

3-15-2021

Epidemiology and the SIR Model: Historical Context to Modern Applications

Francesca Bernardi
Worcester Polytechnic Institute

Manuchehr Aminian
California State Polytechnic University, Pomona

Follow this and additional works at: <https://scholarship.claremont.edu/codee>



Part of the [Mathematics Commons](#), and the [Science and Mathematics Education Commons](#)

Recommended Citation

Bernardi, Francesca and Aminian, Manuchehr (2021) "Epidemiology and the SIR Model: Historical Context to Modern Applications," *CODEE Journal*: Vol. 14, Article 4.

Available at: <https://scholarship.claremont.edu/codee/vol14/iss1/4>

This Article is brought to you for free and open access by the Journals at Claremont at Scholarship @ Claremont. It has been accepted for inclusion in CODEE Journal by an authorized editor of Scholarship @ Claremont. For more information, please contact scholarship@cuc.claremont.edu.

Epidemiology and the SIR Model: Historical Context to Modern Applications

Francesca Bernardi
Worcester Polytechnic Institute

Manuchehr Aminian
California State Polytechnic University, Pomona

Keywords: SIR model, plague, Ebola, epidemic, India, West Africa
Manuscript received on May 17, 2020; published on March 15, 2021.

Abstract: We suggest the use of historical documents and primary sources, as well as data and articles from recent events, to teach students about mathematical epidemiology. We propose a project suitable — in different versions — as part of a class syllabus, as an undergraduate research project, and as an extra credit assignment. Throughout this project, students explore mathematical, historical, and sociological aspects of the SIR model and approach data analysis and interpretation. Based on their work, students form opinions on public health decisions and related consequences. Feedback from students has been encouraging.

We begin our project by having students read excerpts of documents from the early 1900s discussing the Indian plague epidemic. We then guide students through the derivation of the SIR model by analyzing the seminal 1927 Kermack and McKendrick paper, which is based on data from the Indian epidemiological event they have studied. After understanding the historical importance of the SIR model, we consider its modern applications focusing on the Ebola outbreak of 2014-2016 in West Africa. Students fit SIR models to available compiled data sets. The subtleties in the data provide opportunities for students to consider the data and SIR model assumptions critically. Additionally, social attitudes of the outbreak are explored; in particular, local attitudes towards government health recommendations.

1 Introduction

It is increasingly evident that the younger generations of students are actively involved in pushing for justice on their college campuses [10, 5, 8]. Students, the general public, and even fellow mathematicians are often skeptical that STEMM (Science, Technology, Engineering, Mathematics, and Medicine) can be a tool for social good. We believe it is part of our mission as faculty members to correct this misconception and guide students in

understanding the role that ethics, equity, and social justice have in mathematics research and education.

Recently, the COVID-19 pandemic has awakened a sudden, growing interest from students in epidemiology and disease modeling from a mathematical point of view as well as from public health and sociological perspectives. Countries' responses to this global crisis have differed widely due to varying access to resources, trust and value given to scientific expertise, and societal norms. Consideration of the local cultural and historical contexts as well as the lessons learned from previous epidemics has been crucial to planning local and country-wide approaches to this recent international threat. For example, differences in culture and traditions centering the collective good over the individual (or vice-versa) had a great impact on policy decisions and approaches to contain the spread of COVID-19 [13, 14].

Students should be trained to consider that epidemiological models were indeed developed to aid containment of disease spreading and to plan public health responses to epidemics. These models have real effects on actual populations, often in real time. While the disease itself will have common characteristics that span countries, regions, ethnicities, religions, and more, certain characteristics affecting epidemics development are geographically focused and subject to local values and access to resources. In this paper, we describe a project for undergraduate students designed to teach them about the Susceptible-Infected-Recovered (SIR) model in an historical and social context, and then let them explore its application to the 2014-2016 Ebola epidemic in West Africa.

In this project, students learn first about the Indian Plague epidemic of the early 1900s through historical primary sources [3]. Then, they are guided through the derivation of the SIR model by analyzing the seminal 1927 Kermack and McKendrick paper [9], which utilizes data from the Indian Plague epidemic itself (Section 2). Students are then asked to consider modern applications of this model by focusing on the Ebola outbreak of 2014-2016 in West Africa. Fitting the SIR model to real, available, compiled datasets [12] leads students to consider the subtleties of working with real data and confronting model assumptions critically. Additionally, students learn about local attitudes towards government health recommendations and how those affected the spread of the Ebola epidemic [4] (Section 3). Finally, we discuss possible implementations and variations of this project and conclude by reflecting on potential improvements and future directions (Section 4).

An Appendix follows the Reference section, giving some of the materials we prepared for students.

2 Historical Context: The Indian Plague Epidemic of the Early 1900s and the SIR Model

Our goal is to guide students in understanding the actual applicability of the SIR model, starting from its inception. The 1927 Kermack and McKendrick paper [9] utilizes data from the Indian Plague epidemic of the early 1900s, so students start off the project by learning about this event through historical public health records. Early in the epidemic, the British Empire instituted the so-called Indian Plague Commission to study the spread

of plague in India, understand the causes of the diseases, and help stop the epidemic. We select several key excerpts of the Commission's extensive report (freely available to the public [3]) for students to read and discuss. We provide them with questions (see Appendix A for our *Student Guide*) and supplementary materials from the Centers for Disease Control [6, 7] to reflect on the Commission's epidemiological observations while highlighting the intrinsic imbalance of power between the British colonialists authoring the report, and the Indian population being observed [2]. We help students in building mathematical intuition regarding what key aspects of the epidemic could be modeled and should be considered and what others could be omitted.

