# Predicting COVID-19 Cases and Subsequent Hospital Burden in Ohio

OSU / IDI* COVID-19 Response Modeling Team

April 7, 2020

## 1 Introduction

Coronavirus 2019 (COVID-19) disease has resulted in 1.13 million confirmed cases and 60,115 deaths reported globally as of 4 April 2020 [31]. As evidenced by epidemics in other countries, particularly in Italy and Spain, COVID-19 patient case loads have the potential to overwhelm healthcare systems [15].

The Ohio State University IDI working in conjunction with CPH[1], OSU Department of Mathematics, and SI[2] established a working relationship with ODH[3] to act as a service to the State, beginning in 2018. Based on this initial collaborative relationship, the EEPH program[4] within IDI took the initiative to offer epidemic modeling and decision analytics support to ODH preparatory planning and response to the COVID-19 pandemic, specifically the Ohio epidemic.

Just as in any computational modeling effort, there are many approaches and methods to choose from to model emerging epidemics. Different methods are suited to different situations in terms of data, scenarios, conditions, and urgency. This is no different for modeling the COVID-19 pandemic, either nationally or for the state of Ohio. Modeling approaches include: compartmental models [2,12,30], statistical models [17], and agent-based simulations [14]. For the state of Ohio, the OSU-IDI group has approached the modeling challenge on two fronts by developing: 1) predicted statewide estimates of a times series of COVID-19 incidence, and 2) a geographic component that transforms the output of the statewide model to hospital burden by county or "hospital catchment" area.

## 2 Methods

### 2.1 Model Overview

The predictive statewide model for Ohio is comprised of a dynamic network model [23], where nodes correspond to individuals and network edges correspond to connection through proximity. We highlight three key distinguishing features of the statewide model:

1. The model considers a dynamic network where the edges can be deactivated over time—supporting social distancing impacts more accurately.

*Infectious Disease Institute - The Ohio State University
[1]College of Public Health - The Ohio State University
[2]Sustainability Institute - The Ohio State University
[3]Ohio Department of Health
[4]Ecology Epidemiology and Population Health

2. We use a law of large numbers to derive a set of differential equations that describes the disease process on a large network [19] without requiring simulation methods.

3. The solutions of these differential equations can be used to estimate model parameters using a principled statistical approach based on survival analysis. This approach is based on writing an explicit likelihood for these parameters given data on times of illness onset [29]. Critically, this allows for more accurate quantification of the uncertainty in the model predictions.

Because of these features, this approach retains the tractability of an analytical model while incorporating features of complex contact networks to better represent social interactions and distancing.

The predictive statewide model takes data on illness onset dates of confirmed cases as input and produces estimates of COVID-19 incidence (*i.e.*, new cases) in Ohio over time. This output is not age-stratified, but the age distribution of the new cases is assumed to match the age distribution of confirmed cases when predicting the number of hospitalizations, which occurs in the next step of the model.

Estimates of the number of hospitalized COVID-19 cases are derived in the geographic portion of the model. Because illness severity and the risk of hospitalization for COVID-19 patients varies by age and comorbidities [10, 28], we use age structure and population density to distribute case counts from the statewide epidemic model across smaller geographic areas within the state. We use the age distribution within each geographic unit to predict hospitalization rates over time. Consequently, counties or hospital catchment areas that have a different demographic structure (*e.g.*, older or younger populations) will differ in their COVID-19 hospitalizations over time.

## 2.2 Detailed Methods of Statewide Model

There are challenges to using traditional compartmental models [27] to estimate future COVID-19 incidence. This includes challenges in estimating the effective population size on which the epidemic takes place (directly applying compartmental models using e.g. the entire state's population size can lead to dramatic over-estimates of illness), and lack of data on mild or asymptomatic infection.

In our statewide model, we use an approach called *Dynamic Survival Analysis* (DSA), which is an extension of *survival dynamical systems* [6, 29]. Details on this approach and the model development can be found in Appendix A.

Three key strengths of the DSA approach for predicting novel virus epidemics such as COVID-19 are:

- It does not require knowing the size of the susceptible population,
- It does not require information on overall disease prevalence in the population,
- It does not require prior knowledge of the shape of the epidemic curve.

Since testing has focused on the most symptomatic and severe cases, we have almost no information on the number of asymptomatic infections or those with less severe symptoms who are not tested. Because of the novel nature of this virus and the resulting pandemic, analysis of the epidemic curve cannot be based on previous epidemics caused by other viruses such as SARS-CoV-1 or influenzas. Because it relies on illness onset times rather than counts of new cases, the DSA method can incorporate a partial epidemic curve like the one produced by testing for COVID-19 within Ohio, which is affected by both limited testing capacity and undetected asymptomatic infections. The DSA approach has been developed to handle such limited data in a manner that allows accurate quantification of uncertainty. The method is strengthened by its simplicity; in

most practical circumstances (see [19]), it requires only a single differential equation. Parameters for the model are optimized using empirical temporal data on new illnesses, and the model output is a time series of expected future illnesses.

The optimization is accomplished using maximum likelihood estimation (MLE). This method provides for a consistent estimator for both small and large data sets, a valuable asset in numerical optimization. This optimization is done using standard quasi-Newtonian algorithms that reach convergence using the negative log-likelihood, which is more numerically stable than the likelihood function itself. Due the availability of reliable and efficient MLE algorithms in Python and other languages, the MLE is a commonly used method of function optimization.

