marginal treatment effects, i.e., how mean outcomes under different treatment sequences are connected. In the next section we introduce an MSM motivated by an epidemic SIR model.

4. Models. Epidemics are often modeled using differential equations that describe the evolution of certain subgroups over time. Perhaps the most common is the SIR (Susceptible, Infected, Recovered) model (Kermack and McKendrick (1927), Brauer, Castillo-Chavez and Castillo-Chavez (2012), Bjørnstad (2018)) described by the equations

$$\begin{aligned} \frac{dS_t}{dt} &= -\frac{\alpha I_t S_t}{N} \\ \frac{dI_t}{dt} &= \frac{\alpha I_t S_t}{N} - \gamma I_t \\ \frac{dR_t}{dt} &= \gamma I_t, \end{aligned}$$

where N is population size, S_t is the number of susceptibles, I_t is the number of infected and R_t are the removed (by death or recovery) at time t. Solving the second equation conditional on S_t yields $I_t = I_{t-1} e^{\alpha \frac{S_t}{N} - \gamma}$, showing that, without any interventions, I_t grows exponentially. There are numerous generalizations of this model including stochastic versions, discretized versions and models with more states besides S, I and R.

4.1. *The Mobility Model.* Our proposed MSM is

$$g(\overline{a}_t, \nu_0, \beta, f) = \sum_{s=1}^{t} f(s,t) e^{\nu_0(s) + \sum_{r=1}^{s} \beta(a_r)} \tag{9}$$

with nuisance functions f, ν_0 and β. The model is motivated by the SIR model.

The basic idea of the SIR model is that there is a natural tendency for an epidemic to increase exponentially at the beginning. But there are also elements that reduce the epidemic such as the depletion of susceptible individuals due to recovery and death. At the beginning of a pandemic, reduction of susceptibles will play a negligible role. On the other hand, interventions like lockdowns, school closings etc can have a drastic effect. These considerations lead us to the following initial model.

Let I_t denote *new* infections in week t. Let

$$\begin{aligned} A_t &\sim Q_t \\ I_t &= e^{c_t + \beta(A_t)} I_{t-1} + \delta_t \\ Y_t &= \sum_{s=1}^{t} f(s,t) I_s + \xi_t \end{aligned} \tag{10}$$

where Q_t is an arbitrary distribution depending on $(\overline{A}_{t-1}, \overline{I}_{t-1}, \overline{Y}_{t-1})$, δ_t and ξ_t are mean 0 random variables (independent of the other variables), $f(s,t)$ denotes the probability that someone infected at time s dies at time t, the parameter c_t is a positive number and $\beth$ is a smooth function. Here, c_t represents the evolution of the epidemic without intervention and $\beth(A_t)$ is the effect of mobility. We allow c_t to vary with t to make the model more general and to allow the spread of Covid-19 to depend on the availability of susceptibles. We write

$$f(s,t) = d(s) f_0(s,t) \tag{11}$$

where $d(s)$ is the probability that someone infected at time s will eventually die and $f(s,t)$ is the probability that someone infected at time s and who will eventually die, will die at time t. Following Unwin et al. (2020) we take $f_0(s,t)$, on the scale of days, to be the density of $T_1 + T_2$ where T_1 (time from infection to symptoms) is Gamma with mean 5.1 and coefficient of variation 0.86 and T_2 (time from symptoms to death) is Gamma with mean 18.8 and coefficient of variation 0.45. The resulting distribution can be accurately approximated by a Gamma with mean 23.9 days and coefficient of variation 0.40. Finally, we integrate this distribution over 7 day bins to get $f(s,t)$ on a weekly scale.

At this point, we might use (10) as our model. But the I_t's are not observed. Furthermore, a non-linear, sequentially specified parametric outcome model can suffer from serious anomalies when used for causal inference. In particular, such a model can suffer from the *null paradox* (Robins (1986, 1989); Robins and Wasserman (1997)). This means that there may be no parameter values that satisfy (i) Y_t is conditionally dependent on past values of A_s and such that (ii) the null hypothesis of no treatment effect holds. We explain this point in more detail in the discussion in Section 7. We could also find $\mu_t = \mathbb{E}[Y_t | \overline{A}_t = \overline{a}_t, \overline{Y}_{t-1} = \overline{y}_{t-1}]$ and use μ_t to infer the effects of A on Y. But finding μ_t and applying the g-formula in (1) to μ_t may be intractable and there is still the danger of a null paradox.

Instead, we apply the g-formula to the model specified by (10) to find $\mathbb{E}[Y_t^{\overline{a}_t}]$ and use the resulting function as an MSM. This yields

$$\mathbb{E}[Y_t^{\overline{a}_t}] = \sum_{s=1}^{t} f(s,t) e^{\nu_0(s) + \sum_{r=1}^{s} \beth(A_r)} \equiv g(\overline{a}_t, \nu_0, \beth, f) \tag{12}$$

where $\nu_0(s) = \log I_1 + \sum_{r=1}^{s} c_r$. (We treat I_1 as an unknown parameter that is the absorbed into ν_0.) Now we abandon the initial model and just interpret $g(\overline{a}_t, \nu_0, \beth, f)$ directly as a model for the counterfactual $\mathbb{E}[Y^{\overline{a}_t}]$, that is, as an MSM. Put another way, we start with the model (10), find $g(\overline{a}_t, \nu_0, \beth, f) = \mathbb{E}[Y^{\overline{a}_t}]$, and then expand the model to include all joint distributions that satisfy $\mathbb{E}[Y_t^{\overline{a}_t}] = g(\overline{a}_t, \nu_0, \beth, f)$.

The MSM can be fit with the estimating equation (6), which corrects for confounding due to past deaths not by modeling the entire conditional outcome process, but by weighting by propensity weights W_t given by (7). This MSM approach allows us to be agnostic about whether it is our motivating model that holds, or some other much more complicated data-generating process. In fact, one can go further and take a completely agnostic view, in which the marginal structural model is not assumed correct at all, but only viewed as an approximation to the true, and possibly very complex, underlying counterfactual mean (Neugebauer and van der Laan, 2007).

4.2. *Simplified Models.* The MSM is not identified without further constraints. We will take $\beth(A_s) = \beta A_s$ so that

$$\mathbb{E}[Y_t^{\overline{a}_t}] = \sum_{s=1}^{t} f(s,t) e^{\beta \sum_{r=1}^{s} A_r + \nu_0(s)}.$$

This model is difficult to deal with numerically so we consider two approximations. First, we take $f_0(s,t)$ in (11) to be a point mass at $\delta = 4$ weeks (approximately its mean). Then we get

$$\mathbb{E}[Y_t^{\overline{a}_t}] = e^{d(t-\delta)+\nu_0(t-\delta)+\beta M_t}$$

where $M_t \equiv M(\overline{a}_t) = \sum_{s=1}^{t-\delta} a_s$. If we approximate $\log \mathbb{E}[Y_t^{\overline{a}_t}]$ with $\mathbb{E}[\log(Y_t^{\overline{a}_t})]$ we further obtain

$$\mathbb{E}[L_t^{\overline{a}_t}] = \log d(t-\delta) + \nu_0(t-\delta) + \beta M_t \tag{13}$$

where $L_t = \log(Y_t + 1)$. Finally, we take

$$\nu(t) \equiv \log d(t-\delta) + \nu_0(t-\delta) = \sum_{j=1}^{k} \beta_j \psi_j(t)$$

where $\psi_1, \ldots, \psi_k$ are orthogonal polynomials starting with $\psi_1(t) = t$. This model is easy to fit and will be used in Section 6. Note that the probability of dying $d(t)$ is allowed to change smoothly over time, which it likely did as hospitals were better prepared during the second wave. Interestingly, we have consistently found that using $k = 1$ leads to unreasonable results as we discuss in Section 6 but taking $k > 1$ solves this problems. The method for choosing k is described in Section 6.2. Note that $\partial \mathbb{E}[L_t^{\overline{a}_t}]/\partial a_s = \beta$ for any $s \le t - \delta$ so β has a clear meaning.

The model in (13) was used independently in Shi and Ban (2020) with $k = 1$. They used the model for curve fitting and they showed that this simple model fits the data surprisingly well. However, we find that making $\nu(t)$ non-linear seems to be important.

