DATA SCIENCE FOR YOUTH IN THE TIME OF COVID

Jan Mokros, Science Education Solutions, jmokros@scieds.com
Jacob Sagrans, Science Education Solutions, jsagrans@scieds.com
Pendred Noyce, Tumblehome, pendrednoyce@gmail.com

Through the "COVID-Inspired Data Science through Epidemiology Education" project, 400 underserved middle-school youth across the United States are engaging in a 20-hour out-of-school data club centered on a novel. The narrative is integrated with hands-on data activities and modeling (e.g., creating graphs of infections over time in CODAP; modeling disease transmission rates in Net Logo). Youth learn to: 1) Use data tools to track the spread of a variety of infectious diseases; 2) Ask and address their own questions of data; and 3) Use data to communicate to local audiences about epidemiological patterns and challenges. The project breaks new ground in integrating data science with epidemiology education for 11–14-year-old youth.

BACKGROUND

Worldwide, COVID-19 has become part of the fabric of life. It is far-reaching, it continues to dominate the news, and it is likely to be with us for some time (Osterholm & Olshaker, 2021). The coronavirus has disrupted education and family life as it leaves many of us, including young people, struggling with feelings of uncertainty and helplessness. At the same time, the pandemic offers one of the best examples in our lifetimes of the relevance of data science: Rich data resources are updated daily, a plethora of data tools and visualizations are available to examine these data, and the data matter. The epidemiological data visualizations that are a component of major newspapers each day have become a socio-technical tool (Pea & Cole, 2019) that convey more about data than most adults ever learned in school. How can the data movement inspired by COVID-19 take hold among youth, and what will they learn from their involvement? Our work presents one path to addressing these questions.

IMPORTANT DATA CONCEPTS IN EPIDEMIOLOGY

Epidemiology is the study of health patterns in populations, including the distribution of diseases or states of health in populations and subpopulations, and the causes or risk factors for these states. For infectious diseases in particular, the time course of individual cases and of disease spread are key concepts. Epidemiology rests on large amounts of data gathered repeatedly over time.

While epidemiology concerns populations, it requires multivariate data, with individuals as the "cases." To make epidemiological inferences (e.g., the relationship of lung cancer to smoking, or likelihood of severe COVID infection to age) requires datasets that include multiple attributes associated with any given individual. The World Health Organization and national health organizations provide these types of data to the public for COVID daily new infections, active infections, hospitalizations, deaths, vaccinations, and other outcome variables. In the United States, the data are typically provided by state and county and are often accompanied by individual demographic data on race, ethnicity, age group, and gender.

Knowing how many people are sick in a given population is not very useful unless we also know the size of the population. Denominators are important for understanding rates, which allow one to compare disease burden among different populations and sub-populations. Another epidemiological concept where rate is important is measuring the safety and efficacy of public health measures such as social distancing, medications, or preventive measures such as vaccination. Another rate-based concept for infectious diseases is the reproductive rate (R_0, or R-naught), or the average number of people a given sick person infects. R_0 is a key concept but one that is not easy to observe or derive, as it depends on how long an infectious period lasts, innate characteristics of the infectious agent, and how often and how closely a sick person encounters susceptible others.

INTEGRATING EPIDEMIOLOGY AND DATA SCIENCE EDUCATION: A POWERFUL CONFLUENCE

In the past, youth in the US have typically not learned much about either data science or epidemiology in school (Finzer, 2013; Bracken, 2014). When students do learn about epidemiology,

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

there can be a mismatch between the real tasks of epidemiology and the way that students learn about the field (see, for example, the US Centers for Disease Control and Prevention's epidemiology curricula, which involve mostly small datasets, not real-world, big/"messy" data—CDC, 2021). In our view, when teaching about epidemiology, data science should be at the forefront given the centrality of data to the field, as discussed above. Data science has two defining characteristics (Erickson, 2021): 1) The work often begins with a feeling of being *awash in the data*; and 2) One makes *data moves* to manipulate and tame actual datasets. Erickson asserts that the first task in examining any new dataset is to "explore techniques for finding the patterns and stories in the ocean of data." This can be done by using tools like CODAP (Common Online Data Analysis Platform), a free web-based data tool developed by the Concord Consortium that serves as a platform for educators and a powerful application in the hands of young people.

The availability of public datasets, combined with the evolution of easy-to-use web-based data tools, means that even young students can study infectious disease data in a manner that simulates the work of epidemiologists. In our project, middle school youth explore time-series data about COVID infections and vaccinations, examine relationships between COVID infections and demographic variables, and make comparisons between COVID infections in various countries and US states. CODAP can provide tables listing cases (rows) and attributes (columns); users decide on attributes of interest and drag and drop them onto graphs. For example, as seen in graph below where a student decided to compare China and Italy early on in the pandemic (Figure 1), date can be dragged onto the X-axis and daily new COVID infections can be dragged onto the Y-axis in order to examine how total infections change over time. A third attribute, in this case "country," is shown with different colors.