Our predictive statewide model provides robust estimates of cases over time given partially observed daily counts of new illnesses. This is ideal in a setting where testing capacity is limited or changing due to constraints on lab capacity or detection limits. The counts of new illnesses are often known as an observed *epidemic curve* in the literature [32]. Our approach is derived from the general stochastic model of a pathogen spread across a contact network where the nodes represent individuals in a community [19]. As a working model of a contact network we use a type of random graph called a dynamic configuration model (CM) [7], which is sometimes referred to as a *pairwise model* [18].

**Assumptions.** All models have specific assumptions used to develop and implement them. First, we assume that each individual in the network (network node) has a number of neighbors (their degree) drawn from a degree distribution. Subsequently, local Markovian infectious pressure changes their status from *susceptible* ($S$) to *infective* ($\underline{\mathrm{I}}$) to *removed* ($R$). We further assume that the $R$ individuals are no longer able to transmit the infection and cannot be reinfected—an unknown, but readily assumed component of contemporary models. For the dynamics of the network model, individuals transition between the $S$, $I$, and $R$ compartments based on the following principles that allow the extraction of a mathematical model of the ongoing epidemic based on observable data:

1. Disease spread occurs over a network of contacts, that is, an infectious individual can infect their immediate neighbors at a fixed rate (the rate must be greater than 0). It is implicitly assumed that the average number of a person's contacts is also greater than 0.

2. Each infected individual recovers (or dies) from infection at rate that is $> 0$, **or** is restricted from contacting their network neighbors through mandatory or voluntary isolation or social distancing of its neighbors at a rate that is greater than 0.

3. Once infected, an infected individual has an infectious period that has an exponential distribution.

4. People who are ill remain infectious, and a partial count of new illnesses is observed over time with a negligible chance of misdiagnosis (*i.e.*, false positives).

**Estimating Transmission Rates.** The complete model description and detailed development can be found in Appendix A. In brief, the statistical model is used to estimate the number of people and timing of transfer between states $S$, $I$, and $R$. Then through substitution, a term $S_t$, the number of susceptible people at time $t$ is developed. Embedded within this $S_t$ is an improper survival function for the *time to the onset of illness in a random susceptible person.* Using this survival function approach, the time to infection of a randomly selected susceptible person within a large population follows a temporal pattern determined by probability laws. After we estimate the time series of susceptibles, we can estimate the probability of a randomly selected susceptible

individual being infected *during the lifetime of an epidemic*. From this we develop a conditional probability of remaining susceptible past time $t$.

**Estimating Dropout and Recovery Rates.** The impact of social distancing is accomplished via a generalized approach to a person being dropped from the the network. As has been visualized elsewhere, when a person is removed from the network, then their neighbors and contacts within the network are removed, which limits transmission. This is accomplished by estimating the rate of infectious contact within a network ($\tilde{\beta}$ in Appendix A), and then considering drop outs as a function of recovery rate (details in Appendix A).

## 2.3 Geographic Modeling Methods

### 2.3.1 Estimating Hospital Census over Time

After the statewide model produces time series estimates of cases across the state, these are then translated into estimates of hospitalizations and subsequent ICU admission. Details for this portion of the model are given in Appendix B. The conversion of statewide illness onsets from the statewide model to hospital census over time in each geographic region involves three steps:

1. Estimation of case onset time series for each age group based on age stratification in each geographic area modeled.
2. Estimation of the number of severe cases that will need hospitalization, using age and other risk factors [10, 11].
3. Use probability distributions for time from illness onset to hospitalization, and for length of stay in hospital, to estimate hospital census numbers over time. This provides a means of modeling the time lag between symptom onset and when hospitalization would occur based on observed rates of hospitalizations nationally, and also incorporates different lengths of stay for non-ICU vs. ICU patients.

Once daily hospital counts were estimated, we compared these to the reported number of COVID available beds pooled across all hospitals in a geographic area. This allowed us to understand when and where hospital bed need might exceed current capacity.

### 2.3.2 Description of Data

**Demographic data.** Demographic data on the age distribution for Ohio counties and ZIP Codes were obtained from the U.S. Census Bureau's 5-year American Community Survey 2014-2018 estimates [8]. The ACS data were used because the sample size is large enough for small geographies for reasonable standard errors and stable estimates. The Census Bureau does not develop data products for the USPS ZIP Codes. Rather, ZIP Code Tabulation Areas (ZCTAs) are generalized areal representations of United States Postal Service (USPS) ZIP Code service areas. We mapped the Census ZCTAs to ZIP Codes and used population estimates for ZCTAs. We used table B01001 which breaks out population counts by sex and 5 year age groups. Figure 1 shows the distribution of the high risk population (age 55+) by county and Hospital Catchment Area in Ohio (definition of Hospital Catchment Areas provided in Section 2.3.3).