Reading a historical primary source will most likely be a new experience for students outside of the context of history or literature courses, especially for those enrolled in STEMM majors. Having them reflect on the Commission's report and related documents achieves four goals:

- Reading primary sources reporting the real, drastic effects of these events on the local population will open the students' eyes to the potential impact of mathematics and epidemic modeling. Given their own recent experiences with the COVID-19 pandemic, this is probably less needed now than it would have been just last year.
- Having students apply critical thinking in a context traditionally associated with a history class, rather than a math class, may provide an extra 'buy-in' to the project for participants. We want them to be challenged in this project in ways they probably haven't been before in a mathematics context.
- Allowing students to read about this epidemiological event without the burden of having to connect it to new mathematical concepts allows them to be braver when it comes to stipulating model assumptions and hypotheses. Students are often subject to anxiety as well as to self-imposed and external pressures when confronting new mathematical challenges [11]; initially separating the two aspects of this project aids them in gaining confidence in themselves.
- The supplementary materials provide guiding questions for students (Appendix A), a brief overview of the disease in question [6, 7], and a way to appreciate the intrinsic biases in the historical writings [2]. It is important for students not to take the readings at face value but rather understand who the authors and the subjects are and how their positioning in the epidemic and colonial contexts affects their public health choices and outcomes. We will return to similar ideas later when working on the 2014-2016 Ebola epidemic in West Africa.

Continuing with the historical part of the project, students navigate through the original 1927 Kermack and McKendrick paper, following prompts and focusing on selected sections of the manuscript (for details and example questions, see Appendix A). As they read through the chosen parts of [9], they are asked to answer questions in writing and follow along some of the mathematical derivations. There are several parts of the manuscript where a knowledge of single-variable calculus is enough to work through the equations. After digesting the Introduction and General Theory of [9], students jump

ahead to one of the Special Cases described in the paper, i.e., *Part B. Constant Rates*. This section includes a figure with data from the Indian plague epidemic, so students are asked to analyze the plot and compare it to what they know from reading the primary source material. This is also where they first see the full SIR system (albeit in 1927 notation). Then, students spend time thinking through more technical aspects of the problem, such as choice of variables and units, the meaning of each term and each equation, and the system end-behaviors. From here, they work towards the modern notation of the SIR model and learn about the basic reproduction number along the way. Finally, they are asked to solve analytically the SIR system for this special case and analyze its behavior as it compares to their earlier qualitative predictions.

3 Modern Application: The 2014-2016 Ebola Epidemic in West Africa

We believe that incorporating real data into mathematical models is an essential part of an applied mathematics curriculum in the modern day, and this project provides an excellent opportunity to do so. The core mathematics of what students are asked to do here is applying a few approaches to identifying initial conditions and/or SIR model parameters. However, this part of the project goes well beyond these mathematical tasks. Students get experience in loading and preprocessing data, posing questions and ‘subsetting’ the data, analyzing their results critically, and finally making careful conclusions and considering further analyses.

The mathematics of ‘fitting’ parameters in an ODE model becomes exponentially more difficult as the number of parameters increases, the model dependencies become nonlinear, or both. Depending on the preparation of students, the instructor may consider beginning with a simplification for the early period of the epidemic, where an approximate solution form and simple linear least squares may be used. If the students are more advanced, or a longer term project is intended, the instructor may consider having the students apply a nonlinear least squares solver to work with the full SIR model; this can be an alternative as well as an addition to the linear fit previously discussed. Below we report a possible first approach to fitting the SIR model to data from the 2014-2016 Ebola epidemic in West Africa [12], starting with a simpler SI approximation. Python codes used for fitting and plotting are available in [1].

3.1 The ‘SI’ Approximation

The purpose of this first possible step is to obtain an approximate solution for the ‘Infectious’ population, $I(t)$, which is simple enough to allow students the use of ordinary least squares to estimate β , the disease contact rate. To achieve this, a few assumptions need to be made, and students are asked to think them through. Removing the ‘Recovered’ category from the typical SIR essentially results in the model only having a single transfer

between compartments:

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \quad (3.1)$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} \quad (3.2)$$

This is the first simplification for students to grapple with as the missing category can be interpreted as either being included in the ‘Infectious’ group, or being assumed to be so small as to be insignificant (an infectious person has an expected number of days until they recover). In any case, applying the conserved quantity $S + I = N$, then solving for the ‘Susceptible’ population and substituting, gives an equation depending only on the Infectious population and two parameters, β and N . Students enrolled in an ODE course (and even some who have only taken calculus) will have likely encountered this as the well-known logistic equation:

$$\frac{dI}{dt} = \beta \frac{I}{N} (N - I), \quad I(0) = I_0. \quad (3.3)$$

As a side exercise, this can be solved using partial fraction decomposition; one solution form is:

$$I(t) = \frac{N e^{\beta t}}{e^{\beta t} - 1 + N/I_0}. \quad (3.4)$$

The last step of the derivation here leads students towards the assumptions that, first, I_0 is small relative to N (that is, the epidemic begins with a small number of infected people in the population), and second, we are interested in fitting the contact rate β during the initial phases of the infection (this meshes well with being able to ignore the ‘Recovered’ compartment for being zero, or close to it). Then, one can make an asymptotic approximation considering $e^{\beta t} - 1$ to be very small relative to N/I_0 , so that this term can be crossed off. Hence, the expression reduces to what is often understood as the ‘exponential growth’ phase of the SIR model:

$$I(t) \sim I_0 e^{\beta t}, \quad \text{with } t \text{ small.} \quad (3.5)$$

If the instructor or students are time constrained, an alternative approach can be to start at this approximation and argue its reasonableness from a modeling perspective. Recall that the advantage of following this process is to obtain an explicit formula for $I(t)$, so that β can be fit to data. When estimating β from available time series data $(t_k, \log I(t_k))$, students should apply a log-transformation: $\log(I) \sim \log(I_0) + \beta t$, and use a linear least squares fit to find a value of β . Here, there is the option of simplifying this process even further by treating I_0 as a known value; alternatively, I_0 can be viewed as an unknown to be inferred from the data. We expect students to grapple with a few modeling concerns at some point during this process:

1. Defining initial conditions. What does $t = 0$ refer to here? Similarly, what should I_0 be? One option is to start time at the first observed case, and use initial condition $I_0 = 1$ (1 person), but these choices can cause fitting difficulties down the road. **The**

important thing for students to realize here is that there is no perfect answer. Trying to make an appropriate choice for the time frame of reference and I_0 is a modeling challenge of its own, involving difficult mathematics especially when working with real, noisy data.

