Bayesian detection and uncertainty quantification of the first change point of the COVID-19 case curve in the Midwest: Timeliness of non-pharmaceutical interventions

Alessandro Maria Selvitella [1] Kathleen Lois Foster [2]

[1]Purdue University Fort Wayne [2]Ball State University

1. Introduction

Motivation.
The first case of COVID-19 was identified in Wuhan, China, after which it spread throughout Europe and the US, leading to an ongoing pandemic, as officially determined by WHO in March 2020. Since the very first stages of the pandemic, the global research community mobilized and started to study the evolution of COVID-19 to understand its virology, pathophysiology, and epidemiology. The complexity of the problem requires the development of new methodologies and the collaboration of large interdisciplinary teams.

Our Effort.
Our team joins this interdisciplinary research effort with the interest of understanding the dynamics of the disease from a machine learning perspective. We want to understand the time evolution of COVID-19 and in particular its changes with respect to non-pharmaceutical interventions (eg. lockdowns, social distancing, face mask, stay at home, and many others). In this poster, we will concentrate on understanding the relationship between qualitative changes in the curve of COVID-19 cases and two government policy orders: "Face Mask" and "Stay at Home".

5. Our Analysis

We ran the algorithm described in the methods with $K = 9000$ iterations and 3 chains to estimate the parameter ψ. Our outcome variable Y is taken on the log scale and represents the natural logarithm of the cumulative case counts. We will have one Y for each of the twelve states in the Midwest. We estimated the posterior distribution of the change point parameter ψ, computed its posterior mean and its corresponding 95% credible interval for each of the twelve states in the Midwest. We compared this with the dates of the first case detected in each state and the dates of the "Stay at Home" and "Face Mask" orders. We performed the Savage-Dickey density ratio test to make this comparison.

State	Illinois	Indiana	Iowa	Kansas	Michigan	Minnesota
First Case	24-01	06-03	08-03	08-03	10-03	06-03
Stay at Home	21-03	25-03	NO	30-03	24-03	28-03
Mask	01-05	27-07	16-11	03-07	27-04	24-07
First CP	28-02	07-04	29-04	11-04	01-04	27-04
LB CI	22-03	06-04	27-04	10-04	31-03	21-04
UB CI	23-04	08-04	01-05	14-04	02-04	02-05
State	Missouri	Nebraska	North Dakota	Ohio	South Dakota	Wisconsin
First Case	07-03	06-03	12-03	10-03	10-03	03-03
Stay at Home	06-04	NO	NO	24-03	NO	25-03
Mask	NO	04-05	14-11	23-07	NO	01-08
First CP	04-04	02-05	14-04	06-04	21-04	01-04
LB CP	03-04	30-04	10-04	04-04	19-04	03-04
UB CP	05-04	04-05	18-04	07-04	23-04	05-04

Table 1:This table provides the dd-mm-2020 dates for all 12 Midwest states for: First Case of COVID-19 (Row 1) , Stay at Home order (Row 2), Face Mask order (Row 3), First Change Point (CP) ψ (Row 4), Date of the Lower Bound (LB) of the 95% Credible Interval (CI) for ψ (Row 5), Date of the Upper Bound (UB) for the 95% CI for ψ (Row 6). NO indicates when an order was not executed.

2. Dataset and Software

- The case counts by state were taken from CDC, beginning with the first case in Washington reported on January 22, 2020 until February 21, 2021. The state policies, including dates and information on the "Stay at Home" and "Face Mask" orders, were taken from the COVID-19 US State Policy Database (CUSP) curated by Boston University.
- The analysis was performed using the software R and its packages *mcp* and *patchwork*.
- All data is publicly available and code is available upon request.

3. Bayesian Change Point Estimation

To estimate the change point we will use a Bayesian perspective. Although, the methodology can be adapted to multiple change points, we will concentrate on the case of one single change point. Consider a sequence of observations of an outcome variable Y (in our case the COVID-19 case counts), given by $y_1, \dots, y_T$ with $T > 0$ the time extension of our study (January, 22nd 2020 to February 21st, 2021) and $t = 1, \dots, T$ the corresponding time component. We model the mean response $\mu = E[Y]$ with a piece-wise linear function such as $\beta_1 t + \beta_2 (t - \psi)_+$, where $(t-\psi)_+ := (t-\psi)\, I\,(t > \psi)$ and $I(\cdot)$ representing the indicator function. Here β_1 is the slope at the left of the change point ψ and β_2 is the difference-in-slopes between the slopes at left and right sides of ψ.