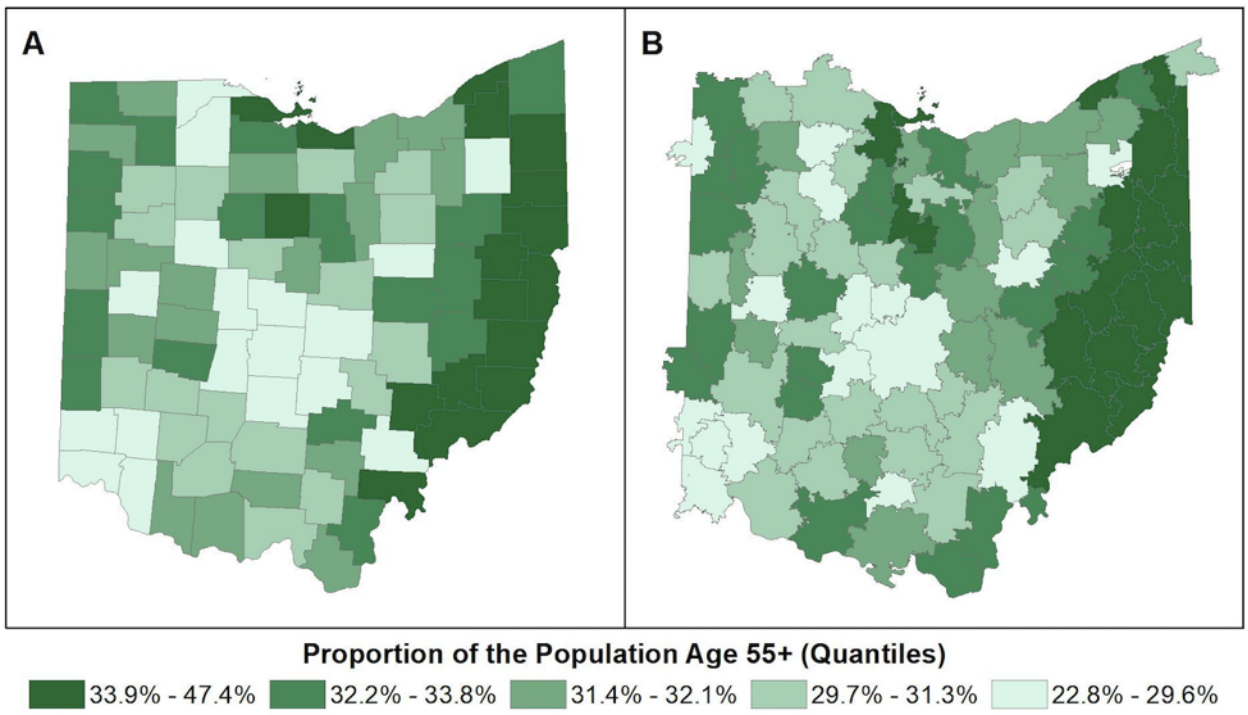


Figure 1: Proportion of the population over the age of 55 in A) counties and B) hospital catchment areas.

**OHA Hospital and Bed Data.** The Ohio Hospital Association (OHA) provided data on the location of all hospitals in the state and the number of registered beds for each facility. Beds were broken out into several categories: Airborne Isolation, Critical Care, General Medical/Surgical Beds, and Extracorporeal Membrane Oxygenation (ECMO) beds. These bed types represent two levels of care required for COVID-19 patients: general hospital care for severe disease, and ICU beds for patients requiring ventilation. The OHA data contains information that constitutes a trade secret under Ohio Revised Code Section 1333.61.

**Description of OHA hospital market share data.** OHA provided data on hospital market share derived from administrative hospital claims from the past year (January 2019-December 2019, inclusive). For each hospital, patient encounters were grouped by patient ZIP code. A hospital's market area included ZIP codes that represented the top 80% of all encounters at that hospital.

### 2.3.3 Definition of Hospital Catchment Areas

We developed small area estimates from state-level model predictions for counties and hospital catchment areas (HCA). Boundary files for the 88 Ohio counties were obtained from the U.S. Census Cartographic Boundary Files [9]. We developed hospital catchment areas using an approach modified from the Dartmouth Atlas Project [1]. We define a hospital catchment area as a collection of ZIP codes whose residents receive most of their hospital care from the hospital(s) in that region. Figure 1-B shows the HCAs developed and mapped. Note that in figure 1-B there are HCAs that cross over into neighboring states. This is explained in more detail below.

HCA definitions depend on the integrated use of geospatial methods that grouped each ZIP code in the state with the most geographically proximate hospital and modified these groupings

using data on hospital market share by ZIP code. Only hospitals with acute care beds that could be used for COVID-19 patients were included in the analysis. We excluded facilities such as long-term acute care (LTAC), hospice, orthopedic, rehabilitation or psychiatric/behavioral health hospitals and freestanding ERs. We defined HCAs using three steps:

1. The locations of all hospitals in Ohio, and those in Michigan, Indiana, West Virginia, Kentucky and Pennsylvania located on the border with Ohio, were used to generate a Voronoi diagram with hospitals as generating points. [25] We included hospital in neighboring states in the Voronoi analysis to avoid edge effects which would attribute patients to Ohio hospitals that typically use hospitals in other states. This yielded 233 distinct areas, one for each hospital generator. 28 areas were subsequently deleted because they included no area within OH.

2. Voronoi polygons were overlaid with ZIP codes. ZIP codes were assigned to the Voronoi polygon if their centroid fell within the polygon. This ensured that that each ZIP code was associated with the most geographically proximate hospital and divided the state into groupings of ZIP codes assigned to each hospital.