We will also consider a different approach to fitting the model. Specifically, we will use deconvolution methods to estimate the unobserved infection process $I_1, \ldots, I_T$. The first equation in (10) implies $\mathbb{E}[I_t] = e^{\nu(t)+\beta \sum_s A_s}$ suggesting the MSM

$$\mathbb{E}[L_t^{\overline{a}_t}] = \nu(t) + \beta M_t$$

which is the same as (13) except that now $L_t = \log(I_t)$ and $M_t = \sum_{s=1}^{t} a_s$ rather than $M_t = \sum_{s=1}^{t-\delta} a_s$.

REMARK. *We have regularized the model by restricting $\nu(t)$ to have a finite basis expansion. We also considered a different approach in which $\nu(t)$ is restricted to be increasing which seems a natural restriction if $\nu(t)$ is supposed to represent the growth of the pandemic in lieu of intervention. Using the methods in Meyer et al. (2008, 2018); Liao and Meyer (2018) we obtained estimates and standard errors. The results are very similar to the results in Section 6.*

Counterfactual Estimands. Now we discuss some causal quantities that we can estimate from the model. Let $\overline{a}_t = (a_1, \ldots, a_t)$ be a mobility profile of interest. After fitting the model we will plot estimates and confidence intervals for counterfactual deaths

$$\theta_t = \exp\Big\{\mathbb{E}[L^{\overline{a}_t}]\Big\} \tag{14}$$

under mobility regime $\overline{a}_t$, $t = 1, \ldots, T$.

We will consider the following three interventions:

$$\text{Start one week earlier} : \overline{a}_T = (A_2, A_3, \ldots, , A_{T+1})$$

$$\text{Start two weeks earlier} : \overline{a}_T = (A_3, A_4, \ldots, A_{T+2})$$

$$\text{Stay vigilant} : \overline{a}_T = (A_1, A_2, \ldots, A_9, A_{10}, A_{10}, A_{11}, A_{11}, A_{12}, A_{12}, A_{13}, A_{13}, \ldots)$$

The first two interventions aim to assess COVID-19 infections if we had started sheltering in place one and two weeks earlier. The last intervention halves the slope of the rapid decrease in stay at home mobility after the initial peak in week 9 that is clearly visible in Fig.1.

5. Fitting the Model. Now we discuss the method for estimating the model.

5.1. *Fitting the Semiparametric Model.* Recall the MSM

$$\mathbb{E}[L_t^{\overline{a}_t}] = \nu(t) + \beta M(\overline{a}_t) \tag{15}$$

where $\nu(t) = \sum_{j=1}^{k} \beta_j \psi_j(t)$. We estimate $\nu(t)$ and β by solving the estimating equation

$$\sum_t h_t(\overline{a}_t) W_t [L_t - (\widehat{\nu}(t) + \widehat{\beta} M(\overline{a}_t))] = 0 \tag{16}$$

corresponding to (6). We discuss the estimation of the weights W_t in Section 5.2. As is often done for MSMs we choose

$$h_t(\overline{a}_t) = (1, \psi_1(t), \ldots, \psi_k(t), M(\overline{a}_t))^T$$

since solving the estimating equation then corresponds to using least squares with weights W_t. Note that the estimating equation is the derivative of the weighted sum of squares set to zero.

Recall from (14) that $\theta_t = e^{\psi(\overline{a}_t)} = e^{\nu(t)+\beta M(\overline{a}_t)}$ which we estimate by $\widehat{\theta}_t = e^{\widehat{\nu}(t)+\widehat{\beta} M(\overline{a}_t)}$. We obtain approximate confidence intervals using the delta method. The asymptotic variance is based on the sandwich estimator.

5.2. *Estimating the Stabilized Weights.* To estimate the marginal structural model we need to estimate the weights

$$W_t = \prod_{s=1}^{t} \frac{\pi(A_s | \overline{A}_{s-1})}{\pi(A_s | \overline{A}_{s-1}, \overline{Y}_{s-1})}.$$

(We remind the reader that W_t denotes the stabilized weights which includes the numerator density; see eqs.(5) and (7)). One approach is to plug in estimates of the densities into the formula for W_t. But estimating these densities is not easy and ratios of density estimates can be unstable. The problem is exacerbated when we multiply densities. Instead we use a moment-based approach as in Fong et al. (2018); Zhou and Wodtke (2018). The idea is to estimate the vector of weights $W_1, \ldots, W_T$ by noting that they need to satisfy certain moment constraints. Our method is similar to the approach in Zhou and Wodtke (2018).

We rewrite $W_t = \prod_{s=1}^{t} V_s$ where

$$V_s \equiv V_s(\overline{A}_s, \overline{Y}_{s-1}) = \frac{\pi(A_s | \overline{A}_{s-1})}{\pi(A_s | \overline{A}_{s-1}, \overline{Y}_{s-1})}.$$

Let $\widetilde{h}_1(a_t)$ and $\widetilde{h}_2(y_{t-1})$ be arbitrary functions and define their centered versions by

$$h_1(a_t) = \widetilde{h}_1(a_t) - \mu_t$$

$$h_2(y_{t-1}) = \widetilde{h}_2(y_{t-1}) - \nu_t$$

where the conditional means are

$$\mu_t \equiv \mu_t(\overline{A}_{t-1}) = \mathbb{E}[\widetilde{h}_1(A_t) | \overline{A}_{t-1}]$$

$$\nu_t \equiv \nu_t(\overline{A}_{t-\delta-1}, \overline{Y}_{t-2}) = \mathbb{E}[\widetilde{h}_2(Y_{t-1}) | \overline{A}_{t-\delta-1}, \overline{Y}_{t-2}].$$

Weighted products of these functions have mean zero since

$$\begin{aligned}\mathbb{E}[h_1(A_t)h_2(Y_{t-1})W_t] &= \int\cdots\int h_1(a_t)h_2(y_{t-1})p(\overline{a}_t,\overline{y}_{t-1})W_t(\overline{a}_t,\overline{y}_{t-1})\,d\overline{a}_t\,d\overline{y}_{t-1}\\ &= \int\cdots\int h_1(a_t)h_2(y_{t-1})\pi(a_t|\overline{a}_{t-1},\overline{y}_{t-1})p(y_{t-1}|\overline{a}_{t-1},\overline{y}_{t-2})p(\overline{a}_{t-1},\overline{y}_{t-2})\\ &\qquad\times\frac{\pi(a_t|\overline{a}_{t-1})}{\pi(a_t|\overline{a}_{t-1},\overline{y}_{t-1})}\left(\prod_{s=1}^{t-1}V_s\right)\;d\overline{a}_t\,d\overline{y}_{t-1}\\ &= \int\left\{\omega(\overline{y}_{t-2},\overline{a}_{t-1})\int h_1(a_t)\pi(a_t|\overline{a}_{t-1})da_t\int h_2(y_{t-1})p(y_{t-1}|\overline{a}_{t-1},\overline{y}_{t-2})dy_{t-1}\right\}d\overline{a}_{t-1}\,d\overline{y}_{t-2}\\ &= 0\end{aligned}$$

from the definition of h_1 and h_2, where

$$\omega(\overline{y}_{t-2},\overline{a}_{t-1}) = p(\overline{y}_{t-2},\overline{a}_{t-1})\prod_{s=1}^{t-1}V_s.$$

Thus, the weights are characterized by the moment constraints

$$\mathbb{E}[h_1(A_t)h_2(Y_{t-1})W_t] = 0. \tag{17}$$

As in Zhou and Wodtke (2018) we estimate the weights by finding W_t to satisfy $\mathbb{E}[h_1(A_t)h_2(Y_{t-1})W_t] = 0$ for a set of functions h_1, h_2. This requires estimating these moments and estimating μ_t and ν_t. To proceed, we make a Markov assumption, namely

$$\mathbb{E}[\widetilde{h}_1(A_t)|\overline{A}_{t-1}] = \mathbb{E}[\widetilde{h}_1(A_t)|A_{t-1},\ldots,A_{t-k}]$$

and

$$\mathbb{E}[\widetilde{h}_2(Y_{t-1})|\overline{A}_{t-\delta-1},\overline{Y}_{t-2}] = \mathbb{E}[\widetilde{h}_2(Y_{t-1})|A_{t-1-\delta},\ldots,A_{t-k-\delta},Y_{t-2},\ldots,Y_{t-k}]$$

for some k. We will use $k=1$ in our analysis. Moreover, we assume homogeneity so that the functions μ_t and ν_t do not depend on t. Under the homogeneous Markov assumption, μ and ν can be estimated by regression. For example, if $k=1$, μ can be estimated by regressing $\widetilde{h}_1(A_2),\ldots,\widetilde{h}_1(A_T)$ on $A_1,\ldots,A_{T-1}$. (We tried both linear and nonparametric regression and obtained similar weights from each approach so we have used linear regression in our results.) The sample versions of the moment conditions (17) are then

$$\frac{1}{T}\sum_t H_{tj}W_t = 0$$

where

$$H_{tj} = (\widetilde{h}_{1j}(A_t) - \widehat{\mu}_j))(\widetilde{h}_{2j}(Y_{t-1}) - \widehat{\nu}_j)$$

and $\{(\widetilde{h}_{1j},\widetilde{h}_{2j}) :\ j=1,\ldots,J\}$ are a set of pairs of functions, $\widehat{\mu}_j$ is the estimate of $\mathbb{E}[\widetilde{h}_1(A_t)|A_{t-1},\ldots,A_{t-k}]$ and $\widehat{\nu}_j$ is the estimate of $\mathbb{E}[\widetilde{h}_2(Y_{t-1})|A_{t-1-\delta},\ldots,A_{t-k-\delta},Y_{t-2},\ldots,Y_{t-k}]$.