2. Short-time approximations. What does “ t small” mean? What does “initial phases of infection” mean? A possible answer from a risky interaction-type model considers this period lasting as long as the chance of two infectious people interacting with each other is very small. Once again, the question of ‘smallness’ is difficult to address, but it can lead to interesting conversations regarding nondimensionalization and epidemiological context.
3. Total population. How do we decide what the total population N should be? Each choice of N , whether *ad hoc* or informed by the data (e.g., the population of a neighborhood, city, or country) has critical modeling assumptions built in and comes with implications for data fitting. Depending on the focus and scope of the project, the instructor may guide students to make a suitable choice (e.g., the city’s population) and leave other options as potential directions to explore for a project extension.

3.2 Loading and Processing Data

We suggest students work with the Ebola data sets compiled in the Github repository [12] which includes data from the Ministries of Health of several West African countries and data sets from the World Health Organization, among other things. This data includes very fine-grained information which is useful for study by epidemiologists, and to the despair of mathematicians. There are two general routes when working with data in the `case_products` folder of [12] :

1. Work with the Excel file `case_data_consolidated_sl_and_liberia.xlsx`. In Python, we recommend using the `pandas`¹ package to load this file and access its sheets of data. The sheet `Sierra_Leone_transposed`, for example, has daily Ebola case data grouped for various regions throughout the country, and allows students the opportunity to further narrow their focus, or aggregate across the country.
2. The file `country_timeseries.json` is a so-called JSON file, which stores data in a flexible format allowing for potentially heterogeneous, messy data. One may use the `json` package in Python to load it quickly, then take further steps to process it enough to plot and analyze it. Here, data is stored per-day, with daily cases and deaths aggregated by country. This allows students to get started as quickly as possible in a data fitting exercise. The examples presented in the next section utilize this version of the data.

¹`pandas` is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

3.3 Some Example Results

Depending on students’ interest and focus, the data analysis can follow many different paths. We discuss a few nuances working with the JSON per-country, per-day case data so that instructor and students have a guidepost in doing their own work. We have made the associated data and code available in our public GitHub repository [1].

The data were loaded and arranged in a pandas DataFrame in Python, with each row being a calendar date. We have transcribed a small sample of this data in Table 1, which was the result of the following processing. Each row has an integer time (days

Date	Day	Cases_Liberia	Cases_SierraLeone	Cases_Guinea
6/1/2014	71	13	79	328
6/3/2014	73	13		344
6/5/2014	75	13	81	
6/10/2014	80	13	89	351
6/16/2014	86	33		398
6/17/2014	87		97	

Table 1: A small sample of Ebola cases by country contained in the JSON data file. Missing values in this Table represent the data source not having a report from a country on a given day. Nigeria and Senegal did not have any cases to report until July 23, which is why we do not include them here.

since a reference), and death and case counts in five different countries spanning about six months. This data is stored entirely as strings, initially, so our loading script casts values to datetime objects (for simplified plotting of dates on the horizontal axis) and to integers for case counts and time for fitting. Imputing empty strings with NaN (“not a number”) allows `matplotlib`, the plotting software, to bypass missing data in a natural way.

Initial examination of the data reveals several facts. First, the behavior is vastly different from country to country and a difference in time of initial outbreak can be observed. More interestingly, countries with significant outbreaks – Guinea, Liberia, and Sierra Leone – show very different growth rates during this period. Given this, we felt further aggregation across some or all of the countries lost too much information.

With cleaned data, we applied the approximations and data fitting methodology described in the previous section. For the purposes of modeling and fitting, we chose to restrict our focus to Sierra Leone and Liberia (Guinea would also be a reasonable choice, but we do not explore it here). We define $t = 0$ as June 1, 2014. Where data was missing, we ignored the corresponding $(t, I(t))$ pair by utilizing a mask to restrict focus to time points for each country where data was available. Finally, we obtained fits of the model parameters for the Sierra Leone and Liberia case data as reported in Table 2.

	β	$\log I_0$	I_0 (predicted cases on June 1)	Doubling time (days)
Liberia	0.0494	2.8820	18	14
Sierra Leone	0.0278	4.5892	98	25

Table 2: Example fit of the model parameters for Liberia and Sierra Leone.

