Modeling SARS-CoV-2 infection dynamics among residential undergraduates at The Ohio State University

Matthew Wascher, Wasiur KhudaBukhsh, Patrick Schnell, Joseph Tien, Grzegorz Rempala
and the OSU / IDI COVID-19 Response Modeling Team

September 30, 2020

Abstract

This report describes efforts to model SARS-CoV-2 infection dynamics among undergraduates living on The Ohio State University's Columbus campus. The model is simple, yet flexible enough to accommodate changes in behavior over time. Model parameters are estimated using an approach that utilizes individual results of weekly SARS-CoV-2 testing of residential undergraduate students. Model output serves several purposes, including estimating the effective reproduction number ($\mathcal{R}_t$) and providing predictions of disease prevalence that can help inform decisions about isolation and quarantine bed capacity.

1 Overview

The Ohio State University has undertaken a robust testing and contact-tracing program as part of mitigation and surveillance efforts for COVID-19. Here we describe modeling efforts to assess intervention efficacy, estimate key quantities such as disease prevalence and the reproduction number over time, and provide forecasts of cases in the near term.

Some important features of the model are:

- The modeling approach is simple, with basic 'SEIR' (Susceptible-Exposed-Infectious-Removed) compartments in discrete time as the foundation.
- The modeling framework is flexible, allowing for changes in contact patterns, transmissibility, and social distancing over time.
- Our estimation procedure makes use of individual-level testing data for the on-campus population (time of last negative and first positive test).
- Statistical estimation allows for uncertainty quantification, and in particular gives credible bounds for forecasts.

Limitations of the model include:

- The model treats the on-campus population as decoupled from the off-campus population.

- Contacts are treated as well-mixed. In particular, we ignore social network structure and heterogeneity in activity patterns.

The remainder of this report is organized as follows. Section 2 describes the model formulation. Section 3 describes data sources and our process for integrating the model with data. Section 4 gives some sample results of the model, including model fit, model forecasts, and estimates of the effective reproduction number over time. Some concluding remarks are given in Section 5.

2 Model

The base of our model is a Reed-Frost type SEIR model [1]. We treat residential undergraduate students on OSU's Columbus campus as a closed, well-mixed population of known size n. The population is divided into susceptible (S), exposed (E), infectious (I), and removed (R) compartments according to immunological status. Time is treated as discrete, with units of days. We denote the counts of individuals in different compartments at time t by S_t, E_t, and I_t and assume that they evolve according to the following rules:

- Each pair of one individual from S_{t-1} and one from I_{t-1} has probability $\beta_t(n)$ of contact, and each individual in S_{t-1} who experiences such a contact is infected beginning at t,
- Following infection, an individual enters the exposed compartment and remains there for three days, with $E_t^{(j)}$ denoting the number of individuals at time t in the jth day of their exposure period,
- Each infectious individual in I_{t-1} is removed beginning at t with probability $\gamma_t(n)$.

The three day incubation period used here is comparable to but slightly shorter than the median 5.1 day incubation period reported in [5]. We assume that individuals in E are not yet infectious, nor are they detectable as infected by RT-PCR testing. By contrast, individuals in I are both infectious and detectable as infected by RT-PCR testing. While the time between exposure to detectability is not well established, it is likely shorter than five days [2, 4], hence our use of three days for the exposed period here.

Let δ_t be the daily decrease of susceptibles and ϵ_t be the daily decrease of infectious individuals. Under this rule, we have the following probability laws for the daily increments of infection $\delta_{t+1} = -\left(S_{t+1} - S_t\right)$ and recovery $\varepsilon_{t+1} = R_{t+1} - R_t$, respectively:

$$\begin{aligned} \delta_{t+1}|S_t, I_t, \beta_t(n), n &\sim \text{Binomial } \left[S_t, 1 - (1 - \beta_t(n))^{I_t}\right] \\ \varepsilon_{t+1}|I_t, \gamma_t(n), n &\sim \text{Binomial } \left[I_t, \gamma_t(n)\right]. \end{aligned} \tag{1}$$

Note that recovery here encompasses not only biological clearance of infection, but also removal from infectiousness due to isolation following positive test, as well as removal due to quarantine of infected contacts of positive cases.

An important feature of the model is that the transmission parameter β_t can potentially change at each time point. The model thus can accommodate behavioral changes over time, for example in response to perceived risk of infection, policy changes, weekend or holiday effects, and more.