The moment conditions do not completely specify the weights. As in the above references we add a regularization term, in this case, $(1/2)\sum_t(W_t-1)^2$ and we require $\sum_t W_t = T$. This leads to the following minimization problem: minimize $W_1,\ldots,W_T$ in

$$\frac{1}{2}\sum_t(1-W_t)^2 + \lambda_0\sum_t(W_t - T) + \sum_{j=1}^{J}\lambda_j\sum_t W_tH_{tj} \tag{18}$$

1. Choose the order k of the Markov assumption.
2. Choose J pairs of functions $\left\{(\widetilde{h}_{1j}(a), \widetilde{h}_{1j}(y)) : \ j = 1, \dots J\right\}$.
3. Estimate $\mu_j = \mathbb{E}[\widetilde{h}_{1j}(A_t)|A_{t-k}, \dots, A_{t-1}]$ and $\nu_j = \mathbb{E}[\widetilde{h}_{2j}(Y_{t-1})|A_{t-k-\delta-1}, \dots, A_{t-\delta-1}, Y_{t-1-k}, \dots, Y_{t-2}]$ by regression.
4. Compute the weights $W_1, \dots, W_n$ from (19).
5. Fit the model $L_t = \beta \sum_{i=1}^{t-\delta} A_s + \nu(t) + \epsilon_t$ using weighed least squares with weights $W_1, \dots, W_n$.

FIG 2. *Steps for fitting the model.*

where the λ_j's are Lagrange multipliers. The solution to the minimization is

$$W = \mathbf{1} - H(H^T H)^{-1}[H^T \mathbf{1} - \mathbf{D}] \tag{19}$$

where $W = (W_1, \dots, W_T)$, $\mathbf{1}$ is a vector of $1's$, $\mathbf{D} = (T, 0, \dots, 0)^T$ and

$$H = \begin{pmatrix} 1 & H_{11} & \cdots & H_{1N} \\ 1 & H_{21} & \cdots & H_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & H_{T1} & \cdots & H_{TN} \end{pmatrix}$$

and N is the total number of moment constraints. In our case we choose $h_{11}(a) = a$, $h_{12}(a) = a^2$, $h_{21}(y) = y$, $h_{22}(y) = y^2$.

To include other time varying confounders X_t one should replace $h_2(y_{t-1})$ with two functions:

$$h_2(y_{t-1}) = \widetilde{h}_2(y_{t-1}) - \mathbb{E}[\widetilde{h}_2(y_{t-1})|\overline{X}_{t-1}, \overline{A}_{t-1}, \overline{Y}_{t-2}]$$

and

$$h_3(x_{t-1}) = \widetilde{h}_3(x_{t-1}) - \mathbb{E}[\widetilde{h}_3(x_{t-1})|\overline{X}_{t-2}, \overline{A}_{t-1}, \overline{Y}_{t-2}].$$

The steps for fitting the model are summarized in Fig.(2).

6. Results. In this section we give results for the mobility measure 'proportion of people staying at home.' We begin by showing the results of fitting the MSM to each state. Then we report on various types of sensitivity analysis.

6.1. *Main Results.* Figure 3 shows 95 percent confidence intervals for $\widehat{\beta}$ for each state from the marginal structural model in (15). We computed standard errors as if the weights were known, which results in valid but potentially conservative inference as long as the weight models are correctly specified (Tsiatis, 2007). The estimates are mostly negative, as would be expected, since higher A_s means less mobility. Interestingly, we find that there turns out to be little confounding due to past deaths, as the fits with and without the estimated weights (not shown) are very similar. Nevertheless, we keep the weights in all the fits as a safeguard. In Section 6.2 we investigate this further by doing a sensitivity analysis.

Figure 4 shows the estimated smooth function $\widehat{\nu}(t)$ in (15) for four states. The functions are strictly increasing and nearly, but not quite, linear. This is consistent with the usual epidemic dynamics where it is assumed that this component should grow linearly on the log-scale. The non-linearity probably reflects the fact that the probability $d(t)$ of dying decreases over time due to better hospital treatment, and the number of susceptibles to COVID-19 also decreases over time as recovered patients are likely immune for some period post-infection. Interestingly, if $\nu(t)$ is forced to be linear for all states (constant exponential growth), it causes many

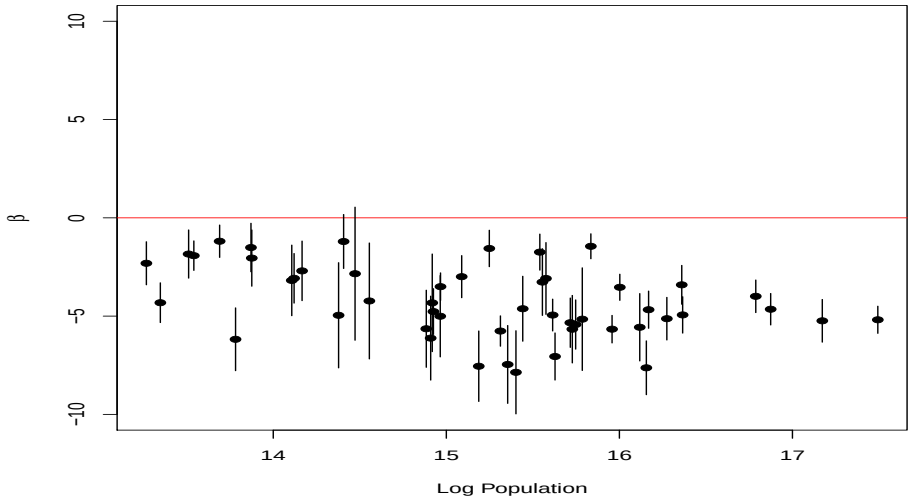

FIG 3. *Plot of $\widehat{\beta}$ and 95 percent confidence interval from the marginal structural model (15) for each state, versus state log population. A value of $\beta = -5$, for example, means that log deaths are reduced by 5 if A_s is increased by one percent at any time s.*

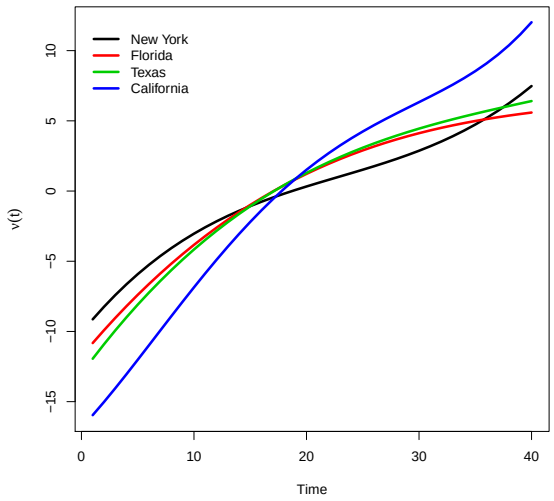

FIG 4. *Plot of $\widehat{\nu}(t)$ for four states.*

of the $\widehat{\beta}$'s to become positive which seems impossible. Allowing some non-linearity means that we only assume locally constant exponential growth which is both a weaker assumption than is usually made and leads to more reasonable estimates of β.