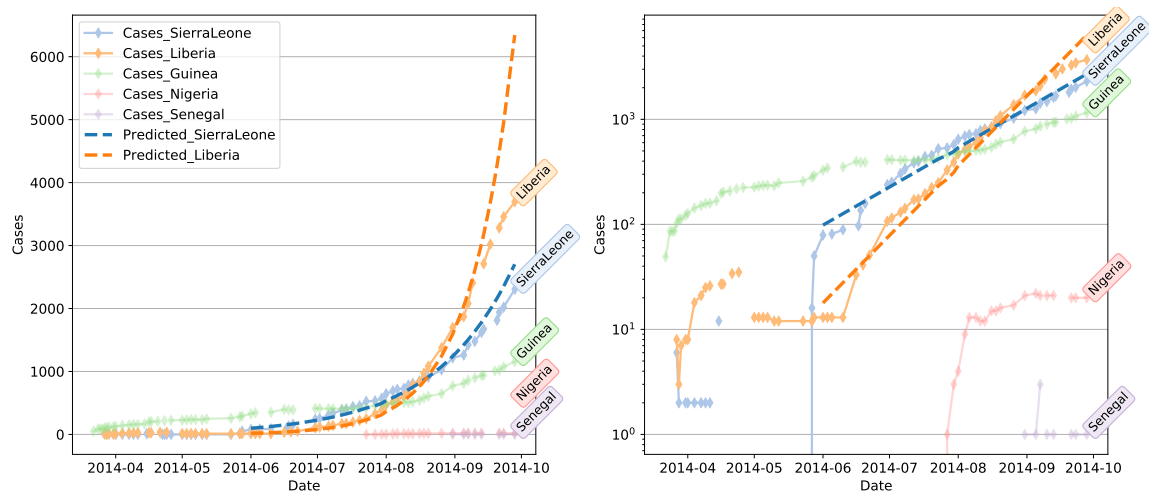


Figure 1: Example of data visualization for Ebola cases count during April-October 2014 in several countries in West Africa. Predicted Ebola cases count for Sierra Leone (shades of blue) and Liberia (shades of orange) starting on June 1, 2014. Left and right figures illustrate the same data with a linear scale (left) and logarithmic scale (right) for the cases. Code to produce this plot is available our own public GitHub repository [1].

We confirm reasonableness of our fits by including the predicted model on top of the actual data in Figure 1. We have observed that this last step is often challenging for students. While obtaining parameters is ultimately a sequence of five commands in the script once the data is cleaned, correctly applying those parameters in a model forces students to tackle a few challenges. To name a few: they should revisit how these parameters appear in the model, they should consider how the choice of $t = 0$ associates to the model built, and they should understand the relationship between calendar dates and ‘time’ used in the model. If guided past these challenges, students get a deep satisfaction when they see their model curve overlapping well with observed data in a way that merely obtaining a numerical value for a parameter cannot show.

An interesting feature to note in Figure 1 is that the data curves for Sierra Leone and Liberia intersect between August and September 2014. In Sierra Leone there were about a factor of 10 more cases in June 2014, but Liberia had a much shorter doubling time in their cases (which we showed in Table 2), so it passed Sierra Leone in the number of cases in about two months. We do not know the reasons for this; there could have been spatial differences in how the disease was spreading (e.g., higher density areas versus rural areas), or public health policy differences, or a combination of causes. There is an excellent opportunity here to dig deeper, for example within the context of an undergraduate research project.

3.4 Putting the Modern Ebola Epidemic in Context

In conjunction with studying this data, as with historical documents on the Indian Plague, we have encouraged students to read related materials to expand their thinking beyond the mathematical exercise. As an example, one such article from the *New York Times*

provides a cultural lens on the Ebola epidemic [4]. Public health efforts coming from primarily Western aid organizations (including the World Health Organization), clashed with local customs in Liberia during the 2014-2016 epidemic, especially in regard to the burning of corpses of those afflicted with Ebola. Local people working in crematories were ostracized by their families and communities for going against Liberian tradition. When reading these documents, students learn about other cultures and are faced with how the complex realities of epidemiological modeling, data collection, and analysis influence public health decisions and policy that potentially affect the lives of millions.

4 Possible Implementations and Future Work

This project has been implemented at multiple US institutions in different versions as part of the syllabus and as an extra credit assignment for an introductory ODE course, as well as an independent study project. Most students who worked on this project belonged to STEMM and social sciences majors, but not mathematics majors. A first version of this project was ideated for an introductory modeling course in 2014, and various versions of the project have been used almost every year since. Most recently, in the 2019/2020 academic year, two undergraduate students at Florida State University (majoring in Biological Sciences and Economics and Statistics, respectively) worked on a year-long version of this project as part of the Undergraduate Research Opportunity Program (UROP) on campus.

When asked to think back on their experience working through this project, all students reported enjoying reading primary sources and using them as a basis for mathematical exploration. This was particularly true for students majoring in the life sciences. Students were surprised to realize how many factors need to be considered to design appropriate mathematical modeling to fit the behavior of real-world populations and cultures. Students appreciated the chance to consider the importance of context as well as mathematical modeling in making policy decisions and evaluating whether a chosen approach is working as hoped. Unsurprisingly, participants found reading the Kermack and McKendrick paper [9] most challenging. They often got lost in the details of the manuscript and were not able to follow the steps of the calculation reported in the paper, even when the mathematics involved should have been accessible to them. The challenges of reading a technical mathematics paper (especially one from 1927) became apparent very quickly; this is the part of the project where we as instructors had to step in more consistently during the implementations. Some students liked the data analysis part of the project more than others. In this final part of the project, students were faced with subtleties in the data collection and analysis and with having to confront the reality of messy and incomplete data sets. This helped them realize the stark difference between the pre-arranged class exercises they are used to and the realities of modeling with actual data.

While part of the reason for developing this project is to get students outside of mathematics excited about differential equations and modeling, it would be interesting to see if the more technical aspects of the 1927 Kermack and McKendrick paper could be appreciated and explored further by participants majoring in mathematics. Nonetheless, analyzing the selected parts of the manuscript is enough for students to understand where

the model comes from and do some mathematical experimentation on their own. The solid historical connections between the Indian Plague epidemic of the early 1900s and the seminal SIR paper of 1927 make them an ideal place to start this exploration, but given the wealth of data freely available online, this project can be adapted to include virtually endless other modern applications. Even within our focus of the 2014-2016 Ebola epidemic in West Africa, there are numerous avenues that we leave unexplored. We mentioned one such example at the end of Section 3.3. Additionally, first and second-hand accounts of the difficulties reported on the ground when dealing with health officials handling the epidemics could be further researched. From the data analysis and parameter fitting perspective, estimating multiple parameters in addition to β involves complex, nonlinear data fitting. The identifiability of the recovery rate γ primarily relates to the medium and long-time dynamics of SIR and can only be considered when including the ‘Recovered’ compartment; asymptotic analysis could be applied to compute long-time approximate solutions in a similar fashion to the short-time analysis done for the infection rate β .