We will estimate change points and their level of uncertainty with the mean and standard deviation of their posterior distribution via **Monte Carlo Markov Chain** methods. The priors of all parameters are uninformative, with the the exception of the prior for the change point which is restricted to be ordered monotonically while otherwise remaining uninformative.

3. Savage-Dickey Ratio Test

Suppose you observe data D and have the vector of parameters $\theta = (\theta_1, \theta_2)$ with θ_1 the parameters of interest, and θ_2 nuisance parameters. Consider a null hypothesis, $H_0 : \theta_1 = h$, with h a fixed vector of hypothesized values of θ_1. The alternative hypothesis is $H_1 : \theta_1 \neq h$. Denote p_0 and p_1 the probability density distributions under H_0 and H_1, respectively. Suppose that $\lim_{\theta_1 \to h} p_1(\theta_2|\theta_1) = p_0(\theta_2)$, then $p_1(\theta_2|\theta_1 = h) = p_0(\theta_2)$. Consider the Bayes factor

$$BF_{01} := p(D|H_0)/p(D|H_1) = p_0(D)/p_1(D).$$

Then

$$p_0(D) = \int p_0(D|\theta_2)p_0(\theta_2)d\theta_2 = \int p_1(D|\theta_2, \theta_1 = h)p_1(\theta_2|\theta_1 = h)d\theta_2 = p_1(D|\theta_1 = h),$$

which by Bayes' rule leads to

$$p_0(D) = \frac{p_1(\theta_1 = h|D)p_1(D)}{p_1(\theta_1 = h).}$$

In this way, we obtain the **Savage-Dickey density ratio**, namely the ratio between posterior and prior distributions:

$$BF_{01} = \frac{p_0(D)}{p_1(D)} = \frac{p_1(\theta_1 = h|D)}{p_1(\theta_1 = h)}.$$

In our case, we are interested in the parameter $\theta_1 = \psi$, the change point, although other parameters (eg. the two intercepts and two slopes) will be estimated as well. The observed data is $D = \{(t, y_t)\}_{t=1}^T$. Note also that the hypothesis we are interested in is actually one sided $H_0 : \psi > h_i$ with $i = 1, 2$. In particular, we want to test if the change point ψ arrives after the "Stay at Home" order h_1 or not, and if it arrives after the "Face Mask" order h_2 or not.

6. Results and Discussion

Figure 1 illustrates the results of the Bayesian Change Point Analysis with comparison to the dates of the "Stay at Home" (Red Bar) and "Face Mask" (Blue Bar) orders for each of the 12 states in the Midwest.