Next we consider counterfactual deaths $\theta_t = \exp(\mathbb{E}[L^{\bar{a}_T}])$ in (14) for the three mobility scenarios described at the end of section 4; two of them are shown in Figure 5 for four states. Figure 6 shows the estimates and pointwise 95 percent confidence bands for θ_t for these four states. The plots for all states are in the Supplement.

Finally, Figure 7 shows 95 percent confidence intervals for $\sum_t \exp\left(\mathbb{E}[L^{\bar{a}_t}]\right) - \sum_t Y_t$ and for $\left(\sum_t \exp\left(\mathbb{E}[L^{\bar{a}_t}]\right) - \sum_t Y_t\right) / \sum_t Y_t$ under the 'stay vigilant' scenario. We refer to these

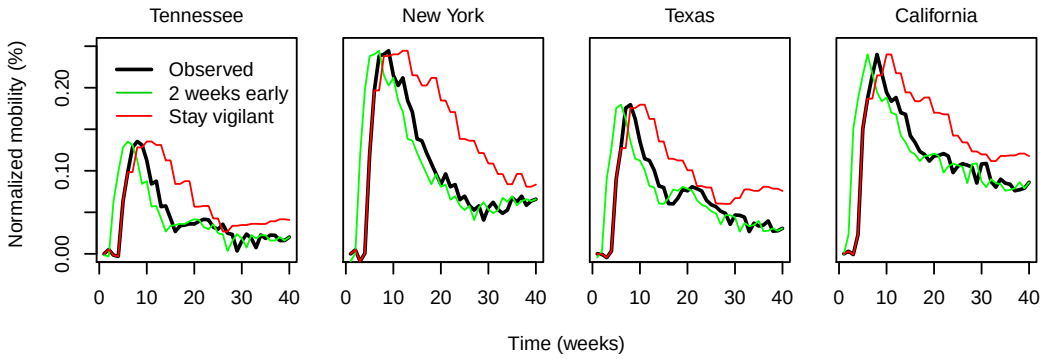

FIG 5. *The observed mobility curves and hypothetical interventions for four states. Mobility has been standardized to have value 0 at the beginning of the series. All plots are on the same scale.*

as total and relative excess deaths, where a negative excess means that lives would be saved. Of course, this number is larger for more populous states, although relative to the total number of observed deaths, all states small and large would have benefited equally from more sustained vigilance. Note that the confidence interval for New York (fourth from right) is very large. New York experienced the pandemic early and responded with large values of A_s so it is believable that further vigilance may not have a large effect.

6.2. *Sensitivity Analysis.* We have made a number of strong assumptions in our model. Our preference would be to weaken these assumptions and use nonparametric methods but the data are too limited to do so. Instead, we now assess the sensitivity of the results to various assumptions. We consider various perturbations of our analysis. These include: (1) changing the model/estimation method (we replace the MSM with an outcome model), (2) assessing the Markov assumption (which was used to estimate the weights), (3) checking the accuracy of the point mass approximation (which was used in Section 4.2 to simplify the model) and (4) assessing sensitivity to unmeasured confounding (we have assumed that the only time varying confounders are past values of mobility and death).

1. An Alternative Model. Here we compare the results from the MSM in (15) to the time series AR(1) outcome model:

$$L_t = L_{t-1} + \beta A_{t-\delta} + r(t) + \epsilon_t \tag{20}$$

where $r(t)$ is a polynomial of degree $k-1$. This says that, apart from random error, L_t differs from L_{t-1} for two reasons, mobility $A_{t-\delta}$ and the natural increase $r(t)$ due to epidemic dynamics. If we apply the g-formula in (1) to this model, we find $\mathbb{E}[L_t^{\bar{a}_t}] = \beta M(\bar{a}_t) + \nu(t)$ where $\nu(t) = \sum_{s=1}^{t} r(s)$ is a polynomial of order k. Hence, this outcome model is consistent with the MSM. In other words, this model is contained in the semiparametric model $\mathcal{P}$ defined in (8). This model resembles Robins' *blip models* (Robins (2000); Vansteelandt et al. (2014); Robins (2000)) as it measures the effect of one blip of treatment $A_{t-\delta}$ so we will refer to (20) as the blip model. We will fit (20) by least squares. There are three reasons for fitting this model. First, it as a point of comparison for the MSM. Second, as an outcome model, we are able to check residuals and model fit. Third, since it is an outcome model, we can use AIC to choose the degree $k-1$ of $r(t)$. We also use this choice of k in the MSM. The degree k chosen by AIC is typically $k=1$ for small states and $k=3$ or $k=4$ for the larger states. A plot of the selected degree versus log population and versus log deaths is in the supplementary material.

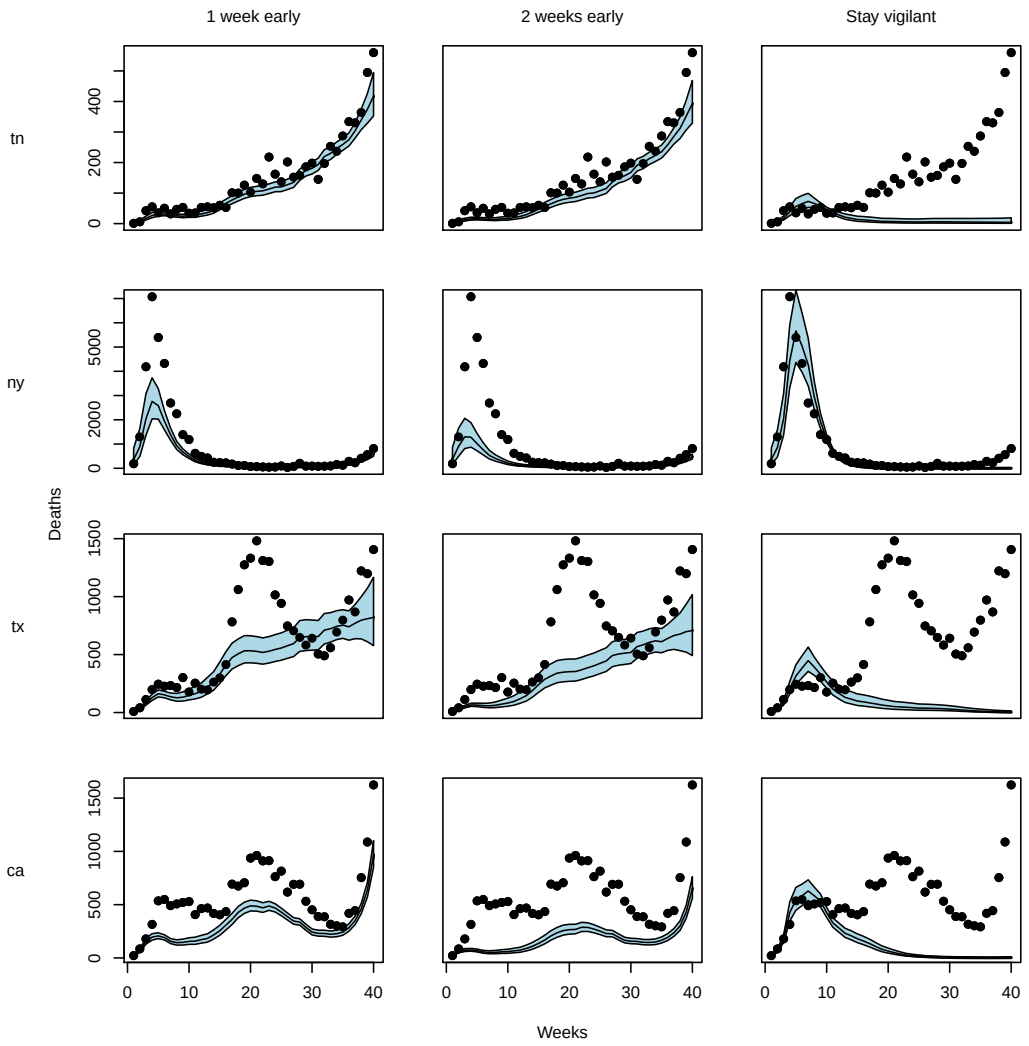

FIG 6. *Pointwise 95 percent confidence bands for deaths* $\theta_t = \exp(\mathbb{E}[L^{\overline{a}_T}])$ *for the three mobility scenarios* $\overline{a}_T$ *described at the end of section 4; see also Figure 5. Each row is a different state. Each column is a different scenario, start one week early, start two weeks early and stay vigilant. The epidemic in NY started early so staying at home sooner had a large impact. The same is true for PA, IL, MI, NJ, MA. Staying home earlier would not have had as much of an impact in states such as TN that did not suffer the epidemic early. Staying more vigilant would have had a large impact during every wave of infections. Some lack of fit in the early time period is evident in Texas where counterfactual deaths exceed observed deaths under 'stay vigilant' where mobility has not yet been changed.*

The left plot in Figure 8 shows the estimates of β and 95% confidence intervals for all the states from the blip model in (20), and the right plot compares the estimates of β from the MSM and blip models, where we see the similarity of the inferences. Since the blip model is an outcome model, it makes sense to compare the observed data to the fits. Fig 9 shows the fitted values and the data for four states. The fit is not perfect but is reasonable. We have examined the residuals and found that they show no sign of temporal dependence. There are some large outliers in some states, mostly in the first few weeks of the pandemic where mobility A_t and log deaths L_t change rapidly. Because of this we also fitted a robust regression but the results did not change much.