Parameters for the model are the transmission rates β_t, removal of infectiousness γ_t, and initial conditions S_0, E_0, I_0. In the remainder we treat the initial conditions as fixed, and estimate β_t, γ_t from data.

Given values for S_t, β_t, γ_t, and n, we have the following expression for the effective reproduction number at time t:

$$\mathcal{R}_t = \frac{\beta_t}{\gamma_t}\frac{S_t}{n}. \tag{2}$$

3 Estimation framework

3.1 Survival and hazard functions

The estimation approach builds off of the dynamical survival analysis methods given in [3]. Specifically, we adapt the methods of [3] for the results of individual-level repeat testing.

Consider the survival function $\mathcal{S}_t$ that describes the decay of susceptibles over time, along with its associated hazard function h_t. More precisely, $\mathcal{S}_t$ is the probability that an initially susceptible individual is still susceptible at time t. Define $\beta_t(n) = \beta_t/n$ when n is assumed to be large (i.e., we have a large population of susceptibles). Define also $\gamma_t(n) = \gamma_t$. Note that by the above discussion the probability that an initially susceptible individual stays susceptible until t is

$$\mathcal{S}_t = \prod_{s=0}^{t-1}\left(1 - \frac{\beta_s}{n}\right)^{I_s} \tag{3}$$

and thus the hazard function for a random susceptible being infected in $[t, t+1]$ is

$$h_{t+1} = \frac{\mathcal{S}_t - \mathcal{S}_{t+1}}{\mathcal{S}_t} = 1 - \frac{\mathcal{S}_{t+1}}{\mathcal{S}_t} = 1 - \left(1 - \frac{\beta_t}{n}\right)^{I_t} \approx \frac{\beta_t I_t}{n}. \tag{4}$$

By a similar calculation we obtain that the hazard of recovery in the interval $[t, t+1]$ is

$$g_{t+1} = \gamma_t.$$

In view of the above we may consider a simplified approximation to (1):

$$\begin{array}{rcl} \delta_{t+1}|S_t, I_t, \beta_t, n & \sim & \text{Binomial } \left[S_t, \beta_t I_t/n\right] \\ \varepsilon_{t+1}|I_t, \gamma_t, n & \sim & \text{Binomial } \left[I_t, \gamma_t\right]. \end{array} \tag{5}$$

3.2 Testing data and time of infection

Every residential undergraduate on the Columbus campus undergoes weekly SARS-CoV-2 testing. Thus, for each individual we know

- t_{neg}, the most recent time this individual was known to be susceptible, and
- t_{pos}, the first time this individual was known to be infected.

Note that it is possible that a particular individual was infected the first time they were observed, in which case we set $t_{neg} = 0$. It is also possible that a particular individual has never been observed to be infected, in which case we set $t_{pos} = \infty$.

Given $\mathcal{S}_t$, t_{neg}, and t_{pos}, we can find the probability that an individual became infected on a particular day as follows:

- If $t_{neg} = i$ and $t_{pos} = j$, then for each $i < k \leq j$, the probability that this individual became infected on day k is
$$\frac{\mathcal{S}_{k-1} - \mathcal{S}_k}{\mathcal{S}_i - \mathcal{S}_j}.$$
- If $t_{neg} = 0$ and $t_{pos} = j$, then for each $i < k \leq j$, the probability that this individual became infected on day k is
$$\frac{\mathcal{S}_{k-1} - \mathcal{S}_k}{(1-\rho) - \mathcal{S}_j},$$
where $\rho = I_0/n$.
- If $t_{neg} = i$ and $t_{pos} = \infty$ and we have observed data until present time T, then for each $i < k \leq j$, the probability that this individual became infected before time T is
$$\mathcal{P}_T := \frac{\mathcal{S}_i - \mathcal{S}_T}{\mathcal{S}_i}.$$
- Thus, the probability this individual became infected on day $i < k \leq T$ is
$$\mathcal{P}_T \frac{\mathcal{S}_{k-1} - \mathcal{S}_k}{\mathcal{S}_i - \mathcal{S}_T}. \tag{6}$$

3.3 Parameter estimation algorithm

We use an iterative process to estimate the model parameters. Following initialization, the process uses the current prevalence estimate to compute the survival function (3). The survival function and individual interval censored testing data are then used to compute daily incidence, which is then used to update the prevalence estimate. We assume that exposed individuals remain exposed for $m = 3$ days before moving to the I compartment, and let $E_t^{(j)}$ contain individuals on their jth day of exposure. Specifically, we use the following Gibbs Sampler to estimate model parameters:

