Teaching undergraduate physical chemistry lab with kinetic analysis of COVID-19 in the United States

Dylan K. Smith, Kristin Lauro, Dymond Kelly, Joel Fish, Emma Lintelman, David McEwen, Corrin Smith, Max Stecz, Tharushi D. Ambagaspitiya, Jixin Chen*

Department of Chemistry and Biochemistry, Ohio University, Athens, OH 45701

**ABSTRACT**
A physical chemistry lab for undergraduate students described in this report is about applying kinetic models to analyze the spread of COVID-19 in the United States and obtain the reproduction numbers. The susceptible-infectious-recovery (SIR) model and the SIR-vaccinated (SIRV) model are explained to the students and are used to analyze the COVID-19 spread data from U.S. Centers for Disease Control and Prevention (CDC). The basic reproduction number $R_0$ and the real-time reproduction number $R_t$ of COVID-19 are extracted by fitting the data with the models, which explains the spreading kinetics and provides a prediction of the spreading trend in a given state. The procedure outlined here shows the differences between the SIR model and the SIRV model. The SIRV model considers the effect of vaccination which helps explain the later stages of the ongoing pandemic. The predictive power of the models is also shown giving the students some certainty in the predictions they made for the following months.
INTRODUCTION

The unprecedented times we live in have brought about a call for chemical professionals to provide solutions for instructing laboratory techniques in our rapidly changing world. The onset of the novel coronavirus or SARS-CoV-2 in 2019 (COVID-19) changed the way chemistry is taught and brought about many hurdles to conquer. In the past two years we have seen a shift from traditional learning spaces to remote solutions allowing students to be instructed in their studies without risking contact during the pandemic.\textsuperscript{1–4} This is difficult as an important part of chemical education is physically taking part in chemical experimentation.\textsuperscript{5} Brilliant and dedicated educators developed and employed new procedure to provide that critical part of a well-rounded education in chemical science, for example, Silverberg published an article on an approach to laboratory instruction that allows great flexibility for students who may be in or out of quarantine.\textsuperscript{6} It is this goal that needs meet and, it needs to be kept in mind that not every laboratory or classroom has equal access to equipment nor does every student have equal access to a stable internet connection. This switch from in person instruction to remote across the world has made that fact front and center to programs with a diverse range of students in regards to their race, ethnicity, and socio-economic conditions.\textsuperscript{1,2,7,8} The laboratory procedure outlined here was built on a well understood model in epidemiology, the susceptible-infectious-recovered (SIR) model.\textsuperscript{9–11} Students only need access to a computer capable of running Microsoft Excel, Microsoft Word, and that has access to the world wide web. This procedure seeks to be very accessible with nearly any chemistry laboratory being able to conduct it with their students.

Modeling techniques have been in use in the field of epidemiology for centuries now.\textsuperscript{11–15} With one of the earliest examples found in its use to track the life expectancy in patients who were inoculated for smallpox.\textsuperscript{16} The foundation is then there to take the raw data derived from public health authorities and create models to understand and predict the direction of an ongoing health crisis.\textsuperscript{17} One such model is the SIR model that has its roots in a simple relationship between three possible states of being susceptible, infectious,
The transitions between these three states can be treated in the same manner of chemical kinetics where a concentration of the states exists that change in time depending on various factors. Two of those factors that will be focused heavily on in the procedure will be replacement number $R_t/R_e$ and the basic replacement number $R_0$. The SIR model correlates the replacement number with the relative spreading speed of the virus. By adding additional parameters such as vaccination rates, SIR model can be extended to simulate the real-world data with a reasonable replacement number, especially at key points of public policy or social events.

We have reported a laboratory procedure last year to analyze COVID-19 spreading in a state at the early stage with no vaccinations. This year, we added the vaccination data into the analysis and updated the SIR model to SIR-vaccinated (SIRV) model to analyze and predict the trends of COVID-spreading in the U.S. The skills taught and utilized in this procedure are useful in the physical chemistry classroom and many other places in the educational journey of an undergraduate student. The ability for students to see in real time the impact of public policy, and the statistical parameters they will derive from the model leaves a lasting impression about the usefulness and limitations of such techniques.