3. Using the OHA hospital market share file, we examined the level of agreement between each Voronoi polygon and market share ZIP codes for each hospital. In cases where adjacent Voronoi polygons were generated by hospitals which also shared 60% or more of their market share ZIP codes, we aggregated these polygons to create one hospital catchment area. In large metropolitan areas with many hospitals, we aggregated groupings of Voronoi polygons to create regional catchment areas. In rural areas, this typically resulted in aggregating two adjacent polygons.

Using this procedure, we generated 96 Hospital Catchment Areas for the state of Ohio. Some HCAs included ZIP codes from neighboring states, and some Ohio ZIP codes were included in HCAs for non-Ohio hospitals. HCAs are shown in Figure 2.

# 3 Discussion

## 3.1 Comparison with Other Approaches

Although, due to the significant differences in approach, it is difficult to directly compare our predictive model to other contemporary methods, some indirect comparisons may be offered. The susceptible-infectious-recovered (SIR) framework is the basis for many COVID-19 epidemic models [2, 12, 22, 30]. Several websites provide implementations of the SIR framework and present scenarios with different interventions such as social distancing affecting the transmission parameter $\beta$ [2, 12], thus allowing for different hypothetical scenarios to be explored. Note that the $I$ variable in the SIR model corresponds to infections, rather than symptomatic cases. Consequently, knowledge of asymptomatic to symptomatic infections is needed to translate output from these models to cases and hospitalizations. Other models such as [12] extend the basic SIR model to include additional compartments corresponding to incubation, presymptomatic infectious, stages of symptomatic infectious, and hospitalization. Parameterizing these models is a challenge, given the lack of clear clinical information and detailed data about these compartments.

The current online SIR tools [2,12] are suited for scenario exploration, like illustrating *flattening of the curve* under different levels of social distancing compliance. Unfortunately, they are less

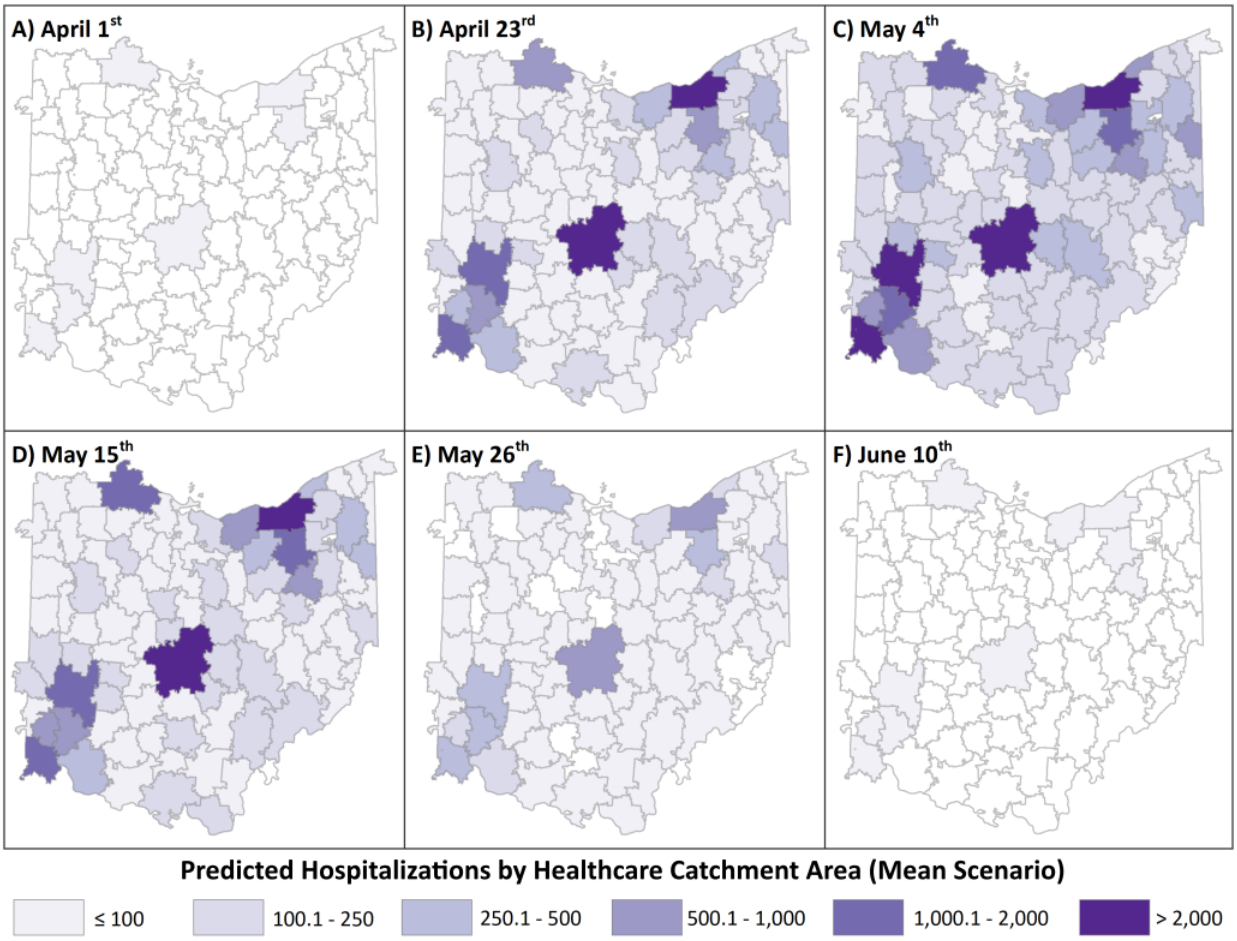


Figure 2: Estimates of hospital burden mapped to hospital catchment areas for 6 days during the epidemic. These estimates are based upon ODH case onsets through March 19, 2020.