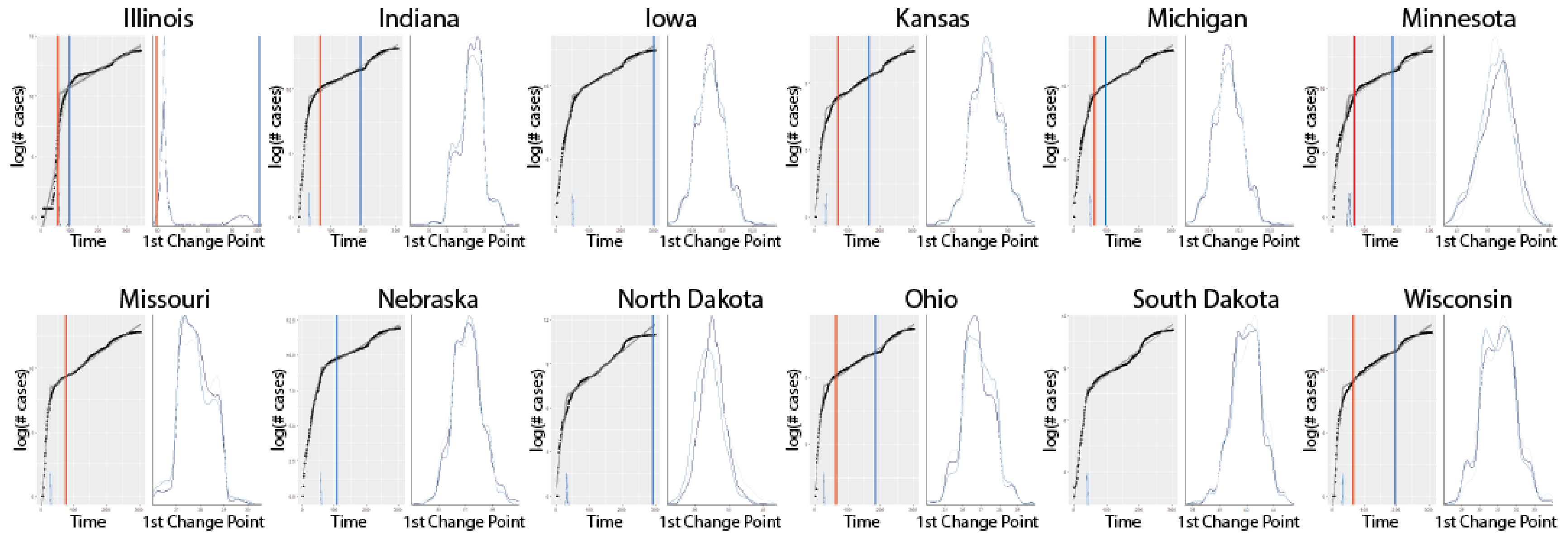

Figure 1:Left Plots: The horizontal axis represents the time variable, while the vertical represents the logarithm of the cumulative number of cases. Right Plots: Represents the posterior distribution of the first change point ψ.

- Illinois is the only state where we cannot exclude the possibility that the first change point is subsequent to the "Stay at Home" order. Note that Illinois saw the first case much earlier than the other states and registered a plateau soon after. Possibly related: Chicago is the biggest airline hub in the Midwest area by far, a fact that speculatively might be responsible for this impetus for the earlier crackdown on mask use and movement outside the home. The higher uncertainty of the estimate of the first change point in Illinois is possibly due to this plateau occurring at the beginning of the epidemic.
- The change points of Indiana, Kansas, Michigan, Minnesota, Ohio, and Wisconsin have been estimated to be before both governmental policies were put in place.
- Iowa and North Dakota did not execute a "Stay at Home" order, while the "Face Mask" order arrived much later than the estimated first change point.
- Missouri's policy recommended rather than required mask use, while its "Stay at Home" order was much later than the change point.
- Nebraska did not have a "Stay at Home" order and they mandated face mask use by employees only in public-facing businesses, and the first change point arrived before that.
- In South Dakota, there hasn't been any "Stay at Home" order, while masks were encouraged, but not required.

Altogether our results suggest that important government non-pharmaceutical interventions restricting movement outside the home and mandating the use of masks were put in place after a qualitative change in the COVID-19 case trajectory had already taken place. Thus, these government mandated policies were not a likely contributor to the observed first flattening in the curve of COVID-19 cases.

Conclusions

We studied the problem of detecting the first change point in the curve of COVID-19 cases in the twelve Midwest states. We found evidence that there has been qualitative rate changes in the diffusion of COVID-19 before the "Stay at Home" and "Face Mask" orders were implemented, in all states but Illinois. This calls for possibly quicker governmental actions. The analysis described in this manuscript is descriptive and not predictive, associative and not causal.

COVID-19 INCIDENCE IN THE INDIANA'S SECONDARY SCHOOL SYSTEM THROUGH A CONDITIONAL GAUSSIAN MODEL AND AN AGE-STRUCTURED COMPARTMENTAL MODEL

Kale Menchhofer [1], Nathan Mills [1], Kathleen Lois Foster [2], Alessandro Maria Selvitella [1]

[1]Purdue University Fort Wayne [2]Ball State University

Motivation:

- COVID-19 emerged in Wuhan (China) at the end of 2019 and was declared a pandemic by the World Health Organization in March 2020.
- With more than 2.5 million deaths worldwide as of late February 2021, COVID-19 has been a defining health crisis and has impacted people's everyday lives in countless ways.
- One of the most noteworthy circumstances of the COVID-19 outbreak in the United States was the closure of virtually all schools throughout the country.
- Since their closure, one of the most pressing issues pertaining to COVID-19 is how to properly reopen schools without sparking a surge in cases throughout the community.
- Currently, the situation is highly heterogeneous with even nearby schools adopting alternative strategies.
- The prolonged school closure has been shown to negatively affect student learning experience and to be the cause of serious mental illnesses, such as anxiety and depression.