**PROTOCOL**

This laboratory procedure can be completed in two weeks (4 three-hour labs). The students are instructed to collect literature research, process data, write reports, and receive additional instruction outside of class time if needed. The SIR model and a modified version SIRV model which considers vaccination rates are introduced and practiced. Pre-laboratory instruction about the SIR model along with information about the history of mathematical modeling in the study of pandemics is necessary to explain the practical usefulness to the students. Microsoft Excel is the only computational program required to conduct this laboratory. While it could be conducted just the same with other programs that allow for data modeling, some students at the undergraduate level find themselves more familiar with Excel than other programs. The loan mortgage calculations are practiced in Excel at the beginning to warm up the students about model simulations in Excel, which has been explained in details in our previous publication. Students will also need to source and download data of COVID case numbers and vaccination rates of the region or regions they are designated to. This could be a state in the United States, it could be just the city of New York, or it could be an entire nation. The only real requirement is that the region being examined has reliable and easily accessed data on COVID case rates per day and vaccination rates. We instructed students to use a state in the United States as their region of interest for this procedure. Both COVID-19 cases and vaccination data in the U.S. are originated from U.S. Centers for Disease Control and Prevention (CDC). The COVID cases are summarized and downloaded from Github site of the New York Times. The vaccination data is downloaded from Our World in Data’s State by State data on Covid-19. Both are accessed and downloaded around Jan. 27, 2022 with no further modifications except for sorting over states and time. The pretreatment of the data includes running smooth for a 7-day period from three days before to three days after a given day and filling the gaps of data by copying the available ones before them. Five days are added before the first date and assume one case during these five days. Data is checked and download again on April 14, 2022 after the analysis has been done to check the predictions.
The SIR model follows a flow quite like the kinds of simple diagrams you may see drawn out in a modern chemistry textbook showing you the simplest model of chemical reaction kinetics (Figure 1). In the model (Figure 1), given a set of data with $S$, $I$, $R$ in the unit number of people in a community, the two rate constants $\beta/N$ and $\gamma$ are the key parameters to predict the spreading trend of the disease. $\beta$ refers to the average infection frequency of any given carrier and $\gamma$ the frequency of removing an infected carrier. Both will have units of days$^{-1}$ as the data provided by public and private sources on the spread of the virus will be most often in days. Students will be working with $N$ refers to the total population of their assigned region.

Historically, the two rate constants in the SIR model are combined to a new parameter called the reproduction number. Three different forms of reproduction number are examined in this lab, namely the time-dependent reproduction number $R_t$, the time-dependent effective reproduction number $R_e$, and the basic reproduction number, $R_0$ (Equations 1-2). The time-dependent reproduction number, $R_t$, reflects the infectivity of the virus and the social interaction frequency of the society and is defined:

$$R_t = \frac{\beta(t)}{\gamma(t)}$$

which is proportional to the ratio between the second-order reaction rate constant of infecting and the recovering rate. For a given variant of the virus, this value should be mainly proportional to the frequency of social interactions and virus transportation rate which are affected by various conditions such as public awareness and government regulations on social frequency, social distance, air circulation, washing hands, and wearing masks.

An effective reproduction number, $R_e$, has been defined in the literature to reflect the up-bending and down-bending of the daily new cases, which is the quasi-first order reaction rate over the recovering rate:

$$R_e = \frac{\beta(t)S(t)}{N(t)} \frac{1}{\gamma(t)}$$

In Equation 2, both the growth and decrease part are first order dependent on the number of infectious people, thus, $R_e>1$ reflects an exponential increase in the daily cases, $R_e=1$ a steady state, and $R_e<1$ sees an exponential decay of the number of infectious people, corresponding to the upside and the downside of the waves of infected cases each day.