Students participating in this project are exposed to clear examples of how mathematics, and STEM more broadly, can be tools in service of public health and social problems. We believe students would be more interested in working towards technical STEM degrees if made aware of the many ways they can use them to serve society. Young students, and in particular those from underrepresented groups in STEM, find strength in helping others and advocating for social justice. We advocate for teaching the younger generations how to use mathematics ethically to serve their broader goals. We believe the approach showcased in this paper incorporating historical and social context can be adopted for a variety of projects focused on ODE modeling. While admittedly not all differential equation models have as rich a history and widespread a use as the SIR, ODEs are so often used to model the real world that they are an ideal avenue for this type of project. We hope our work can be viewed as a guiding example of how to inject some historical and social context in a mathematics classroom.

References

- [1] Manuchehr Aminian. Supplementary materials for a lesson plan for a math epidemiology project. URL https://github.com/maminian/codee_ebola.
- [2] David Arnold. *Colonizing the body: State medicine and epidemic disease in nineteenth-century India*. University of California Press, 1993.
- [3] Indian Sanitary Commissioner. *The Etiology and Epidemiology of Plague – A Summary of the Work of the Plague Commission*. Superintendent of Government Printing, 1908. URL <https://babel.hathitrust.org/cgi/pt?id=uc1.b5626368>.
- [4] H. Cooper. They helped erase Ebola in Liberia. Now Liberia is erasing them. *New York Times*, 2015. URL <http://www.nytimes.com/2015/12/10/world/africa/they-helped-erase-ebola-in-liberia-now-liberia-is-erasing-them.html>.

- [5] Nature Editorial. Teenage activist and an IPCC triumph. *Nature*, 572:283, 2019. URL <https://doi.org/10.1038/d41586-019-02425-7>.
- [6] Centers for Disease Control and Prevention. Plague. URL <https://www.cdc.gov/plague/index.html>.
- [7] Centers for Disease Control and Prevention. Protect yourself from Plague. URL https://www.cdc.gov/plague/resources/235098_Plaguefactsheet_508.pdf.
- [8] Wesley C. Hogan. *On the Freedom Side: How Five Decades of Youth Activists Have Remixed American History*. UNC Press Books, 2019.
- [9] William O. Kermack and Anderson G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927.
- [10] Dawn Laguens. Planned parenthood and the next generation of feminist activists. *Feminist Studies*, 39(1):187–191, 2013.
- [11] María Isabel Núñez-Peña, Macarena Suárez-Pellicioni, and Roser Bono. Effects of math anxiety on student success in higher education. *International Journal of Educational Research*, 58:36–43, 2013.
- [12] Caitlin Rivers. Data for the 2014 global Ebola outbreak. URL <https://github.com/cmrrivers/ebola>.
- [13] Robert Simon. Is there a trade-off between freedom and safety? A philosophical contribution to a COVID-19 related discussion. *Philosophy*, 10(7):445–453, 2020.
- [14] Jay J. Van Bavel, Katherine Baicker, et al. Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, 4:460–471, 2020.

A Appendix. The SIR Model in Historical Context: Student Guide

The following is a packet of exercises and readings most recently tailored by Francesca Bernardi to use as the introduction of a year-long Undergraduate Research Opportunity Program (UROP) project on the SIR model at Florida State University in the academic year 2019/2020. It is an expanded version of a shorter document developed by Manuchehr Aminian in 2014. We provide this as an appendix for a more concrete picture of the types of questions that have been asked of students working on this project. Not all readings and options discussed in the main manuscript appear in the appendix.

There is a range of difficulty in questions. Some sections of this packet assume knowledge of basic ordinary differential equations (separable equations and the method of integrating factors, in particular). Students will have an easier time with this document if they have some understanding of how ODEs are manipulated and solved. However, some sections can be used with students without any prior differential equations experience. These typically involve little to no calculation, but rather mostly working with the readings and understanding the model and its modern applications conceptually. These have been used successfully in an introductory mathematical modeling course geared towards students majoring in fields other than mathematics.

For both math and non-math majors, it may be useful to provide a sheet of definitions of technical terms to help them in readings. For an example of this, see section [A.4](#) of the Appendix.

This project revolves around mathematical epidemiology.² This document is a starting point for our project; it is a guided exploration of the SIR model for disease spreading. It consists of reading assignments, trying a few mathematical challenges on your own, and keeping track of what you're learning by answering the questions listed to form a written report.

A.1 The Indian Plague Epidemic of the Early 1900s

In the early 1900s, an outbreak of plague erupted in India which was then part of the British Empire.³ Take a look at the *Wikipedia* article on the *Third Plague Pandemic*⁴ and pay particular attention to the section titled *Political Impact in Colonial India*.

The Indian Plague Commission was instituted by the British government to study the situation, understand the causes of the diseases, and help stop the epidemic. The written report produced by the Commission was very thorough (for full document, see [\[3\]](#)).

²For more information on mathematical epidemiology, see *Computational and Mathematical Epidemiology*, written by Dr. Fred S. Roberts. Science Magazine | Careers, 2004.

³For more information on colonial India, see https://en.wikipedia.org/wiki/Colonial_India.

⁴See https://en.wikipedia.org/wiki/Third_plague_pandemic.

- ★ Please read the excerpt from *The Etiology and Epidemiology of Plague - A Summary of the Work of the Plague Commission* available at [1] (for full version, see [3]). Note that the “General Conclusions” (on page 12 of the linked PDF) are taken from the very end of the report.