suited for forecasts as model parameters are not calibrated based upon case or outcome data. Pei et al [22] use a metapopulation approach with SEIR dynamics in each patch and estimate model parameters (including proportion of asymptomatic infection) based upon case counts from China. This approach gives a spatially explicit model without age-structure. Conversely, [30] gives an example of a model with age structure but without space: Weitz extends the SEIR framework to include age groups, and estimates model parameters based upon hospitalization and death counts from Georgia. An alternative approach [17] also uses mortality data as model inputs, but takes a purely statistical approach in fitting a sigmoidal curve to cumulative COVID-19 deaths using a mixed-effects model. Another alternative is an agent-based model such as that used in [14]. Similar to [22], [14] forecasts a very large number of infections with COVID-19.

## Software

The software to perform model fit along with a data example is available from [21].

## Additional Materials

The video presentation describing the methodology used for epidemic size predictions is available from [26].

# 4 Appendix A - Predicting Statewide Cases of COVID-19

## 4.1 Detailed Description of Statewide Model Development

This derivation is taken in whole from the manuscript in preparation [5]. We will outline the derivation of a simple but powerful general modeling framework that provides robust estimates of the quantities relevant to monitoring local outbreaks where only limited amount of information is available through partially observed daily counts of new (symptomatic) infections - *e.g. illnesses*. Our approach is derived from the general stochastic model of a contagion spread across a contact network of nodes representing individuals in a community [19]. As a working model of a contact network we use a dynamic configuration model (CM)-type random graph (*e.g.*, [7]). Such a model is often also referred to as a *pairwise model* [18].

Briefly, we assume that each node has their degree and that the nodes may change their status from the initial "Susceptible" ($S$) to "Infective" ($I$) (or "Infectious") and, finally, to "Removed" ($R$), according to their local Markovian infectious pressure (hazard). We assume that the "Removed" individuals are no longer able to pass infection and cannot be reinfected.

## 4.2 Network Dynamics Assumptions

The dynamic model of individuals transition between the states $S$,$I$ and $R$ is based on several simple principles that allow to extract a mathematical model of the ongoing epidemic and relate it to observable data:

1. the spread occurs over a network of contacts, that is, an infectious individual may only infect his/her immediate neighbors at fixed rate $\beta > 0$; it is assumed that the average number of node's contacts is $\mu > 0$

2. the infected individual may recover at rate $\gamma > 0$ or be restricted from contacting his/her network neighbors either through mandatory or voluntary quarantine at rate $\delta > 0$

3. the infected individuals stay infected for a random amount of time according to an exponential distribution

4. the symptomatic infectives are infectious and the partial count of new infectives is observed over time with a negligible chance of misdiagnosis (false positives).

We note that the model as described here extends previous work studying epidemics on CM-random networks (e.g., [24]) and that this formulation above can also account for extensions of the basic SIR compartments to include a latent period (up to 14 days for COVID-19, see [27]), as well as more general staged progression models [16].

Instead of directly analyzing the stochastic CM model described above, which is challenging due to the heterogeneity in the number of contacts and the connectivity structure evolution (e.g., [4,20]), we will make use of the general results on the mean field approximation [3,13], and the convergence of the random infection hazard in large networks. Specifically, as shown in [19] under the assumption of the Poisson-type degree distribution (for instance the node degree is Poisson or negative-binomial

distributed), the mean field approximation of the dynamics is given by the set of equations (dots denote time derivatives)

$$\begin{aligned} \dot{x}_S &= -\beta x_D x_S \\ \dot{x}_I &= \beta x_D x_S - \gamma x_I \\ \dot{x}_D &= \beta(1-\kappa)x_D^2 + \left(\kappa\mu\beta x_S^{2\kappa-1} - \tilde{\gamma}\right) x_D \end{aligned} \tag{4.1}$$

where: the pair $(x_S, x_I)$ describes the relative number of susceptibles and infected, $x_D = x_{SI}/x_S$ is the relative density of infectious connections, $\kappa$ is the average contact network density[5], and $\tilde{\gamma} = \beta + \gamma + \delta$. The usual initial conditions are $x_S(0) = 1; \quad x_I(0) = \rho > 0; \quad x_D(0) = \mu\rho$.