The basic reproduction rate, $R_0$, is the initial reproduction number which is the same as the $R_t$ and $R_e$ at the beginning when the susceptible population is assumed to be the same as the total population and no action is yet taken to mitigate the spreading.
The rate equations of the SIR model are listed here again just for ease of read. The dropping rate of the susceptible population is a second-order reaction proportional to the product of the number of susceptible people \( S \) and the number of infectious people \( I \):

\[
\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{3}
\]

The changing rate of the infectious population is a tug of war between the newly infected population and the recovered population each day. The latter being a rate constant of recovering, \( \gamma \), times the total number of infectious people. The value \( \gamma \) is the one over the number of days before each infectious person is being removed from the infectious category, we used a fixed average value 0.2 Days\(^{-1} \) according to the literature for simplicity,\(^{25,30} \) even though it should slightly vary over different population groups and time.\(^{18} \)

\[
\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \tag{4}
\]

\[
\frac{dR}{dt} = \gamma I \tag{5}
\]

**SIRV Model:**

The SIRV model is nearly identical to the SIR model but includes the effect of vaccination on reducing the number of susceptible people (Figure 2). The vaccination rate is time-dependent for COVID-19 as vaccines become more available and number of people who are willing to take the vaccine varies over time. **Equations (6-10)** show the relationship between the variables in the SIRV model.

In this model, the dropping rate of the susceptible population is a combination of the infected rate with the vaccinated rate. The vaccination rate is the increase of the fully vaccinated population assumed immune over the total population.

\[
r_V = \frac{dV_{full}}{Ndt} \tag{6}
\]

Thus, assuming the vaccination is evenly distributed across the total population (if we had data on how the vaccination is distributed among different groups, a model can be built for more accurate rates),

\[
\frac{dS}{dt} = -\frac{\beta IS}{N} - r_V S \tag{7}
\]
\[
\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I
\]  
(8)

\[
\frac{dR}{dt} = \gamma I
\]  
(9)

The immune population is assumed to be the recovered people, recovered and vaccinated people, or the susceptible then fully vaccinated people by ignoring the exceptions and people who are naturally immune,

\[
\frac{dl_{mm}}{dt} = r_vS + \gamma I
\]  
(10)

All the data analysis in the lab is carried out using discrete numerical methods instead of the continuous analytical model described above. These equations are converted to discrete equations with the time interval \(dt \Rightarrow \Delta t = t_2 - t_1 \equiv 1\) day, where the subscription 1 and 2 meaning two adjacent days, e.g. 2 is today and 1 is yesterday. Examples are provided in the following for the SIRV model. Equations for the SIR model use the same principle.

\[
r_{v1} = \frac{V_2 - V_1}{N\Delta t}
\]  
(11)

\[
S_2 = S_1 - \left(\frac{\beta_1I_1S_1}{N} + r_{v1}S_1\right)\Delta t
\]  
(12)

\[
I_2 = I_1 + \left(\frac{\beta_1I_1S_1}{N} - \gamma_1I_1\right)\Delta t
\]  
(13)

\[
l_{mm2} = l_{mm1} + (\gamma_1I_1 + r_{v1}S_1)\Delta t
\]  
(14)

The model predicted daily new cases number is

\[
n_{ewI_2} = \frac{\beta_1I_1S_1}{N}\Delta t
\]  
(15)

The students start the numerical calculation by guessing the first date of the first infectious person in a state, e.g., five days before the first date of report. The student can practice manual fitting by changing \(R_t\) values piecewise as explained previously by comparing the simulated new cases per day with the raw data.\(^{25}\) They can also directly calculate the real-time \(R_t\) and \(R_e\) values by estimating \(\beta\) from Equation 15 by replacing the simulated new case with the smoothed raw data of daily case, \(n_{ewI_{raw}}\):

\[
R_{t2} = \frac{n_{ewI_{raw2}} \cdot N}{S_1I_1\gamma_1}
\]  
(16)