1. Initiate $\mathcal{S}_t, (\beta_t)_{t=1}^T, (\gamma_t)_{t=1}^T, I_0 = 1, E_0^{(1)} = E_0^{(2)} = E_0^{(3)} = 1, S_0 = n - 4$.
2. Given $\mathcal{S}_t$ and the data, draw the $(\delta_t)_{t=1}^T$ using the probabilities described in (6).
3. Given $(\delta_t)_{t=1}^T$ and $(\gamma_t)_{t=1}^T$, draw $(\varepsilon_t)_{t=1}^T$ and compute $(I_t)_{t=1}^T$, $(S_t)_{t=1}^T$, and $(E_t^{(j)})_{t=1}^T$ for $j = 1 \ldots 3$ using
$$\varepsilon_1 \sim \text{Bin}(I_0, \gamma_1),$$
$$\varepsilon_t \sim \text{Bin}\left(I_0 + E_0^{(3)} + \sum_{i=1}^{t-1} \left(E_i^{(3)} - \varepsilon_i\right), \gamma_t\right) \text{ when } t = 2 \ldots T.$$

$$I_t = I_0 + E_0^{(3)} + \sum_{i=1}^{t} \big(E_i^{(3)} - \varepsilon_i\big),$$

$$S_t = S_0 - \sum_{i=1}^{t} \delta_i,$$

$$E_t^{(3)} = E_{t-1}^{(2)},$$

$$E_t^{(2)} = E_{t-1}^{(1)},$$

$$E_t^{(1)} = \delta_t,$$

for each $t = 1 \ldots T$.

4. Update $(\beta_t)_{t=1}^T$ by drawing $\beta_t I_t/n \sim \text{Beta}(\delta_t + 1, S_{t-1} - \delta_t + 1)$.
5. Update $(\gamma_t)_{t=1}^T$ by drawing $\gamma_t \sim \text{Beta}(\varepsilon_t + 1, I_{t-1} - \varepsilon_t + 1)$.
6. Given $(\beta_t)_{t=1}^T$ and $(I_t)_{t=1}^T$, update $\mathcal{S}_t$ using (3).
7. Go to step 2.

Note that the updating $(\beta_t)_{t=1}^T$ and $(\gamma_t)_{t=1}^T$ here uses a Beta-Binomial conjugate prior model. For the transmission parameters β_t we use an uninformative prior Beta(1,1). For recovery parameter γ_t we use a prior of Beta(3,6), reflecting time from test result to isolation.

The estimation scheme yields posterior samples for $\mathcal{S}_t$, $(\beta_t)_{t=1}^T$, $(\gamma_t)_{t=1}^T$, $(\delta_t)_{t=1}^T$, $(\varepsilon)_{t=1}^T$, $(I_t)_{t=1}^T$, and $(S_t)_{t=1}^T$.

3.4 Testing gaps and backfill

Two challenges with the data are testing gaps and 'backfill'.

Weekly testing is conducted via sign-up slots, typically available Monday through Thursday. Little to no testing is done on Friday, Saturday, or Sunday, thus leaving gaps in the testing data. We addressed gaps by treating the day before a gap, the gap itself, and the day after the gap as a single time period and using $\mathcal{S}_t$ to estimate the number of infections that should fall in this time period, rather than each day of this time period. We then distribute the infections in the time period over the individual days uniformly. An example of such a period is 9/5-9/8. We first use $\mathcal{S}_t$ to estimate the number of infections in the four day period 9/5-9/8, then distribute them uniformly over those 4 days when updating $(\delta_t)_{t=1}^T$.

An additional challenge is the so-called backfill problem, which is a well-known challenge in fitting epidemic models. Because there is a delay between when individuals become infected and when they are observed to be infected (i.e. test positive), we have only a fraction of the information about the most recent days on which we have data. A standard solution, which we implement here, is to use data up to time T but only fit the model to some earlier time $T - s$. Here we fit the model until time $T - 4$ and use the forward prediction method outlined above to generate counts for $T - 3, \ldots T$. Note that there is an additional two to three day delay from when tests are administered to when the results are available, so that in practice there is typically a six or seven day lag between the current date and dates for which parameter estimates can be made.

