Statistical Inference. Date of this draft: May 26, 2020.

Data

Available data vary by state and time window. Commonly available data are

$$\begin{aligned} x_{a,t} &\quad \text{daily new case counts by age group } a \\ y_{a,t} &\quad \text{daily new hospitalizations by age group } a \\ z_{a,t} &\quad \text{daily new hospital deaths by age group } a \\ w_{a,t} &\quad \text{daily new out-of-hospital deaths by age group } a \\ u_{a,t} &\quad \text{daily new hospital discharges by age group } a \end{aligned} \tag{1}$$

and

$$\begin{aligned} h_t &\quad \text{current number of hospitalized individuals at time } t \text{ (includes } c \text{ and } v \text{ below)} \\ c_t &\quad \text{current number of individuals in the ICU at time } t \text{ (includes } v \text{ below)} \\ v_t &\quad \text{current number of individuals on a ventilator at time } t \end{aligned} \tag{2}$$

For some states, and some time windows, age-structured data are not available, but summed data over all age groups is available. This is common for daily new case counts and daily new hospitalizations and can occur for all variables. We refer to these summed data as

$$\begin{aligned} x_t = \textstyle\sum_a x_{a,t} &\quad \text{daily total new case counts summed over all age groups} \\ y_t = \textstyle\sum_a y_{a,t} &\quad \text{daily total new hospitalizations summed over all age groups} \end{aligned} \tag{3}$$

Age-structured counts of new cases $x_{a,t}$ and new hospitalizations $y_{a,t}$ can be available at the same time as the corresponding total counts x_t and y_t, and it is currently common for the age-structured counts to record fewer individuals than the summed total counts. That is,

$$x_t > \sum_a x_{a,t} \tag{4}$$

In these cases, we assume that individuals recorded in the total summed count data but not in the age-

structured count data are missing at random, and impute them as part of the inference. More details are below in the Inference section.

Models and Likelihoods for Observations

Daily New Count Data

Daily new case counts are assumed to follow a negative binomial distribution with dispersion parameter (shape parameter) k_1. The expected mean number of new symptomatic cases, $\rho \cdot \tilde{J}_{a,t}(\theta)$, is obtained from the differential equations described in **Mathematical Transmission Model**, with θ being the parameters that are input into the ODE model (i.e., contact rate, length of hospital stay, etc) and ρ is the fraction of symptomatic cases that are expected to be observed by the health system. The $\sim$ symbol is used to indicate that

$$\tilde{J}_t \;=\; J_t \;-\; J_{t-1}$$

is the daily incidence while the J_t-variables are the cumulative incidence. The likelihood of $\{x_{a,t},\ t \in 1,\ldots,t_f\}$ given the ODE model parameters, the dispersion parameter k_1, and the reporting rate ρ is a negative binomial likelihood and can be written

$$L_1(\,\{x_a,t\}\,|\,\tilde{J}_{a,t}(\theta)\,,\,k_1\,,\,\rho\,) \;=\; \prod_{t=1}^{t_f}\prod_{a=0}^{8}\frac{\Gamma(\,k_1+x_{a,t}\,)}{\Gamma(\,x_{a,t}+1\,)\cdot\Gamma(\,k_1\,)}\left(\frac{\rho\tilde{J}_{a,t}(\theta)}{\rho\tilde{J}_{a,t}(\theta)+k_1}\right)^{x_{a,t}}\left(\frac{k_1}{\rho\tilde{J}_{a,t}(\theta)+k_1}\right)^{k_1}$$

where the products are taken over our nine age classes (a) and over all time points t for which data are available. The time $t = 1$ is January 1 2020.

Daily New Hospitalization Data

Similarly, let $\tilde{K}_{a,t}$ represent the daily incidence of hospitalization by age group a. No reporting parameter is associated with hospitalization incidence since all cases are assumed to be reported. The likelihood of the daily new hospitalization data $\{y_{a,t},\ t \in 1,\ldots,t_f\}$ given the parameters is

$$L_2(\,\{y_{a,t}\}\,|\,\tilde{K}_{a,t}(\theta)\,,\,k_2\,) \;=\; \prod_{t=1}^{t_f}\prod_{a=0}^{8}\frac{\Gamma(\,k_2+y_{a,t}\,)}{\Gamma(\,y_{a,t}+1\,)\cdot\Gamma(\,k_2\,)}\left(\frac{\tilde{K}_{a,t}(\theta)}{\tilde{K}_{a,t}(\theta)+k_2}\right)^{y_{a,t}}\left(\frac{k_2}{\tilde{K}_{a,t}(\theta)+k_2}\right)^{k_2}$$

Death Data

Let $\tilde{D}_{a,t}$ be the model's daily incidence of hospital deaths and $\tilde{W}_{a,t}$ the model's daily incidence of non-hospital deaths. Similar likelihood equations can be written down for the model's death counts (L_3 and L_4). If hospital and non-hospital deaths are not broken down into these two groups, a single likelihood function can be used for the death count.

Our current model fits do not utilize death count data in model fitting. This will be added soon.