Dataset and Software

- Data is taken from the Indiana Data Hub, updated to Dec. 28th, 2020. This dataset includes COVID-19 student cases broken down by school.
- The analysis and the simulations utilized the software *R* and the package *deSolve*.

Our Analysis

We will concentrate on a couple of distinct models with the intent of capturing important factors in the diffusion of the coronavirus in Indiana's secondary school system. For the sake of interpretability, we confined our analysis to the simplest models capturing the phenomenon under study.

Conditional Gaussian Model.
In the first model, we analyze the number of cases in each school, subdividing them by county. The distribution of the number of cases in schools within a given county is modeled with a **Conditional Gaussian Distribution**; namely, we model the number of cases in each county as a linear function of the sum of the student cases in that county plus a Gaussian error.

Age Structured Compartmental Model.
The second model is a compartmental model with age structure (4 compartments of young interacting with 4 compartments of adults). **Compartmental models** are models in which the population is divided into mutually exclusive and exhaustive classes, and the spread is modeled through a system of coupled ODEs describing the evolution of the disease across compartments.

Conditional Gaussian Model

We considered the number of student cases y_i in Indiana's county i and the sum number of cases per secondary school x_i in county i for $i = 1, \dots, 92$, with 92 the number of counties in Indiana. Our model is a simple linear regression model of the form $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with the x_i considered non-stochastic, $E[\epsilon_i] = 0$, $\epsilon_i \sim N(0, \sigma^2)$ and ϵ_i independent and identically distributed for $i = 1, \dots, 92$. Although our analysis was comprehensive of 1) mean and sum for students/teachers/employees/all of them [8 models], 2) Conditional Gaussian/Poisson/Negative Binomial for each model with Outlier detection at 1-2-3 st. dev., 3) Cooks distance for all models, and 4) Non- parametric outlier detection tests for all models, for space reasons, we report in this poster only the result on the relationship between the sum of student cases per county.

Age-structured SEIR model

We considered the following SEIR model with two age-groups: children vs adults. We have the following system of coupled differential equations:

$$(SEIR2) \quad \begin{cases} \frac{dS_1}{dt} = -S_1\left(\beta_{11}I_1 + \beta_{21}I_2\right) \\ \frac{dE_1}{dt} = S_1\left(\beta_{11}I_1 + \beta_{21}I_2\right) - \sigma_1 E_1 \\ \frac{dI_1}{dt} = \sigma_1 E_1 - \gamma_1 I_1 \\ \frac{dR_1}{dt} = \gamma_1 I_1 \end{cases} \qquad \begin{cases} \frac{dS_2}{dt} = -S_2\left(\beta_{12}I_1 + \beta_{22}I_2\right) \\ \frac{dE_2}{dt} = S_2\left(\beta_{12}I_1 + \beta_{22}I_2\right) - \sigma_2 E_2 \\ \frac{dI_2}{dt} = \sigma_2 E_2 - \gamma_2 I_2 \\ \frac{dR_1}{dt} = \gamma_2 I_2 \end{cases} \tag{1}$$

Here $S_i(t), E_i(t), I_i(t), R_i(t) \in C^1([0, +\infty))$. To fix the ideas: $i = 1$ represents the children age group and $i = 2$ the adult age group with $S_i(t), E_i(t), I_i(t), R_i(t)$ the corresponding susceptible, exposed, infective, and removed individuals of age group i. The following theorem implies that $SEIR2$ gives biologically meaningful solutions for all times t.

Theorem

For every $0 \leq S_{i0}, E_{i0}, I_{i0}, R_{i0} \leq 1$ $i = 1, 2$ such that $S_{i0} + E_{i0} + I_{i0} + R_{i0} = 1$ for $i = 1, 2$, there exists a unique solution to system $(SEIR2)$ such that $I_i(0) = I_{i0}, E_i(0) = E_{i0}, I_i(0) = I_{i0}, R_i(0) = R_{i0}$, $0 \leq S_i(t), E_i(t), I_i(t), R_i(t) \leq 1$ for $i = 1, 2$, and $S_i(t) + E_i(t) + I_i(t) + R_i(t) = 1$ for $i = 1, 2$.