4 Results

In this section we show results for model fits in terms of on-campus prevalence together with estimates for the effective reproduction number over time (Section 4.1). We also show forward predictions of the model, together with observed on-campus positivity for comparison (Section 4.2).

4.1 Model fits and effective reproduction number

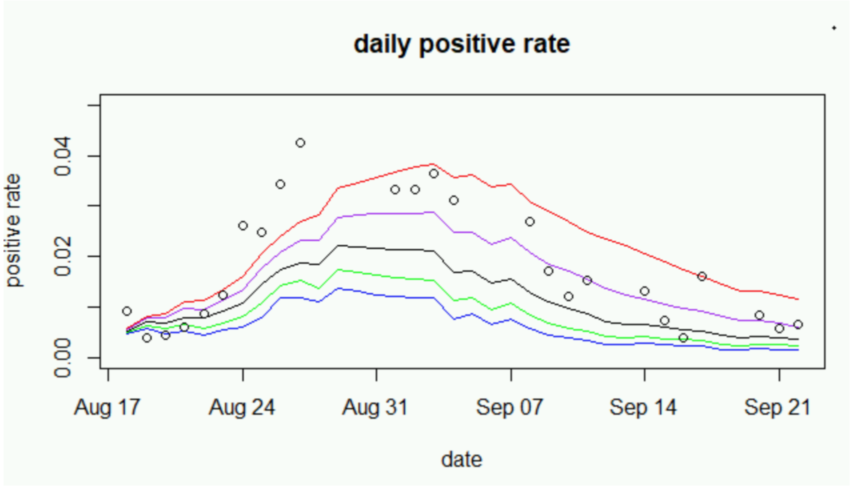

Figure 1: Model trajectories for on-campus prevalence I_t/n based upon tests administered from 8/17/2020 through 9/18/2020. Circles correspond to empirically observed positivity. Lines correspond to 10%, 25%, 50%, 75%, and 90% model quantiles.

Model trajectories for on-campus prevalence based upon fits to Vault testing data through September 18, 2020 are shown in Figure 1. Estimates for the effective reproduction number over time from tests administered from August 17 to September 25, 2020 are shown in Figure 2.

4.2 Model predictions

Because β_t and γ_t are allowed to vary each day, in order to predict forward in time, it is necessary to make some assumptions. For β_t, we average the values of $(\beta_t)_{T-6}^{T}$ for the most recent seven days for each posterior sample, and use the quantiles of the resulting distribution as assumed future values of β_t to generate the model quantiles. For γ_t, we use an informative prior based on the testing and contact tracing scheme by which potentially infected individuals are isolated. Since Vault Health has usually not tested on the weekends, we include a weekend effect that sets the prior for γ_t to Beta$(2, 14)$ on weekends.

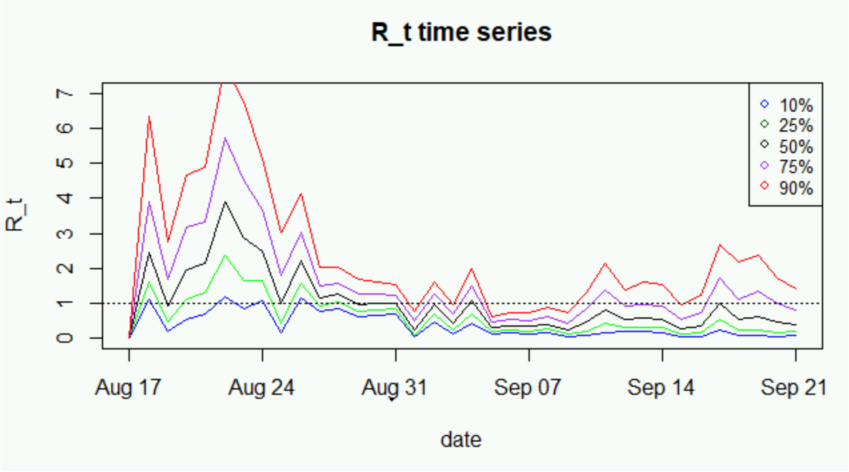

Figure 2: Estimates for the effective reproduction number $\mathcal{R}_t$ over time, from model fits using tests administered from 8/17/2020 to 9/25/2020. Lines correspond to 10%, 25%, 50%, 75%, and 90% quantiles of the posterior $\mathcal{R}_t$ distribution.