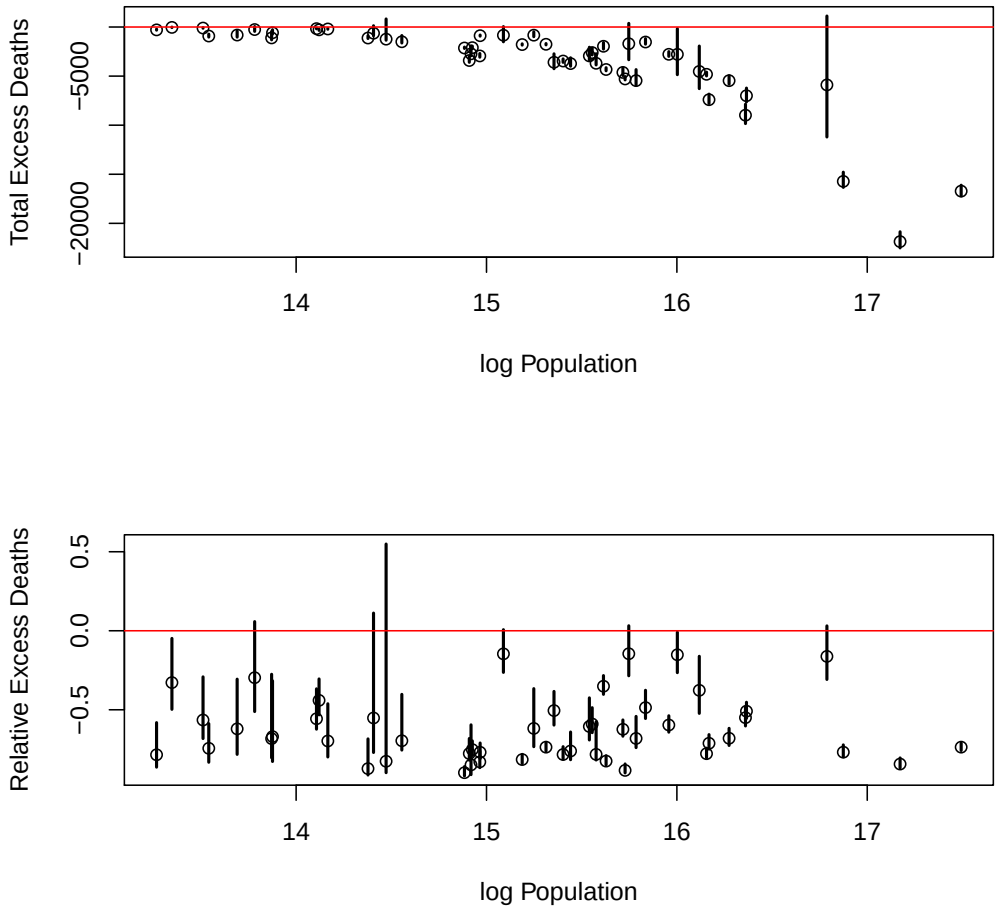

FIG 7. *95 percent confidence intervals for total excess deaths* $\sum_t \exp\left(\mathbb{E}[L^{\overline{a}_t}]\right) - \sum_t Y_t$ *(top) and relative excess deaths* $\left(\sum_t \exp\left(\mathbb{E}[L^{\overline{a}_t}]\right) - \sum_t Y_t\right) / \sum_t Y_t$ *(bottom) under the 'stay vigilant' scenario. The confidence intervals for NY (fourth from right) and a handful of other states include zero and suggests that staying more vigilant would not have significantly impacted the death toll. On the other hand, many states, small and large, could have reduced their death tolls by over a half.*

2. The Markov Assumption. In Section 5.2, to estimate the weights, we have made the Markov assumption that $A_{t-\delta}$ is conditionally independent of the past given $(A_{t-1-\delta}, L_{t-1-\delta})$. We also assumed that L_t is conditionally independent of the past given $(A_{t-1-\delta}, L_{t-1})$. To assess this assumption, we fit the models

$$A_{t-\delta} = \alpha_0 + \alpha_1 A_{t-1-\delta} + \alpha_2 A_{t-2-\delta} + \alpha_3 A_{t-3-\delta} + \beta_1 L_{t-1-\delta} + \beta_2 L_{t-2-\delta} + \beta_3 L_{t-3-\delta} + \epsilon_t$$

$$L_t = \alpha_0 + \alpha_1 A_{t-\delta} + \alpha_2 A_{t-\delta-1} + \alpha_3 A_{t-\delta-2} + \beta_1 L_{t-1} + \beta_2 L_{t-2} + \beta_3 L_{t-3} + \delta_t.$$

Figure 10 shows boxplots of the t-statistics for these parameters. The evidence suggests that the first order Markov assumption is reasonable. The weak dependence of A_t on past values of Y_t is consistent with the fact that we found that the weights W_t do not have a strong effect i.e. there is little confounding due to past deaths, However, this assessment still assumes that

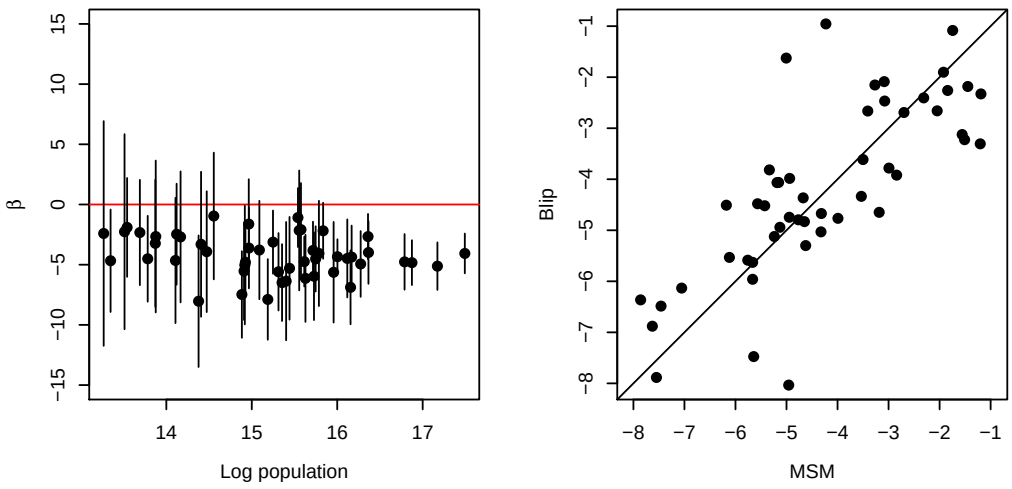

FIG 8. **Left:** *Estimates of β from the blip model in (20) with 95% confidence intervals.* **Right:** *Comparison of estimates of β from the blip model and the MSM in (15).*