When reading, focus your attention on answering the following questions:

1. According to the report, how is the plague transferred to humans?
2. What is definitely not true about the way the plague is transferred to humans?
3. Does any of this surprise you?
4. Do you think there is a parallel between the spreading of the disease among rats and humans?

After this first introduction to the report, choose one Part from it and read it carefully (the table of contents of the full report is on page 7 of the linked PDF). At our next meeting, be ready to present a summary of what you learned to the group. Note that this report was written in 1908 for a technical audience; it is understandable if you don't grasp everything right away. Take some time to think about it and let it sink in.

This short Plague fact sheet [7] from the CDC (Centers for Disease Control and Prevention) answers a lot of questions regarding the disease itself. If you want more information about it, see the full CDC Plague webpage [6].

A.2 The SIR Model

Now that you have some historical context of the plague epidemic, we're going straight to the (mathematical) source.

- ★ Please read the paper that first defined the SIR model as we know it today, by following the instructions below. The manuscript titled *A Contribution to the Mathematical Theory of Epidemics* was written by W.O. Kermack and A.G. McKendrick in 1927 [9].

This is a technical mathematical paper with notation from 1927. You are not supposed to be comfortable with all of its content. It is expected for all students to struggle through the first reading of this manuscript. Skim the entire paper, but pay particular attention to:

- The *Introduction* section.
- The *General Theory* section. This part is long and gets tedious after a while; it's important to understand through the end of page 703. We will discuss this together.

As you're reading, try to take notes of what you're understanding and any question that comes up for you. The ideal way of reading a mathematics paper is to follow the authors' argument by deriving the equations along with them. This may only be possible for parts of this paper, but give it a try! Please answer the following questions:

5. Summarize, in your own words and in a bullet point format, the evolution of an epidemic from beginning to end, as described in the *Introduction*.
6. Look for any assumptions the authors make in the *Introduction*. Do you think each of them is reasonable? Why or why not?
7. How do these assumptions relate to the assumptions or conclusions drawn in the Report? Does anything jump out at you?

We expect students to come to our meetings with questions and comments about what they read. Don't be discouraged if the reading is tough — it's hard for everyone, including the instructors. Be kind to yourself: this is likely your first time reading a technical mathematics paper, and this is a manuscript from 1927!

A.2.1 Derivation

While the early part of the paper should go by more quickly, you may need help in deriving some of the equations, so here are some tips.

The process to derive equation (17) is described on pages 703-706; it involves using infinite series, the method of integrating factors to solve first order linear ODEs, and patience. In particular, the first few lines of the calculation require some leaps of faith on your part, but once the paper shows you the expression for

$$x = f_0(t) + \lambda f_1(t) + \lambda^2 f_2(t) + \dots \quad (\text{A.1})$$

at the top of page 706, you should be able to follow along all the way through to successfully derive equation (17). On page 705, the authors mention that they haven't found a way to solve equation (16) explicitly, but they are going to base their solution process on the observation that (16) is a Volterra-like equation (of the second kind).⁵ Do not focus on this too much, i.e., believe that this is possible and accept their solution form reported in equation (A.1).

A.2.2 Special Cases

After agonizing over some of the details of the *General Theory*, we are now jumping ahead to the *Special Cases* section, Part B. *Constant Rates*, from the bottom of page 712. The "constant rates" referenced in the title of the section are the rate of infectivity $\phi(t) = \kappa$ (pronounced "fee of tee = kappa") and the rate of removal $\psi(t) = \ell$ (pronounced "psi of tee = elle"), where both κ and ℓ are constants. In particular, note:

- (a) The nonlinear system of three first order ODEs on page 713 (equation (29) of the paper) is essentially the first so-called SIR model. The variables x , y , and z represent the number of Susceptible, Infected, and Removed/Recovered individuals in the population, respectively. The total population density is: $x + y + z = N$ (see also *General Theory*).

⁵This is a link to learn more about the Volterra integral equations: https://en.wikipedia.org/wiki/Volterra_integral_equation.

- (b) Pay particular attention to the plot on page 714 and its caption which continues to page 715 (slightly smaller font than the main text). In the plot, the small black dots correspond to the measured data for rat deaths during the Indian Plague epidemic, and the line with the open circles represents the rat deaths per week predicted by this model.

Answer the following questions:

8. Look for any additional assumptions the authors make in the figure on page 714. Do you think these are reasonable? As mentioned, the figure represents rat deaths over time, not human deaths. Does this make sense in the context of the report of the Indian Plague Commission?
9. Overall, is there anything that stood out to you in the parts you read?

Now let's focus on the equations. We have worked on deriving some of them, but their meaning may have gotten lost in the mathematical details. Hence, now we want you to take a step back and really try to understand what the model means. Here are the equations of the SIR (Susceptible, Infected, Removed) model from the 1927 paper by Kermack and McKendrick [9], written using modern notation for the variables, followed by Figure A1, a visualization of the compartment model:

$$\frac{dS}{dt} = -\kappa S(t)I(t), \quad \frac{dI}{dt} = +\kappa S(t)I(t) - \ell I(t), \quad \frac{dR}{dt} = +\ell I(t). \quad (\text{A.2})$$



Figure A1: Visualization of the SIR model described by (A.2). Parameter κ is the rate of infectivity and ℓ is the rate of removal.

Let's look at the equations in more detail. Read below and answer the following questions:

- The dependent variables are S , I , and R , representing the number of individuals (or rats, as in the paper) in each of the Susceptible, Infected, or Removed group, respectively. The independent variable t represents time since the beginning of the infection. The chosen unit for time is selected depending on the situation at hand. The parameters κ and ℓ affect how quickly individuals move from one group to the other. Please answer:
 10. What happens if κ is large? Would people get infected more or less quickly than if κ was small? Explain why.
 11. What happens if ℓ is large? Explain your answer.