Sketch of the Proof By Picard–Lindelöf existence and uniqueness theorem, there is a unique smooth solution local in time. Summing the equations in each system, we deduce that the population is conserved. Since the total population is conserved $S_i(t) + E_i(t) + I_i(t) + R_i(t) = S_{i0} + E_{i0} + I_{i0} + R_{i0} = 1$. Therefore, the solution is global in time. By its equation, S_1 is decreasing. By taking the ratio between $\frac{dS_1(t)}{dt}$ and $\frac{dR_1(t)}{dt}$ and integrating from 0 to $+\infty$, we get by conservation of total population:

$$\frac{S_{1\infty}}{S_{10}} = e^{\left\{-\left[\frac{\beta_{11}}{\gamma_1}R_{1\infty} + \frac{\beta_{21}}{\gamma_2}R_{2\infty}\right]\right\}} \text{ and so } S_{1\infty} \geq S_{10}e^{\left\{-\left[\frac{\beta_{11}}{\gamma_1}R_{1\infty} + \frac{\beta_{21}}{\gamma_2}R_{2\infty}\right]\right\}} > 0.$$

This applies similarly for the second age-group and analogously for the other compartments.

Simulations

In our simulations, we will use the population values for Indiana and the epidemiological parameters in Table 1.

State	Description	Range/Estimate	Base Case
β_{11}	child-to-child	[0.05-2]	0.1
β_{12}	child-to-adult	[0.05-2]	0.5
β_{21}	adult-to-child	[0.05-2]	0.5
β_{22}	adult-to-adult	[0.05-2]	0.5
$1/\sigma_1$	child latent	3	3
$1/\sigma_2$	adult latent	3	3
$1/\gamma_1$	child infectious	4	4
$1/\gamma_2$	adult infections	4	4

Table 1:This table provides the parameter values for our 17 simulations. The β's are the transmission coefficients, whose range are given per day. The latent and infectious periods are in days.

As an example, we report the simulations for Allen County, which is characterized by the following parameters: children population (≤ 17) $n_1 = 97,101$, adult population $n_2 = 282,198$ (> 17), and initial conditions for the eight compartments: $S_{10} = 97,099/n_1, E_{10} = 2/n_1, I_{10} = 0, R_{10} = 0, S_{20} = 282,195/n_2, E_{20} = 2/n_2, I_{20} = 1/n_2, R_{20} = 0$.

Results

Conditional Gaussian Model. Interestingly, the number of cases per county is roughly 30 times the sum of the student cases in each county. This value is stable across counties. The estimate of the slope coefficient $\hat{\beta}_1 = 29.694$ gives significance of the predictor with p-value $< 2 * 10^{-16}$.

Age-structured model. The most interesting models had the smallest or the greatest proportion of each age group having contracted COVID-19. Extreme cases 1, 3, and 4 depicted optimal outcomes in which less than 5% of either age group have contracted the disease by the end of a 90 days period (Figure 1). The most calamitous outcomes, which were exhibited in extreme cases 6, 8, 9, 11, 12, 13, 14, 15, and 16, showed that more than 99% of both age groups contract the disease within 90 days. Cases 12, 15, and 16 showcased the worst potential scenarios with over 99% of both age groups being exposed to or contracting COVID-19 before day 20.