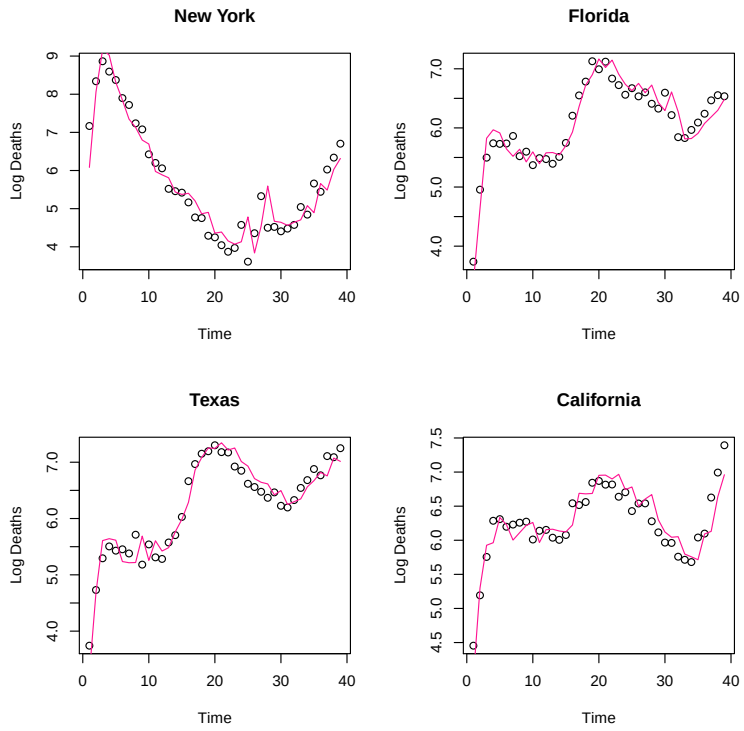

FIG 9. *Observed log deaths in four states as functions of time with estimates (red) from the blip model in (20).*

the Markov assumption is homogeneous, that is, that the law of A_t given (A_{t-1}, Y_{t-1}) is constant over time. This assumption is not checkable without invoking further assumptions.

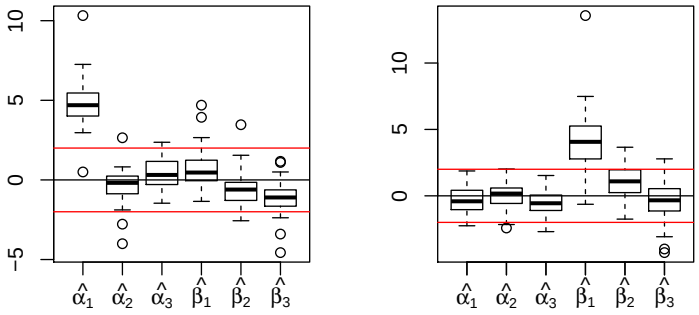

FIG 10. *(Left) Boxplots across states of t-statistics for the parameters in the model for A_t as a function of the past. The horizontal red lines are at ± 2. Only $\widehat{\alpha}_1$ is consistently significantly different from zero across states, suggesting that the times series of at home mobility A_t is a memory one process. (Right) Same for Y_t. Only $\widehat{\beta}_1$ is consistently significantly different from zero across states, suggesting that the deaths times series Y_t is a memory one process.*

3. Point Mass Versus Deconvolution. Recall that in Section 4.2 we approximated $f(s,t)$ with a point mass at $t-4$. An alternative to the point mass approximation is to estimate the number of infections I by deconvolution. From the number of infections, we can estimate the model parameters as in Section 5 without making the point mass approximation, using $\log(I)$ as the outcome variable. We infer $\widetilde{I}_t = d(t)I_t$ from the optimization:

$$\min_{I \geq 0} \|Y - F\widetilde{I}\|_2^2 + \lambda \sum_{r=2}^{T-1} (\widetilde{I}_r - \widetilde{I}_{r-1})^2, \tag{21}$$

where Y denotes the vector of weekly deaths and F is a matrix with (i,j)-entry equal to $f(i,j)$ if $j \leq i$ and zero otherwise; that is, F_{ij} is proportional to the probability of dying at time j given that infection occurred at time i. The parameter λ is user-specified and represents a penalty imposed on non-smooth solutions. Because f is proportional to the density of a Gamma random variable, we have $F_{ii} = f(i,i) = 0$. To ensure nonzero elements on the diagonal of F, we remove the first row and last column (all zeros) from F and solve (eq::lambda) using $Y = (Y_2, \ldots, Y_T)$, thus obtaining an estimate of $\widetilde{I} = (\widetilde{I}_1, \ldots, \widetilde{I}_{T-1})$. To enforce nonnegative values of I, we use the constrained optimization routine `L-BFGS-B` from `optim` in `R`. Using a penalty $\lambda = 1$, we report the inferred infections $\widehat{I}$ (red line) for California, Florida, New York and Texas in Figure 11 along with the implied deaths computed as $F\widehat{I}$. The latter match the observed deaths well, leading credence to this procedure. In Figure 12, we compare the estimates of β from the MSM using the point-mass approximation and those from the MSM using the estimates of infections from the deconvolution step. The estimates are in rough agreement as they lie near the diagonal.

4. Unmeasured Confounding. At time t, we treated $(A_1, Y_1), \ldots, (A_{t-1}, Y_{t-1})$ as confounders. Now suppose there is an unmeasured confounder U. We would like to assess $|\widehat{\beta}_U - \widehat{\beta}|$ where $\widehat{\beta}_U$ is the value of our estimate if we had access to U. This quantity is not identified and so any sensitivity analysis must invoke some extra assumption. Let $\Delta = |\widehat{\beta}_U - \widehat{\beta}|/\mathrm{se}(\widehat{\beta})$ denote the unobserved confounding on the standard error scale. So $\Delta = 0$ corresponds to no unmeasured confounding, $\Delta = 1$ corresponds to saying that the

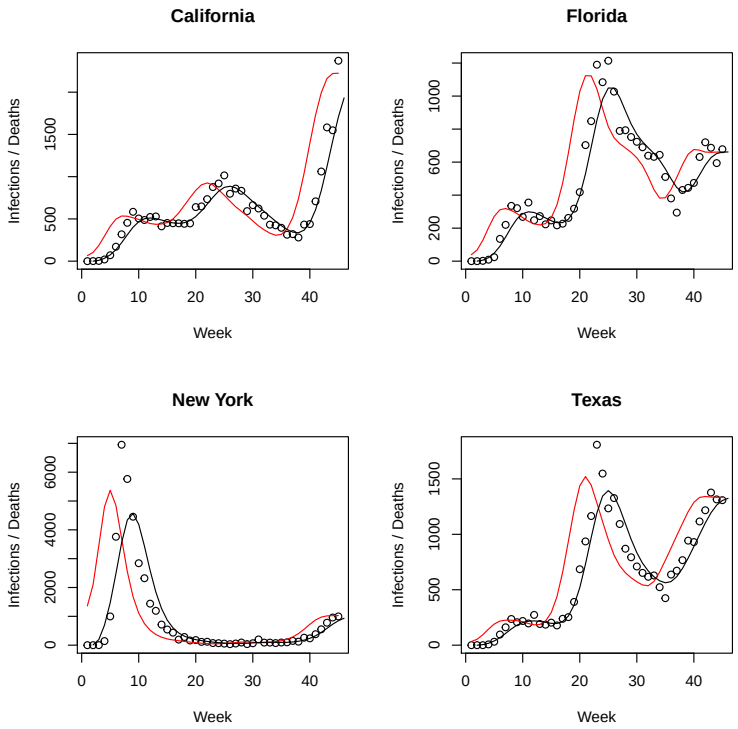

FIG 11. *Inferred infections in four states. The red curve is* $\widehat{\widetilde{I}}_t$*, the estimate of the number of infections times the probability of dying if infected by Covid-19,* $\widetilde{I}_t = d(t)I_t$*. The black curve is deaths* $F\widehat{I}$ *computed from the optimization with* $\lambda = 1$ *in (21), and the dots are the observed deaths.*

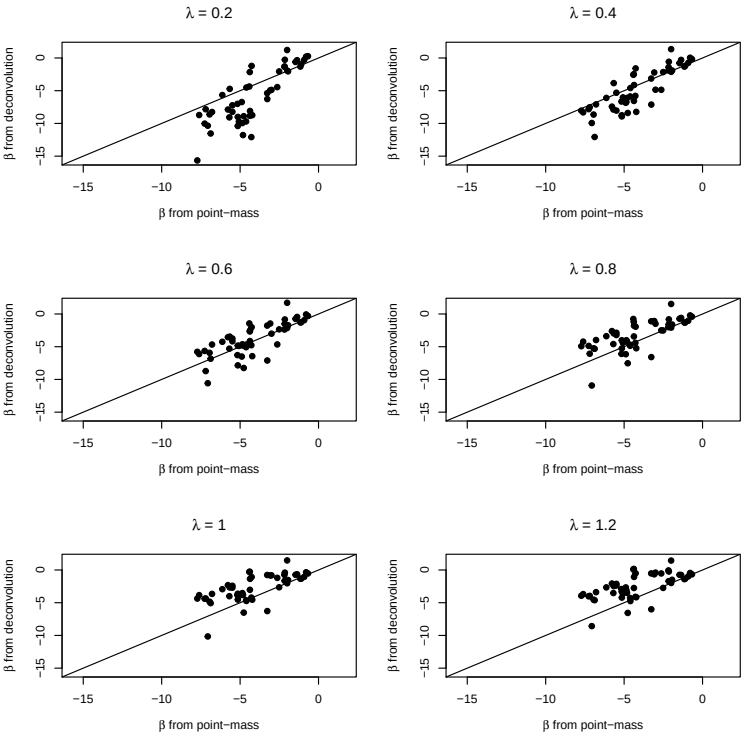

FIG 12. *Comparison of estimates of* $\widehat{\beta}$ *from the MSM using the point-mass approximation vs using estimates of infections via deconvolution for different values of* λ*.*