- The derivatives dS/dt , dI/dt , and dR/dt represent the net rate of change of the population of each of the three groups due to all the factors taken into account in the model.
- The $\kappa S(t)I(t)$ term is based on the **Law of Mass Action**.⁶ This law states that the rate at which people get infected in a population is dependent on the product of the number of healthy and infected people. Please answer:
 12. If there is a very small number of infected people, I , and a large number of healthy people, S , what will the rate of new infections be?
 13. If there is a large number of infected people, I , and a very small number of susceptible people, S , what will the rate of new infections be? Does this make sense to you?
- The $\ell I(t)$ term is more familiar than you think. This assumes that people get removed from the infectious group, I , continuously at a rate ℓ .
 14. What type of mathematical function would describe this decay accurately?
 15. If $\ell = 0.5$, what is the continuous rate at which people are removed from the infected group at each time unit?
 16. What do the positive and negative signs of each term indicate? (See Figure A.2.2 for a hint.)

Now that you have thought through the meaning of all the terms in the equations, please answer the questions below:

17. Summarize with your own words what the second equation is expressing. Use the diagram of the compartment model in Figure A.2.2 to aid your understanding.
18. What does it mean physically if $dR/dt = 0$?
19. What happens if $\ell = 0$ and $\kappa \neq 0$?
20. What happens if $\ell \neq 0$ and $\kappa = 0$?

We would like to *modernize* the parameters in the equations as well, not just the variables. In the modern notation of the SIR model:

- $\kappa = \beta/N$, where β is called the **rate of contact** and takes into account the probability of contracting the disease when there is contact between a susceptible and an infected individual. It is more realistic to consider a force of infection that does not depend on the absolute number of infectious subjects, but rather on their fraction with respect to the total constant population N .

⁶For more information on the Law of Mass Action, see https://en.wikipedia.org/wiki/Law_of_mass_action. In particular, see the section about Mathematical Epidemiology.

- $\ell = \gamma$, where γ is the **rate of recovery or death**. If the duration of the infection is denoted by D , then $\gamma = 1/D$, since an individual experiences one recovery in D units of time.

Hence, the **modern SIR model** is defined as:

$$\frac{dS}{dt} = -\frac{\beta}{N} S(t)I(t) \quad (\text{A.3a})$$

$$\frac{dI}{dt} = +\frac{\beta}{N} S(t)I(t) - \gamma I(t) \quad (\text{A.3b})$$

$$\frac{dR}{dt} = +\gamma I(t) \quad (\text{A.3c})$$

As mentioned earlier, this is a nonlinear system of three first-order ODEs. In general, it cannot be solved exactly. However, given some assumptions, solutions for special cases can be derived. Let's analyze some key aspects of this system.

21. If we wanted to solve this system exactly, how many initial conditions would we need to fix a value for the constants of integration?
22. Add the equations to one another to verify that:

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0. \quad (\text{A.4})$$

What does this imply about the sum of $S(t)$, $I(t)$, and $R(t)$? (Remember that you know what $S + I + R$ is equal to.)

23. How many equations need to be solved to find an expression for $S(t)$, $I(t)$, and $R(t)$?

A.2.3 The Basic Reproduction Number, R_0

The dynamics of the infectious group depends on the **basic reproduction number**,⁷ defined as $R_0 = \beta/\gamma$. This ratio can be interpreted as the number of cases one case generates on average over the course of its infectious period in an otherwise uninfected population. This is a useful metric because it is understood that for $R_0 < 1$ the infection will die out (in this case the disease is referred to as a 'dud'), while for $R_0 > 1$ the infection will spread in a population (and the disease is referred to as an 'epidemic'). That is because for $R_0 > 1$, the infection rate is large relative to the recovery rate γ , and the total number of people to be infected is expected to be large. On the other hand, if $R_0 < 1$ the recovery rate is fast enough that, while a few people may get infected, the spreading is very slow and not considered to be a full-blown epidemic. See Table A1 below from *History and Epidemiology of Global Smallpox Eradication*⁸ to get an idea of typical values for R_0 for well-known infectious diseases.

⁷See https://en.wikipedia.org/wiki/Basic_reproduction_number.

⁸The *History and Epidemiology of Global Smallpox Eradication* is a module of the training course "Smallpox: Disease, Prevention, and Intervention" from the CDC and the World Health Organization, 2001. The table appears on slide 17.

Disease	Transmission	R_0
Measles	Airborne	12-18
Diphtheria	Saliva	6-7
Smallpox	Airborne droplet	5-7
Polio	Fecal-oral route	5-7
Rubella	Airborne droplet	5-7
Mumps	Airborne droplet	4-7
HIV/AIDS	Sexual contact	2-5
Pertussis	Airborne droplet	5-5
SARS	Airborne droplet	2-5
Influenza (1918)	Airborne droplet	2-3
Ebola (2014)	Bodily fluids	1.5-2.5

Table A1: Values of R_0 for well-known infectious diseases. Taken from *History and Epidemiology of Global Smallpox Eradication* (see footnote 7 for more information).

Now it's time to solve the problem! There are a variety of ways to approach the solution to this system. You should have realized in answering the questions above that we only need to solve two differential equations out of the three to obtain an expression for all S , I , and R . That is because the sum of the three populations is always equal to N (the total population density), so, once we have solved two of the three equations, we can take advantage of this N -fact to write the third solution.

Pair together two equations of your choice and try to solve them. Note that while you have a few options here (three equations can be paired in six ways if the order matters), there are some pairings that make solving easier and others that make it very hard or impossible.