### 4.3 Estimating Reproduction Rate

For the purpose of statistical analysis of the system (4.1) we make an assumption that only the empirical counts of the new infected are available in practice. Dividing last equation in (4.1) by the first one, solving for $x_D$ in terms of $x_S$ and substituting back into the first equation we obtain the reduced system with only one equation describing the decay of susceptibles. To simplify notation, denote $S_t := x_S(t)$ to obtain

$$-\dot{S}_t = \tilde{\beta}(1 - S_t^\kappa)S_t^\kappa + \frac{\tilde{\gamma}}{1-\kappa} S_t(1 - S_t^{\kappa-1}) + \tilde{\rho} S_t^\kappa \tag{4.2}$$

$$\text{where: } S_0 = 1 \text{ and } \tilde{\rho} = \beta\mu\rho, \ \tilde{\gamma} = \beta + \gamma + \delta, \text{ and } \tilde{\beta} = \mu\beta.$$

Note that the equation (4.2) is defined for $\kappa = 1$ by taking the limit $\kappa \to 1$. The value of the basic reproduction number $R_0$ for both reduced (4.2) and full (4.1) system is

$$R_0 = \kappa\tilde{\beta}/\tilde{\gamma}.$$

The condition $\kappa = 1$ implies the Poisson degree assumption for the pairwise model and reduces (4.2) to

$$-\dot{S}_t = \tilde{\beta}\left(S_t - S_t^2\right) + \tilde{\gamma}\, S_t \log\left(S_t\right) + \tilde{\rho} S_t \quad \text{and} \quad S_0 = 1. \tag{4.3}$$

Instead of thinking about $S_t$ as a proportion of susceptibles it is convenient to think about $S_t$ as an improper survival function of *the time to infection* of a single random susceptible. Then $S_t$ has an improper density $-\dot{S}_t$ (see [29]). It is improper, since $\int_0^\infty -\dot{S}_t = 1 - S_\infty = \tau < 1$ where $\tau$ is defined below (see also [6], Example 2). Under this survival function interpretation, the infection time for a randomly selected initially susceptible individual (in an infinite population) follows a temporal patterns according to the probability law $S_t$ given by (4.2) or (4.3). When we observe only a partial epidemic trajectory, say until time $T$, then the observed infection time is conditional on the infection occurring by time $T$, that is, on an event that has probability $\tau_T = 1 - S_T$. It is easy to show that as $T \to \infty$ then $\tau_T \to \tau_\infty = \tau$ the probability of a randomly selected susceptible individual *being infected during the lifetime of an epidemic*. (One may think about $\tau$ also as the final proportion of infected in the epidemic in infinite population). Then the conditional density of symptom onset is

$$f_T(t) = -\dot{S}_t/\tau_T$$

[5]The network parameter $\kappa > 0$ is defined as the ratio of network mean excess degree and mean degree, see [19] for details. It is known that $\kappa = 1$ corresponds to the Poisson degree network whereas $\kappa > 1$ corresponds to the negative binomial one.

which is simply the scaled derivative of the probability of staying susceptible past time $t$ (denoted $S_t$). Accordingly, setting $\theta = (\kappa, \beta, \gamma, \rho)$ the approximate likelihood of the joint symptom times (epidemic curve) of $n$ observed new cases by current time $T$ in an infinite population [29]) is given by

$$\mathcal{L}(\theta|t_1, \ldots, t_n, T) = \prod_{i=1}^{n} f_T(t_i) \tag{4.4}$$

Although the expression above looks simple note that the function $f_T(t)$ depends upon the vector of parameters $\theta$ only implicitly through the differential equations (4.2) or (4.3).

### 4.4 Estimating Dropout and Recovery Rates

Recall $\tilde{\gamma} = \beta + \gamma + \delta$. Given $\tilde{\gamma}$, we may estimate the recovery rate $\gamma$ from the recovery density. Then assuming $\beta$ is negligible, we have the approximate expression for drop-out

$$\delta \simeq \tilde{\gamma} - \gamma.$$

To estimate $\gamma$ we consider now the recovery density. As shown in [29], the density of daily recovery times is given by

$$g(t) = \int_0^t f_\infty(u)\gamma e^{-\gamma(t-u)}\, du.$$

For practical model fitting, a shift parameter $\varepsilon \in \mathcal{R}$ may be needed as

$$\bar{g}_\varepsilon(t) = \frac{g(t + \varepsilon \wedge 0)}{\int_0^\infty g(u + \varepsilon \wedge 0)\, du}.$$

The overall density of recovery is then the following mixture

$$\tilde{g}(t) = \frac{1}{1+\rho}\bar{g}_\varepsilon(t) + \frac{\rho}{1+\rho}\gamma e^{-\gamma t}.$$

The analogous likelihood as in (4.4) can be now produced using the conditional recovery density

$$\breve{g}_T(t) = \frac{\tilde{g}(t)}{\int_0^T \tilde{g}(t)\, dt} \quad \text{for} \quad t \in [0, T].$$

$$\mathcal{L}(\gamma|t_1, \ldots, t_k, T) = \prod_{i=1}^{n} \breve{g}_T(t_i) \tag{4.5}$$

This likelihood is used to estimate $\gamma$ directly conditional on estimated $S_t$ (and thus $f_\infty$).

### 4.5 Estimating Size of an Outbreak

We assume that the size of an outbreak $k_\infty$ is a fixed integer representing the likely number of total infections in the contact network of the confirmed cases only. Hence $k_\infty$ is not the prevalence of the disease in the population but rather just an estimate of the total outbreak size in the community of $n$ individuals where we see infections. We estimate $n$ at any given time $T$ by the discount estimator $\hat{n}_T = k_T/(1 - S_T)$ where $k_T$ is the number of cases observed by time $T$. Then we estimate the total number of cases by the end of an epidemic as

$$\hat{k}_\infty = \frac{\tau k_T}{1 - S_T}$$

where $\tau = 1 - S_\infty$ is the final probability of infection defined in Section 4.3.