Forward predictions are obtained by running the process forward in time using the rules outlined in section 1, using as a starting point the state of the process in each posterior sample and β_t and γ_t as described above. We then take empirical quantiles of the forward time simulation of the process as the range of predicted outcomes. Forward predictions based upon tests administered from August 17 to September 18, 2020 are shown in Figure 3.

5 Conclusions

The model fits given in Section 4 appear reasonable, as is agreement between forward predictions from the model with the observed positivity on campus. We note that ability of the model to fit the testing data might be expected, as we allow the transmission β_t and recovery γ_t parameters to vary with time. However, the model does not have an excess of parameters compared with data, as the individual test results are used in the estimation. For example, one week of testing adds on the order of 12,000 data points used for parameter estimation. Additionally, over-fitting in the model would be expected to lead to poor forward predictions, whereas agreement between forward predictions and the data to this point has been good.

Despite its flexibility, there are nonetheless structural limitations to the model, including the absence of importation of infection from outside of the residential undergraduate population. Mixing with students living off-campus or the surrounding non-university community, for example, is not included in this model. Because of this, caution should be used in interpreting $\mathcal{R}_t$ estimates when prevalence is low. For example, steady low-level importation of cases could make $\mathcal{R}_t$ appear to be around one, despite their being little transmission within the residential undergraduate community.

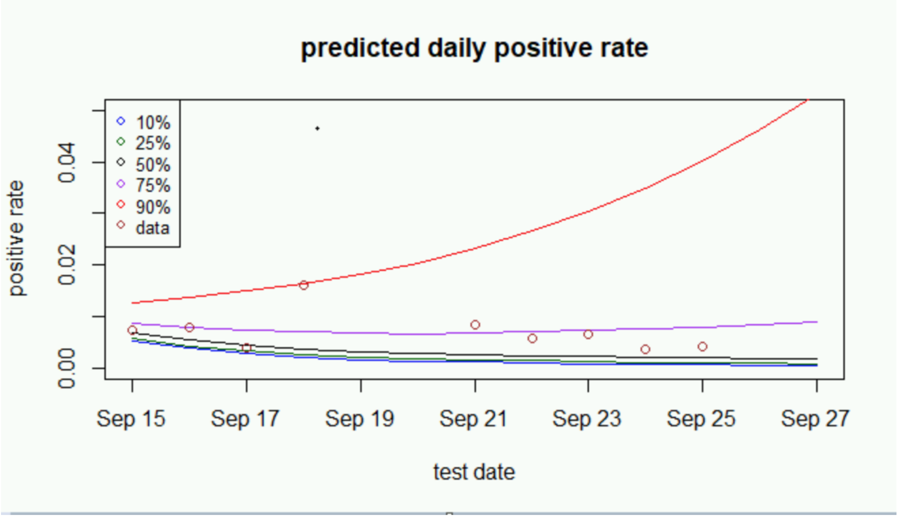

Figure 3: The predicted trajectory of the epidemic, based upon model fits from tests administered from 8/17/2020 to 9/18/2020. Different colors denote different quantiles for model output. Circles denote observed on-campus Vault positivity.

Additional work is needed examining SARS-CoV-2 dynamics in the off-campus student population, and the interaction between off-campus and on-campus students.

References

[1] Hakan Andersson and Tom Britton. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media, 2012.

[2] Xi He, Eric H. Y. Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y. Wong, Yujuan Guan, Xinghua Tan, Xiaoneng Mo, Yanqing Chen, Baolin Liao, Weilie Chen, Fengyu Hu, Qing Zhang, Mingqiu Zhong, Yanrong Wu, Lingzhai Zhao, Fuchun Zhang, Benjamin J. Cowling, Fang Li, and Gabriel M. Leung. Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature Medicine*, 26(5):672–675, 2020.

[3] W. R. KhudaBukhsh, B. Choi, E. Kenah, and G. A. Rempala. Survival dynamical systems: individual-level survival analysis from population-level epidemic models. *Interface Focus*, 10(1):20190048, 2020.

[4] D. B. Larremore, B. Wilder, E. Lester, S. Shehata, J. M. Burke, J. A. Hay, M. Tambe, M. J. Milna, and R. Parker. Test sensitivity is secondary to frequency and turnaround time for COVID-19 surveillance. *medRxiv*, `doi.org/10.1101/2020.06.22.20136309`, 2020.

[5] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incu-

bation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020.