24. Explore the possibilities and report all of your tries, even those that didn't quite work out. Try combining equations by adding, subtracting, multiplying, and dividing them. Remember that our goal is to find a solution in the simplest possible way. These are first-order equations, so let's strive for combining two equations to obtain a simple separable equation to be solved. Use $R_0 = \beta/\gamma$ wherever you can.
25. Can you combine them to solve for $S(t)$? Write a solution for $S(t)$ in terms of $R(t)$.
26. Can you do the opposite? That is, write out a solution for $R(t)$ in terms of $S(t)$?
27. Note that each of these solutions should depend only on one undetermined constant. Take the expression you found for $S(t)$ (in question 25) and set the value of the constant by applying the initial conditions $S(0) = s_0$ and $R(0) = r_0$. Why do we need two conditions for one constant?
28. Now substitute the solution for $S(t)$ found above in equation (A.3b) and solve, using $R(0) = r_0$ and $I(0) = i_0$. You should find an expression for $I(t)$ that depends on $R(t)$ only (i.e., not on $S(t)$).

29. Finally, derive the solution for $R(t)$ based on the relationship between S , I , R , and N (as discussed earlier). You do not need to write this solution explicitly.

If you followed all the steps above, you should now have solutions to all three ODEs, where the susceptible model $S(t)$ depends on $R(t)$ only, the infectious model $I(t)$ depends on $R(t)$ only, and the recovered model $R(t)$ is not written explicitly.

30. Compute the limit as $t \rightarrow \infty$ for $R(t)$. What does this limit represent from an epidemiological point of view?
31. Assume that $t \rightarrow \infty$ represents the *end* of the epidemic and that $S(0) \neq 0$. What does this limit imply with regard to the susceptible population?
32. Based on the conclusion above, how is *the end of an epidemic* defined? What is it caused by?
33. As we discussed earlier, the basic reproduction number R_0 is very important in this model. Rewrite the second equation as

$$\frac{dI}{dt} = \left(\frac{R_0 S}{N} - 1 \right) \gamma I, \quad (\text{A.5})$$

and study the sign of the derivative (i.e., the sign of dI/dt) in terms of R_0 . Can you relate your conclusions to what you learned earlier about R_0 ?

You have now read and understood the original SIR paper, be proud of yourself! This was no small feat! Make sure to take a few minutes to collect your thoughts and summarize the main concepts you learned in a bullet-point list.

A.3 The Ebola Epidemic of 2014-2016 in West Africa

You should be ready to read and thoroughly understand a recent SIAM (Society of Industrial and Applied Mathematics⁹) article discussing the Ebola epidemic of 2014-2016 in West Africa.

- ★ Please read “Emerging Disease Dynamics – The Case of Ebola”, written by Sherry Towers, Oscar Patterson-Lomba, and Carlos Castillo-Chavez. This article appeared in *SIAM News* on November 3rd, 2014.¹⁰ In the article there are a few technical terms. Take a look at section A.4 for some definitions.

⁹The Society for Industrial and Applied Mathematics (SIAM) is an international community of over 14,000 individual members. Almost 500 academic, manufacturing, research and development, service and consulting organizations, government, and military organizations worldwide are institutional members. For more information about SIAM and to learn about student memberships, visit <https://www.siam.org/>.

¹⁰<https://sinews.siam.org/Details-Page/emerging-disease-dynamics>.

Please answer the following questions:

34. What are the variables the article uses in the graphs? Which of the variables in the SIR model you are familiar with do these represent?
35. What is the chosen unit for time?
36. What type of model do they use to fit their data?
37. What do the authors do to validate their model?
38. Was there anything confusing for you in the article? Anything that you think is explained poorly?
39. Can you spot any weaknesses of the SIR model after reading this? What does the basic SIR model not take into account that is discussed in the article?

A.4 Some Vocabulary Related to the SIAM Article

Optimal Control Strategies. A general term indicating the fact that having a limited number of resources means not all possible measures to prevent an outbreak can be implemented. The issue is then: which approaches should be taken to have the greatest impact in slowing the spread of the disease? Full quarantine? Medical treatment? Isolation? Something else? And how much money should be put towards each?

Dimensionless Quantity. Refers to quantities like R_0 which determine the qualitative behavior of a model. Unsurprisingly, the word *dimensionless* specifically refers to the fact that R_0 has no physical units.

Ansatz. A hypothesis, an educated guess, a particular form of a mathematical model. Their model is piecewise exponential. The word *ansatz* comes from German.¹¹

Time series. Basically, a function that depends on time, typically used when referring to repeatedly sampled data over a span of time. Visually, a plot with time on the horizontal axis.

95% Confidence Interval. A statistical term representing the uncertainty in a prediction. Roughly this means that the authors were 95% confident that the true number of new Ebola cases would be somewhere in their predicted interval — though the strict statistical definition of “confidence interval” is more subtle than this.¹²

¹¹For more on the etymology of the word *ansatz* and its use in mathematics, see <https://en.wikipedia.org/wiki/Ansatz>.

¹²For a good explanation of the concept of *confidence interval* see, for example, the video <https://www.youtube.com/watch?v=tFWsu09f74o>.

A.5 Next Steps

Let's recap what you have read and learned:

- Write a short summary of what you've learned. Note the key aspects of an epidemics and what you've learned about modeling it.
- Reflect on the Report and the *SIAM News* article. Draw any parallels you notice and make a list of main differences.
- Think about next steps, taking inspiration from your answers to questions 38 and 39. What is one key aspect of the SIR model and how it is applied that you think needs improvement? What improvements would be most important to you? Spending less money? Saving more lives? Shortening the epidemic? Distributing funds equitably? What else?

We hope you have enjoyed this guided exploration of the SIR model in an historical context. We will now continue working on applying the model to modern data sets of interest based on each student's preference.