### 4.6 Prediction and uncertainty quantification

We follow a semi-Bayesian approach for prediction and uncertainty quantification. We assume (independent) non-informative priors for the parameters $\tilde{\beta}, \tilde{\gamma}$, and $\tilde{\rho}$. This choice of prior distributions has the following important implication: the maximum likelihood estimates (MLEs) of $\tilde{\beta}, \tilde{\gamma}$, and $\tilde{\rho}$, obtained by numerically maximizing the likelihood function in (4.4), correspond to the mode of the posterior distribution. We then estimate the asymptotic covariance matrix of $\tilde{\beta}, \tilde{\gamma}$, and $\tilde{\rho}$ using the bootstrap method. Having estimated the mode of the posterior distribution and the asymptotic covariance, we can now perform a Laplace approximation to the posterior around the mode. This basically allows us to approximate the posterior distribution locally by a gaussian distribution, from which we can now draw posterior samples. While the cumulative epidemic curve corresponding to the MLEs obtained as a solution to (4.3) gives us the most likely trajectory, the posterior samples of $\tilde{\beta}, \tilde{\gamma}$, and $\tilde{\rho}$ are used to generate a pointwise Monte Carlo confidence interval around the most likely trajectory. Therefore, in essence, we generate predicted trajectories corresponding to the posterior samples and then compute appropriate quantiles at desired time points to get the confidence interval.

## 5 Appendix B - Estimating Hospitalizations from Statewide Predictions of Case Numbers

Let $C(t)$ be the trajectory of case onset times as generated by the statewide model. We wish to translate this to hospital census $h_i(t)$ for a geographic region $i$. For simplicity of exposition, we do not discuss separate types of hospital beds (e.g. ICU vs. non-ICU) here. Extension to separate bed-types is straight-forward.

Consider an age group $a \in \mathcal{A}$. Let $n_{i,a}$ be the number of individuals in age group $a$ in geographic region $i$, and let $n_i = \sum_{a \in \mathcal{A}} n_{i,a}$ denote the total population size of $i$. Let $N = \sum_i n_i$ denote the entire population size of Ohio.

Converting $C(t)$ to $h_i(t)$ involves the following steps:

(i) Estimating the case onsets $c_{i,a}(t)$ for age group $a$ in geographic region $i$.

(ii) Deriving from this the onsets of severe cases $s_{i,a}(t)$ that will eventually require hospitalization by age group and geographic region.

(iii) Using probability distributions for time from case onset to hospitalization and length of stay to estimate the hospital census $h_{i,a}(t)$, where $h_i(t) = \sum_{a \in \mathcal{A}} h_{i,a}(t)$.

**Estimating case onsets by age group and geographic region.** We consider the following factors in estimating $c_{i,a}(t)$ from total case onsets $C(t)$: the population size of $i$, age structure of $i$, and the relative likelihood by age of being identified as a COVID-19 case.

We first compute the total case onsets $c_i(t) := \sum_{a \in \mathcal{A}} c_{i,a}(t)$ for $i$ in proportion to the population size of $i$:

$$c_i(t) = \frac{n_i}{N} C(t). \tag{5.6}$$

To distribute these case onsets by age, define the *relative susceptibility* $s_a$ of age group $a$ as the ratio of the proportion of observed cases in $a$ relative to the proportion of the state's population that is in $a$.

A relative susceptibility of one corresponds to the proportion of observed cases across the state matching what would be expected if cases were distributed uniformly at random across individuals. Relative susceptibilities larger than one correspond to more cases identified in $a$ than would be expected at random, while relative susceptibilities less than one correspond to fewer identified cases in $a$ than would be expected at random. For the Ohio data we observe far fewer cases in youth (ages less than 21) than would be expected at random, but more cases in the elderly.

We then compute $c_{i,a}(t)$ by multiplying $c_i(t)$ by a probability incorporating the relative susceptibility of $a$ and the age structure of $i$:

$$c_{i,a}(t) = c_i(t)\frac{n_{a,i}s_a}{\sum_{\ell\in\mathcal{A}} n_{i,\ell}s_\ell}. \tag{5.7}$$

**Onset of cases that will eventually require hospitalization.** It has been documented that case outcome varies strongly with age [11]. We use the mean of the probabilities reported in [10] of severe case outcome by age group. Let $p_a$ be the probability of severe case outcome for age group $a$. Then

$$s_{i,a}(t) = p_a c_{i,a}(t). \tag{5.8}$$

Refinements include incorporating gender and other demographic covariates including comorbidities. We are actively studying these for the Ohio data and will incorporate our findings into future versions of the model.

**Hospital census over time.** The trajectories $s_{i,a}(t)$ correspond to onset times of cases that will eventually require hospitalization. To translate these into hospital census of COVID-19 patients, let $w_a(t)$ be the probability distribution for time between case onset and hospitalization, and $q_a(t)$ the probability distribution for length of stay in hospital for age group $a$. Then the convolution of $s_{i,a}(t)$ with $w_a(t)$ gives hospital admissions over time, and the convolution of admissions with length of stay gives hospital occupancy:

$$h_{i,a}(t) = (s_{i,a} * w_a * q_a)(t). \tag{5.9}$$

# References


[1] The Dartmouth Atlas of Healthcare. `https://www.dartmouthatlas.org/faq/#research-methods-faq`.