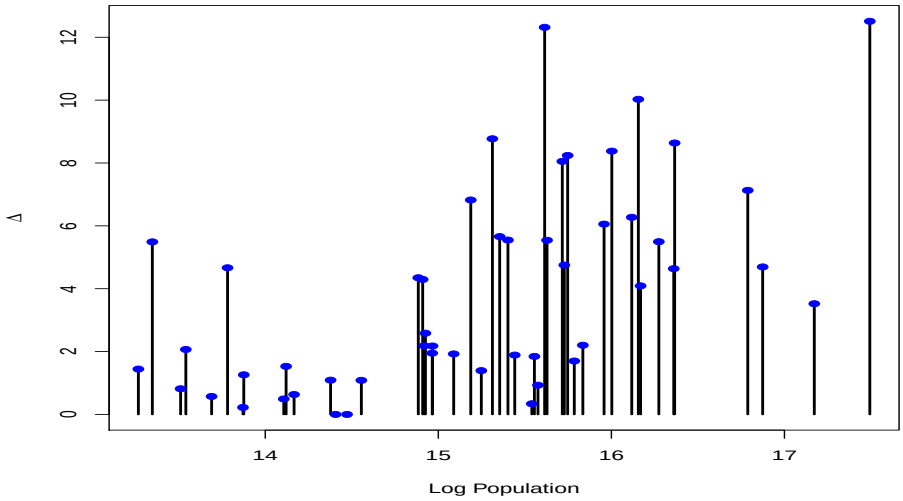

FIG 13. *Unmeasured confounding sensitivity plots. Minimum value of* Δ *versus log-population for each state, such that unmeasured confounding of size* $\Delta \operatorname{se}(\widehat{\beta})$ *causes the confidence interval for* β *to contain 0. For most states, it takes a fairly large* Δ *to lose statistical significance.*

unmeasured confounding is the same size as the standard error, etc. For each state, we enlarge the confidence interval by $\Delta \operatorname{se}(\widehat{\beta})$. We can then ask: how large would Δ have to be so that the enlarged confidence interval would contain 0. Figure 13 shows this critical Δ. We see that for most states, it takes a fairly large Δ to lose statistical significance. Some of the medium size states are the most robust. The larger states are not the most robust but still show substantial robustness.

Adding other potential within state confounders would be desirable but, in a within-state analysis, we can only accommodate time varying confounders. (A fixed confounder is a single variable with no replication and can only be used an a across state analysis.) So far we do not have any within-state time varying variables that would be expected to directly affect both A_t and Y_t. One could imagine that a variable like "the percentage of rural cases" could change over time and possibly affect both variables but we do not have such data.

Next we consider a second style of sensitivity analysis inspired by the approach in Rosenbaum et al. (2010). The effect of unmeasured confounding in our analysis is that the weights W_t are misspecified. If there are unobserved confounders U_t, then the correct weights are

$$\widetilde{W}_t = \prod_{s=1}^{t} \frac{\pi(A_s|\overline{A}_{s-1})}{\pi(A_s|\overline{A}_{s-1}, \overline{Y}_{s-1}, \overline{U}_{s-1})}$$

whereas we estimated the weights

$$W_t = \prod_{s=1}^{t} \frac{\pi(A_s|\overline{A}_{s-1})}{\pi(A_s|\overline{A}_{s-1}, \overline{Y}_{s-1})}.$$

To assess this impact we find the maximum and minimum $\widehat{\beta}$ under the assumption that

$$\frac{W_t}{\Gamma} \leq \widetilde{W}_t \leq \Gamma W_t$$

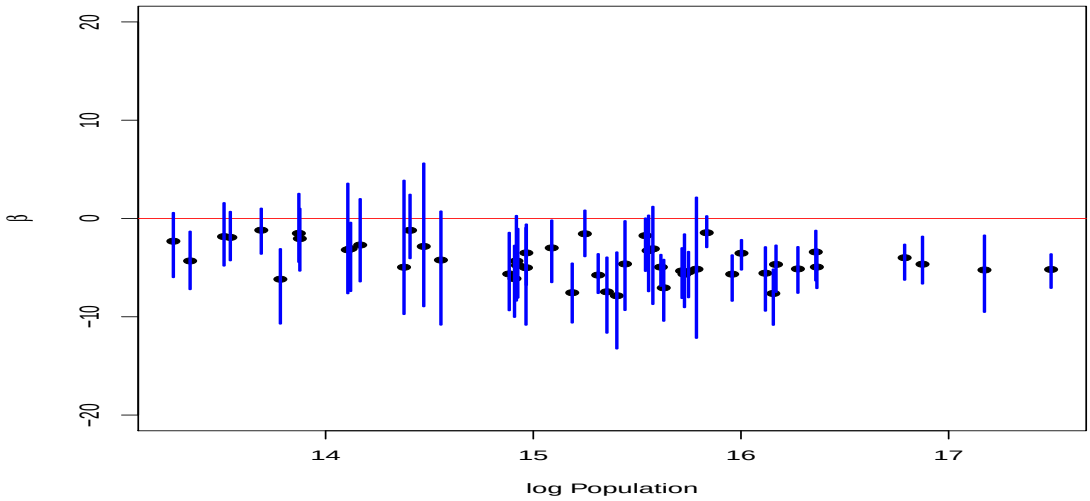

FIG 14. *Unmeasured confounding sensitivity plots. The blue line segments span the lower and upper bounds of* $\widehat{\beta}$ *over the weights* $1/\Gamma \leq \widetilde{W}_t/W_t \leq \Gamma$ *with* $\Gamma = 3$*. The black dots are the original point estimates. The effects for most large and medium states remain significant, indicating robustness to unmeasured confounding.*

for $t = 1, \ldots, T$ and some $\Gamma \geq 1$. Similar ideas for static, binary treatments have been considered in Zhao, Small and Bhattacharya (2019); Yadlowsky et al. (2018). Figure 14 shows the bounds on $\widehat{\beta}$ using $\Gamma = 3$. Even with this fairly large value of Γ the effects for most large and medium states remain significant indicating robustness to unmeasured confounding. (The method for computing the bounds is in Bonvini et al. (2021).)

6.3. *Across Versus Within States.* We have focused on within state estimation. An alternative is to fit a model across states as well. Although we are skeptical of combining data over states we do so here for completeness. We fit the blip model with common β and, rather than include state level covariates such as population size, proportion of residents in cities, etc., we use a fixed effect for each state. The resulting estimates of β and standard errors for $k = 1, 2, 3, 4$ are:

k	$\widehat{\beta}$	standard error
1	-5.20	0.27
2	-4.60	0.27
3	-3.82	0.34
4	-2.83	0.43

The estimates are consistent with the within state models. AIC chooses $k = 1$, which conflicts with the within state analysis with favors larger k for larger states. The likely reason is that combining states adds variability in the combined dataset since β's and $\nu(t)$'s are different between states, so there is less signal compared to the noise to estimate a more complicated relationship than a linear. A natural extension of this model is to use a random effects approach, although we do not pursue that here.

7. Discussion. Our approach to modeling the causal effect of mobility on deaths is to construct a marginal structural model whose parameters are estimated by solving an estimat-

ing equation. We model each state separately to reduce confounding due to state differences. Our approach has several advantages and disadvantages.

Our modeling assumptions are reasonable in the short term but not in the long term. Eventually, the effects of acquired immunity, masks, vaccinations etc might have to be accounted for.