Current Hospitalized Individuals

The model tracks current hospitalizations (H_t, summed across age classes), current ICU occupancy (C_t), and current number of patients using ventilators (V_t). Note that in the model these are disjoint groups, so H_t represents hospitalized patients that are not in the ICU. And, C_t represents ICU patients that are not on ventilators. The data on the other hand are typically nested so that h_t includes all hospitalized patients including ICU patients c_t and patients on ventilators v_t. The likelihood equation accounting for current hospitalizations is

$$L_3(\{h_t - c_t\} \mid H_t(\theta)\ ,\ \sigma_1\ ) \ = \ \prod_{t=1}^{t_f} \varphi_{\sigma_1}(\ H_t - (h_t - c_t)\ )$$

where φ_{σ_1} is a normal PDF with mean zero and standard deviation σ_1. Accounting for the current ICU and ventilator occupancy gives

$$\begin{aligned} L_4(\{c_t - v_t\} \mid\ C_t(\theta)\ ,\ \sigma_2\ ) \ &= \ \prod_{t=1}^{t_f} \varphi_{\sigma_2}(\ C_t - (c_t - v_t)\ ) \\ L_5(\{v_t\} \mid\ V_t(\theta)\ ,\ \sigma_3\ ) \ &= \ \prod_{t=1}^{t_f} \varphi_{\sigma_3}(\ V_t - v_t\ ) \end{aligned}$$

and the product of L_1 through L_5 gives a likelihood based on the seven data types in (1) and (2).

Parameter Models

The ODE model described in **Mathematical Transmission Model** takes a number of parameters as inputs. For almost all parameters we assign uniform prior distributions with values defined either by allowable limits or by reported estimates in the existing literature.

The reporting rate parameter ρ is given a uniform prior with limits between 0 and 1. Parameters that

control hospitalization rates and length of hospital and ICU stays are given uniform priors that include values reported in the existing literature, but allow for variation in these rates across states.

The contact rate parameter β_t controls the rate of contact between individuals, and is allowed to vary over time through a penalized spline expansion. We use a cubic B-spline expansion with one knot every seven days. If $s_\ell(t)$ is the ℓ-th B-spline basis function evaluated at time t, then

$$\beta_t = \sum_{\ell=1}^{L} \alpha_\ell s_\ell(t) \tag{5}$$

with $\alpha_1, \ldots, \alpha_L$ being the spline basis function loadings, which are parameters which need to be estimated. To model temporal consistency in contact rate parameters, we specify a correlated multivariate normal prior distribution on $\alpha_1, \ldots, \alpha_L$, with correlation defined to penalize second differences between the loadings $\alpha_1, \ldots, \alpha_L$. The variance of the multivariate normal prior is controlled by a variance parameter, σ^2_β, which is given an inverse gamma prior. Together, this hierarchical prior provides a flexible model for changing contact rates over time, with smoothness of the contact rates estimated from the data.

Parameters controlling the variance of data models around the mean defined by the ODE model come in two forms. Observed hospitalization data are given Gaussian likelihoods (L_3, L_4, L_5) with variance parameters to be estimated by the data. These variance parameters are assigned inverse-gamma priors. Observed count data are given negative binomial likelihoods (L_1, L_2) have dispersion parameters which are assigned exponential priors.

Inference

Inference on parameters controlling the ODE model, such as contact rate parameters and hospitalization stay parameters, as well as parameters that control variation in the observed data around ODE model means, is conducted under a Bayesian approach, with samples from the posterior distribution of model parameters obtained using a Markov chain Monte Carlo (MCMC) algorithm. All parameters that affect the underlying ODE model are proposed jointly. All other parameters are proposed one at a time. Missing values in age-structured new case counts and age-structured new hospitalization counts are imputed at each MCMC iteration conditioned on the current parameter values and ODE trajectory. Five separate Markov chains are run, and convergence is assessed visually.

Prediction

Prediction of the future state of the epidemic is done using draws from the posterior distribution of model parameters. The ODE model must be run into the future, and this requires future values of the contact rate parameter β_t. As our model allows the contact rate to vary over time, we consider predictions based on a range of contact rate parameters. For each sample of the posterior distribution, we keep future values of β_t fixed at the value of β_t for the last day in which data is used to fit the model. The spline model we use for β_t results in a wide range of possible values in the posterior distribution of the contact rate at the last day of observations, reflecting the possibility of contact rates changing quickly over time, as they have in the past (i.e., large decreases in contact rates as states implemented social distancing guidelines in March).

We summarize predictions by considering the posterior predictive distribution of model predictions conditioned on the contact rate parameters staying within three defined ranges. Under a "low" contact rate scenario, we provide posterior predictions conditioned on β_t remaining below 30% of the average β_t over the first week of March. Under a "medium" contact rate scenario, we provide posterior predictions conditioned on β_t remaining above 30% but below 60% of the average β_t over the first week of March. Under a "high" contact rate scenario, we provide posterior predictions conditioned on β_t remaining above 70% of the average β_t over the first week of March.