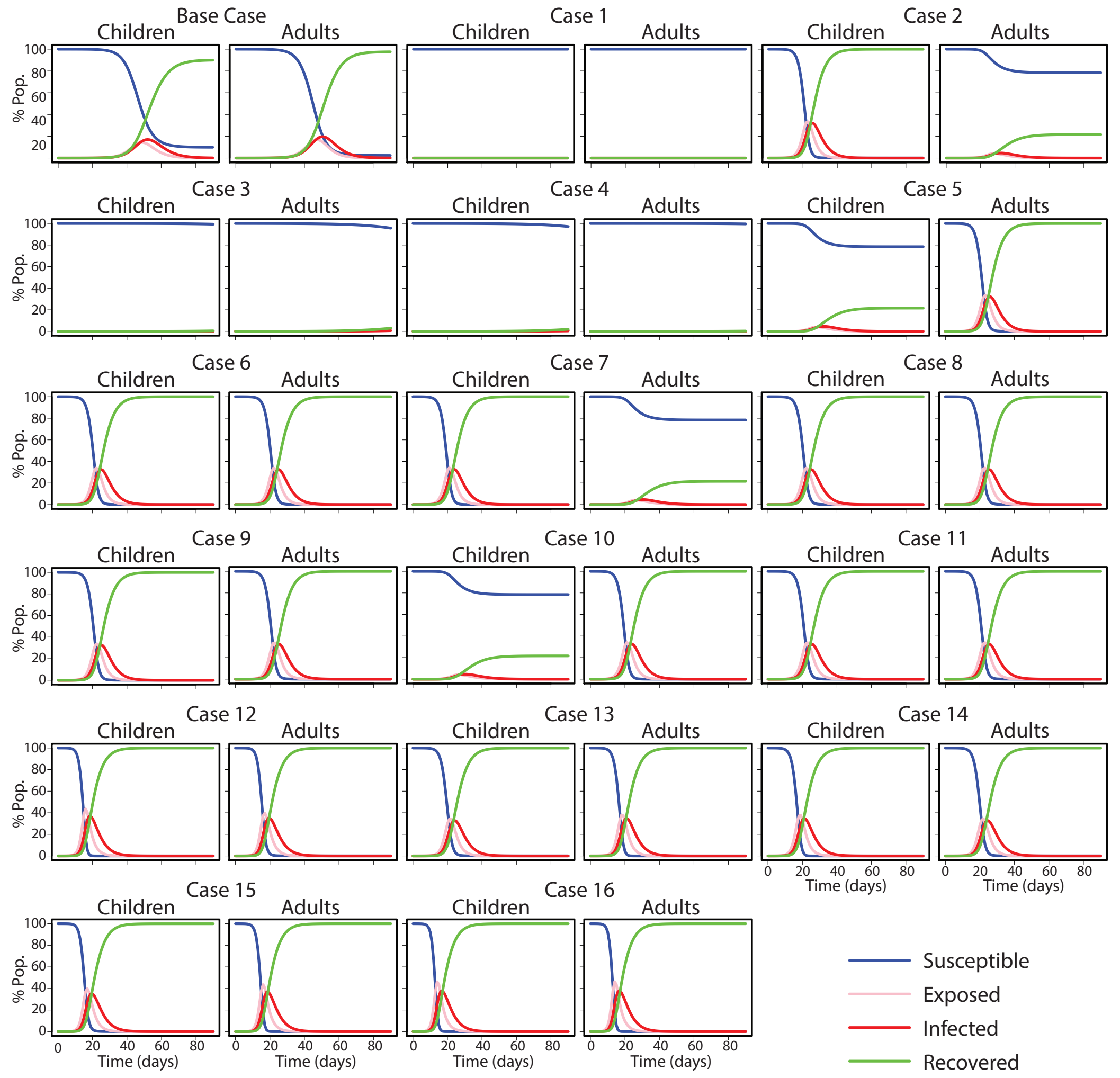

Figure 1:Trajectories of the 8-compartment SEIR models using the parameters of our simulations.

Discussion and Conclusions

Conditional Gaussian Model. The conditional sum of the student cases per county scales linearly with the number of cases of the county. This has speculatively important public policy related consequences, including the possibility of concentrating the testing in schools and using the scaling factor to estimate the incidence of COVID-19 in the full population.

Age Structured Model. The simulations of our models with parameters in line with those of Indiana showed that even if adults keep their contact with other adults to a minimum, transmission from young can present itself to be extremely detrimental to the more at-risk population. This shows that optimal school reopening strategies can potentially benefit not only the school population, but the entire community.

Overall Message. Taken in conjunction, these results underline once more the importance of adopting proper school reopening strategies and how they relate to the diffusion of the coronavirus outside the school environment. The diffusion of the coronavirus among the school population has the potential to not only be a strong determinant of the health of the more at risk population, such as elderly and sick, but also be a proxy for the incidence of COVID-19 in the community.