Estimating the model parameters comes down to solving the estimating equation (16). Computing standard errors and confidence intervals is then straightforward. This is in contrast to more traditional and Icarian epidemic modeling which requires estimating many parameters using grid searches or MCMC. Provably valid confidence intervals are elusive for those methods. On the other hand, the more detailed models might be more realistic and can capture effects that our simple model cannot capture. Moreover, our inferences are asymptotic in nature. When comparing exact Bayesian methods to approximate frequentist methods it is hard to argue that one approach is more valid than the other.

We believe that focusing on weekly data at the state level gives us the best chance of getting data of reasonable quality and helps avoid confounding related to state differences. Further, this allows the causal effect to vary between states. But this results in a paucity of data, a few dozen observations per state. This limits the complexity of the models we can fit and it requires that we make a homogeneous Markov assumption. A natural compromise worthy of future investigation would be to use some sort of random effects model to allow modeling all states simultaneously. This could also permit using data from other countries. At any rate, there is a tradeoff: within state analysis requires stronger modeling assumptions while analyzing all states together requires assuming independence and it assumes we can model all sources of between state confounding.

Detailed dynamic modeling versus the more traditional causal modeling done here (and in Chernozhukov, Kasaha and Schrimpf (2020)) represent two different approaches to causal inference for epidemics. It would be interesting to see a general comparison of these approaches, perhaps eventually leading to some sort of fusion of these ideas.

Finally, we discuss a general issue about causal epidemic modeling that we mentioned earlier, namely, that some epidemic models are subject to a problem known as the *null paradox* which can be avoided by using MSM's. This problem is well known in the causal literature (Robins (1986, 1989); Robins and Wasserman (1997)) but may not be well known in the epidemic literature.

We'll focus on infections I_t. Consider a sequentially specified epidemic model of the form

$$\mathbb{E}[I_t|\overline{I}_{t-1}, \overline{A}_{t-1}, \overline{X}_{t-1}] = f(\overline{I}_{t-1}, \overline{A}_{t-1}, \overline{X}_{t-1}, \theta_t)$$

for some parametric model $f(\cdot;\theta_t)$. Here $(X_1, \ldots, X_T)$ are covariates and $(A_1, \ldots, A_T)$ is the variable of interest which is mobility in this paper but could be mask wearing, school closings, vaccines, etc. Also, let $g(x_t|\overline{X}_{t-1}, \overline{I}_{t-1}, \overline{A}_{t-1}; \alpha_t)$ be a model for X_t given the past. The model is specified by the parameters $\Theta = ((\theta_t, \alpha_t) : t = 1, \ldots, T)$.

Assume that there is no unobserved confounding. Even without confounding, there will typically still be many unobserved baseline variables U that affect the X_t's and the I_t's. For example, overall health could affect infections I_t and covariates X_t. These are not confounders because they do not directly affect $\overline{A}_T$, thus the causal effect is still identifiable and is still given by the g-formula. These baseline variables create conditional dependence between I_T and $\overline{A}_t$ given the X_s's, even when the treatment A has no effect. (For readers familiar with directed graphs, this can be seen by noting that the X_t's are colliders; see Figure 1 in Robins and Wasserman (1997).) We will call this *baseline dependence*.

The causal effect $\psi(\overline{a}_T)$ can be derived by inserting the model into the g-formula. Now suppose that the treatment has no effect which we call the null case. Then $\partial\psi(\overline{a}_T)/\partial a_s = 0$ for all a_s, i.e., $\psi(\overline{a}_T)$ should not depend on $\overline{a}_T$.

Now we can informally state the null-paradox. If the models have non-linear terms then there is no value of the parameters such that (i) there is non-zero baseline dependence and (ii) the null is true. Thus, in some sense, the model is trapped into being misspecified.

We can avoid the null paradox by using nonparametric models but then the model complexity explodes as T increases leading to the curse of dimensionality. Linear models avoid the null paradox but caution is still needed since the causal effect $\psi(a)$ involves complicated nonlinear functions of the regression parameters. Hence, the model is very difficult to interpret and the individual regression parameters do not have a causal interpretation. (Note that our autoregressive blip model is linear but is simple enough so that there is a single, interpretable causal parameter.)

The quickly growing literature on using sequentially specified epidemic models does include such models. MSMs avoid the null paradox, and this is another reason for using MSMs (or some other semiparametric causal model such as structural nested models). In our case we motivated the MSM by starting with a sequentially specified model. This seems like a reasonable approach for using epidemic models to define an MSM but there may be other approaches as well.

Acknowledgments. The authors would like to thank Rob Tibshirani for providing helpful feedback on an earlier draft of the paper. Edward Kennedy gratefully acknowledges support from NSF Grant DMS1810979.

SUPPLEMENTARY MATERIAL

Supplement A: Plots for all states.
(). Plots of the data and counterfactual curves for all states.

Supplement B: AIC plots.
(). Plots of the value of k selected by AIC.

Supplement C: Deconvolution.
(). Plots of the deconvolved data for all states.

REFERENCES

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973.

BJØRNSTAD, O. N. (2018). Epidemics. *Models and data using R: Springer International Publishing* 318.

BONVINI, M., KENNEDY, E., VENTURA, V. and WASSERMAN, L. (2021). Propensity Scores and Sensivity Analysis for Marginal Structural Models with Continuous Treatments. *In preparation*.

BRAUER, F., CASTILLO-CHAVEZ, C. and CASTILLO-CHAVEZ, C. (2012). *Mathematical models in population biology and epidemiology* **2**. Springer.

CHANG, S., PIERSON, E., KOH, P. W., GERARDIN, J., REDBIRD, B., GRUSKY, D. and LESKOVEC, J. (2020). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 1–6.

CHERNOZHUKOV, V., KASAHA, H. and SCHRIMPF, P. (2020). Causal impact of masks, policies, behavior on early COVID-19 pandemic in the US. *arXiv preprint arXiv:2005.14168*.

FONG, C., HAZLETT, C., IMAI, K. et al. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics* **12** 156–177.

IHME (2020). Modeling COVID-19 scenarios for the United States. *Nature Medicine*.

KERMACK, W. O. and MCKENDRICK, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* **115** 700–721.

LIAO, X. and MEYER, M. (2018). cgam: Constrained generalized additive model. *xxxx* **xx** xxxx.

MEYER, M. C. et al. (2008). Inference using shape-restricted regression splines. *The Annals of Applied Statistics* **2** 1013–1033.

MEYER, M. C. et al. (2018). A framework for estimation and inference in generalized additive models with shape and order restrictions. *Statistical Science* **33** 595–614.

NEUGEBAUER, R. and VAN DER LAAN, M. (2007). Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference* **137** 419–434.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure periodâĂŤapplication to control of the healthy worker survivor effect. *Mathematical modelling* **7** 1393–1512.

ROBINS, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS* 113–159.

ROBINS, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials* 95–133. Springer.

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology.

ROBINS, J. M. and WASSERMAN, L. (1997). Estimation of Effects of Sequential Treatments by Reparameterizing Directed Acyclic Graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* 409-420. Morgan Kaufmann.

ROSENBAUM, P. R. et al. (2010). *Design of observational studies* **10**. Springer.

SHI, Y. and BAN, X. (2020). Capping Mobility to Control COVID-19: A Collision-based Infectious Disease Transmission Model. *medRxiv*.

TSIATIS, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

UNWIN, H. J. T., MISHRA, S., BRADLEY, V. C., GANDY, A., MELLAN, T. A., COUPLAND, H., ISH-HOROWICZ, J., VOLLMER, M. A., WHITTAKER, C., FILIPPI, S. L. et al. (2020). State-level tracking of COVID-19 in the United States. *Nature communications* **11** 1–9.

VANSTEELANDT, S., JOFFE, M. et al. (2014). Structural nested models and G-estimation: the partially realized promise. *Statistical Science* **29** 707–731.

XIONG, C., HU, S., YANG, M., LUO, W. and ZHANG, L. (2020). Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences* **117** 27087–27089.

YADLOWSKY, S., NAMKOONG, H., BASU, S., DUCHI, J. and TIAN, L. (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*.

ZHAO, Q., SMALL, D. S. and BHATTACHARYA, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81** 735–761.

ZHOU, X. and WODTKE, G. T. (2018). Residual balancing weights for marginal structural models: with application to analyses of time-varying treatments and causal mediation. *arXiv preprint arXiv:1807.10869*.