[2] COVID-19 Hospital Impact Model for Epidemics (CHIME). `https://penn-chime.phl.io/`, 2020.

[3] Frank Ball and Peter Neal. Network epidemic models with two levels of mixing. Mathematical Biosciences, 212(1):69–87, 2008.

[4] Maurice Stevenson Bartlett. Stochastic population models in ecology and epidemiology. Methuen, 1960.

[5] Caleb Deen Bastian, Wasiur R. KhudaBukhsh, Yuhan Pan, Eben Kenah, and Grzegorz A. Rempala. Predicting the size and duration of the outbreaks of COVID-19 under minimal assumptions. Technical Report, The Ohio State University College of Public Health, April 2020.

[6] Caleb Deen Bastian and Grzegorz A Rempala. Throwing stones and collecting bones: Looking for poisson-like random measures. Mathematical Methods in Applied Sciences (Early View), https://doi.org/10.1002/mma.6224, 2020.

[7] Béla Bollobás. Random Graphs. Springer, 1998.

[8] US Census Bureau. American community survey. https://www.census.gov/programs-surveys/acs.

[9] US Census Bureau. Mapping files A. https://www.census.gov/geographies/mapping-files.html.

[10] CDC COVID-19 Response Team. Preliminary estimates of the prevalence of selected underlying health conditions among patients with Coronavirus Disease 2019 — United States, February 12–March 28, 2020. Morbidity and Mortality Weekly Report, 69:382–386, 2020.

[11] CDC COVID-19 Response Team. Severe outcomes among patients with coronavirus disease 2019 (COVID-19) — United States, February 12–March 16, 2020. Morbidity and Mortality Weekly Report, 69:343–346, 2020.

[12] M. Childs, M. Kain, D. Kirk, M. Harris, J. Ritchie, L. Couper, I. Delwel, N. Nova, and E. Mordecai. Potential long-term intervention strategies for COVID-19. https://covid-measures.github.io/, 2020.

[13] Richard Durrett. Random graph dynamics. Cambridge University Press, 2007.

[14] N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, L. C. Okell, S. van Elsland, H. Thompson, R. Verity, E. Volz, H. Wang, Y. Wang, P. G. T. Walker, C. Walters, P. Winskill, C. Whittaker, C. A. Donnelly, S. Riley, and A. C. Ghani. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. https:/doi.org/10.25561/77482, 2020.

[15] G. Grasselli, A. Pesenti, and M. Cecconi. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy. Lancet, doi:10.1001/jama.2020/4031, 2020.

[16] H. W. Hethcote. The mathematics of infectious diseases. SIAM Review, 42(4):599–653, 2000.

[17] IHME COVID-19 health service utilization forecasting team and Christopher JL Murray. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv, 2020.

[18] Péter L Simon et al. István Z Kiss, Joel C Miller. Mathematics of epidemics on networks. Cham: Springer, page 598, 2017.

[19] K. A. Jacobsen, M. G. Burch, J. H. Tien, and G. A. Rempala. The large graph limit of a stochastic epidemic model on a dynamic multilayer network. Journal of Biological Dynamics, 12(1):746–788, 2018.

[20] M. Kermack and A. G. Mckendrick. Contributions to the mathematical theory of epidemics. Part I. Proc. R. Soc. A, 115:700–721, 1927.

[21] Wasiur R. KhudaBukhsh and Caleb D. Bastian. Python Code for DSA-based Modeling. https://github.com/wasiur/dynamic_survival_analysis.

[22] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (COVID-19). medRxiv, https://doi.org/10.1101/2020.02.14.20023127, 2020.

[23] M. E. J. Newman. Networks. Oxford University Press, 2nd edition, 2018.

[24] Mark EJ Newman. The structure and function of complex networks. SIAM review, 45(2):167–256, 2003.

[25] A. Okabe, B. Boots, , and K. Sugihara. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. Wiley, 2nd edition, 2000.

[26] Grzegorz A Rempala. Mathematical models of epidemics: Tracking coronavirus using dynamic survival analysis. https://mbi.osu.edu/events/seminar-grzegorz-rempala-mathematical-models-epidemics-tracking-coronavirus-using-dynamic.

[27] Aimee Taylor Rene Niehus, Pablo M De Salazar and Marc Lipsitch. Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers. medRxiv, 2020.

[28] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelli, A. Ghani, and N. M. Ferguson. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infectious Diseases, doi.org/10.1016/S1473-3099(20)30243-7, 2020.

[29] Eben Kenah Wasiur R KhudaBukhsh, Boseung Choi and Grzegorz A Rempala. Survival dynamical systems: individual-level survival analysis from population-level epidemic models. Interface Focus, 10(1):20190048, 2020.

[30] Joshua S. Weitz. COVID-19 near-term epidemic risk assessment for Georgia. https://github.com/jsweitz/covid-19-ga-summer-2020, 2020.

[31] World Health Organization. Coronavirus disease 2019 (COVID-19) situation report – 70. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200330-sitrep-70-2020.

[32] Zunyou Wu and Jennifer M McGoogan. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72,314 cases from the Chinese Center for Disease Control and Prevention. Journal of the American Medical Association, 2020.