\(\Delta t\) is held 1 day and \(\gamma\) is held 1/5 day\(^{-1}\) throughout the fitting,\(^{25}\) with it taking 5 to 7 days for an infected individual to become symptomatic according to the CDC and assuming people stay quarantine once they are aware of being infected.\(^{4,30,31}\) From \(R_{t2}\), one can calculate \(\beta\) and \(R_e\)

\[
\beta_2 = \gamma_2R_{t2}
\]  
(17)

\[
R_{e2} = \frac{\beta_2S_2}{NY_2}
\]  
(18)

These values can be used to calculate \(S, I, l_{mm}\) and daily new cases (Equation 11-15) that day. Please see SI Excel sheets for the detailed calculations on both models.
RESULTS
Covid Data Acquisition:

![Graph showing daily cases and smoothed data.](image)

Fig. 3. (a) Example of daily cases produced by subtracting a day’s accumulated positive cases from the previous number, and overlaid with seven days smooth for the state of Ohio from 3/4/2020 to 1/25/2022. (b) Number of accumulated fully vaccinated people in the state of Ohio till 1/25/2022.

Each student was assigned a state and was instructed to use a source whose data can be verified by the CDC. Daily cases can be found by subtracting that day’s positive case count by the day before giving roughly the change in $S$ with the assumption that everyone counted in the $N$ value will be susceptible and ignore infectious people who were not tested. A smoothed daily case count can then be generated with a moving average of the daily change in cases with 3 days before and after (Figure 3).

Data Analysis:

Once students have collected data for their chosen state, they were instructed to find the base reproduction factor $R_o$ by manually fit the SIR model to data from the 1st two weeks. This was done by taking the calculated cases per day generated from the SIR model and fitting it to the cases per day from their smoothed data when minimizing the residual.

$$Residual = I_{raw} - I_{calc}$$  \hspace{1cm} (19)

Typical $R_o$ value of a U.S. state is found to be at $\sim 3^{,25,30}$

Then $R_t$ is manually adjusted in a short period of choice to minimize the residual or directly calculated from the daily cases in a state. An example is shown with the data in the state of Ohio (Figure 4) assuming no vaccination is taken for the SIR model. Both the $R_t$ and $R_o$ values are found to be oscillating around 1 along the waves of outbreak.
When considering vaccination, assuming immune after fully vaccinated, and refit the same data with the SIRV model, the \( R_t \) value gradually bends up after March 2021 and reaches somewhat close to the original basic reproduction value \( R_0 \approx 3 \) and the \( R_e \) values remains similar to those in the SIR model (Figure 5).

The \( R_t \) value of several randomly chosen states are listed in Fig. 6 including Ohio shown above. The real-time fitting at the beginning of the outbreak is very noisy. The states investigated show a similar trend. Some large noises in summer 2021 are from missing data points and tuning backward of the positive cases in some states.
Fig. 6. Comparison of real-time $R_t$ value obtained from SIRV model fitting of several randomly chosen states.

Students were instructed to continue their calculated daily cases well past 1/25/2022 that are not available when they did the analysis using the average reproduction numbers for the last 5 days of available data to predict the trend for the following months. The models overlap well with the data obtained later giving the students some predictive power (Figure 7).

Fig. 7. Prediction of the trend in Ohio after 1/25/2022 with different models overlaid with the data obtained 4/14/2022 from U.S. CDC (red dashed line). An $R_t$ value averaged from 1/20-1/25 is used for all days after 1/25.

**DISCUSSION**

The basic reproduction number of COVID is better to be extracted from the piecewise method of the first 2 weeks of data due to the relatively high sensitivity to the noise at the beginning when case numbers are small. 2 weeks of data give pretty consistent results from state to state with $R_0 \sim 3$. The piecewise method can be applied to the full data set but is time-consuming. Thus, with proper smoothing, the real-time method is better to pull out the $R_t$ values over time for the middle and later stages.

$R_t$ is directly related to the second-order reaction rate constants which reflect the collision frequency of individuals and the energy barriers of the disease infection. $R_e$ is related to the quasi-first-order reaction...
rate constant that has been normalized to the concentration of the susceptible population in the community. The $R_0$ value as designed directly correlated with the growth and decline of the daily cases and its magnitude away from 1 reflects the exponential grow or decline rate like the interest rate or payback rate in a mortgage. However, it does not reflect the effect of social regulations as well as $R_t$, especially in the later stage of the spread. In the later stage, the concentration of the susceptible population has been significantly reduced due to immunity either from recovery or vaccination. Thus, $R_t$ is a more valuable value than $R_e$ to trace the pandemic during these spreading stages. The resulting $R_t$ from the SIRV model (Figure 6) is consistent with our experience that travel and social activities are almost fully recovered around 2022 new year which brings the $R_t$ value back to near $R_0$ in all the randomly chosen states. Both SIR (Figure 4) and SIRV model (Figure 5) give the same $R_e$ values but SIR model significantly lower estimates the $R_t$ values when vaccination is significant.

Breakthrough, natural immune, partial immune, variants of the virus such as delta and omicron, social density, transportations, cultural differences, variations of $\gamma$ values, and population are all ignored during the analysis to simplify the model. However, they can be added and create more reaction pathways with modifications to the calculations, which is beyond the scope of this data analysis lab course. With information on what number of daily cases represent separate variants in the total number of daily new cases and the number of reinfections, The model can be expanded to account for the impact of reported higher rates of reinfection by the Omicron variant.\textsuperscript{30,32}

The model provides the students some prediction power on the spread trend of COVID-19 (Fig. 7). For example, if a mask can block 50% of the virus from spreading, then we can reduce the $R_t$ value by 50% if everybody has put on the mask; reducing 50% of social activity further halves the $R_t$ value as seen in the first several months of the pandemic (Figure 4-5 piecewise $R_t$). In general, the $R_t$ value is unlikely to exceed the $R_0$ value and must be larger or equal to zero for the same virus which set the limits on the upper and lower boundary of the spreading rate. The effect of the vaccination is readily visible in the simulation on the spread rate assuming no change in the social behaviors. There is no obvious evidence observed for the different spreading rates among the different variants such as delta and omicron, which will need data on the variants among the positive cases to be analyzed.

**CONCLUSION**

Extending the kinetic model from SIR to SIRV helps the students to understand the later stages of disease prevention and experience kinetic model selection. Comparing the $R_t$ and $R_e$ values help the students to distinguish the second-order reaction and the quasi-first order reaction models for the same reaction mechanism, which echoes what they have learned in the class well. Comparison between the mortgage model, the COVID-19 kinetics, and the conventional chemical reaction kinetics allow the students to compare the units in different systems, especially on the rate constants, thus helping them to understand the purpose of such analysis and the prediction task better. The real-time $R_t$ values correlate with the social events the best thus provides more prediction abilities, but their value is dependent on various conditions such as models, assumptions, data availability, and data accuracy. Overall, applying what they have learned in chemistry classes to a real-world problem motivates the students to learn and practice various skills.
ASSOCIATED CONTENT
Supporting Information
The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.XXXXXX.
Example excel sheets (XLSX)

AUTHOR INFORMATION
Corresponding Author
*E-mail: chenj@ohio.edu
ORCID: 0000-0001-7381-0918

ACKNOWLEDGMENTS
We thank the National Human Genome Research Institute of the National Institutes of Health (award number R15HG009972) for supporting undergraduate teaching. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We also thank Ohio University Nanoscale & Quantum Phenomena Institute (NQPI).

REFERENCES
(5) Avargil, S.; Bruce, M. R. M.; Amar, F. G.; Bruce, A. E. Students’ Understanding of Analogy after a CORE (Chemical Observations, Representations, Experimentation) Learning Cycle, General Chemistry Experiment. Journal of Chemical Education 2015, 92 (10), 1626–1638. https://doi.org/10.1021/acs.jchemed.5b00230/ASSET/IMAGES/ACS.JCHEMED.5B00230.SOCIAL.JPEG_V03.


2021, 6(43), 29223–29232. https://doi.org/10.1021/ACSOMEGA.1C04842/SUPPL_FILE/A01C04842_SI_002.XLSX.