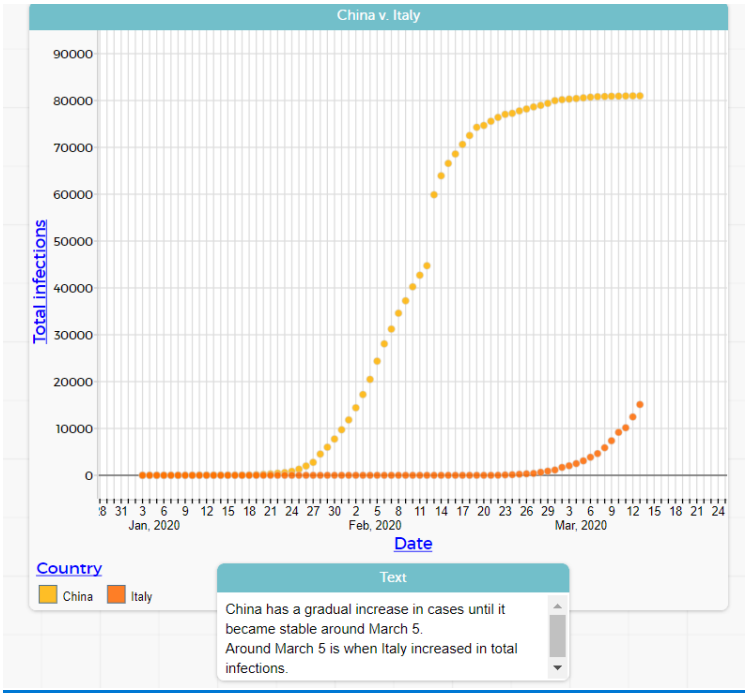

Figure 1. CODAP graph showing total COVID infections in China and Italy, January to March 2020

Our work enables youth to engage with data and mathematics at an appropriate developmental level while they are simultaneously *performing epidemiological work.* The concepts we are helping youth develop include the following:

1) Examining time-series data to determine inflection points, which we describe as places where the graph suddenly turns up or down.
2) Understanding the idea of rate and its importance in epidemiology in order to make fair comparisons across groups of unequal size, such as different US counties or states.
3) Comparing the shape of two or more time-series graphs of infection rates.

4) Looking at relationships between disease outcome variables, such as the relationship between daily infections and cumulative infections.
5) Exploring patterns in relationships between demographic variables and geography and disease outcome variables.
6) Using continually updated data to describe trends and make predictions.
7) Learning about signal vs. noise, primarily through repeated study of time-series graphs that are noisy.
8) Using disease spread computer simulations in conjunction with graphs of infections to understand the ways in which infections are related to R_0 (see the example below taken from one of our CODAP/NetLogo disease spread simulation activities).
9) Learning about the probabilistic nature of modeling disease spread by running repeated simulations with the same starting conditions and noting a range of possible results (see the example below in Figure 2 for three different curves generated by the same parameters).

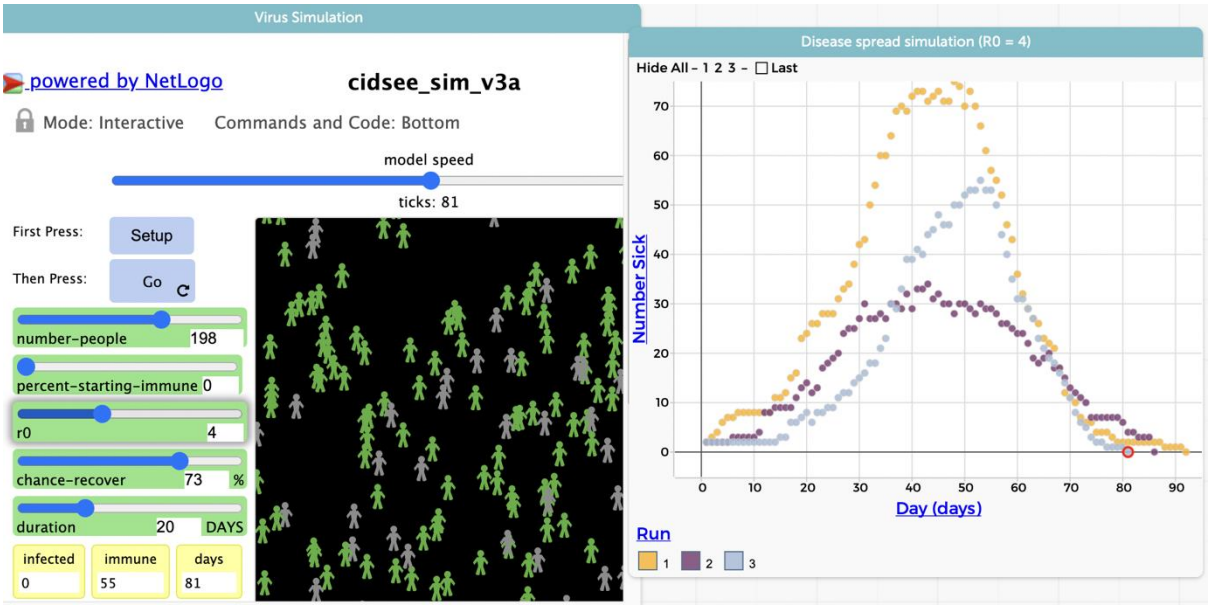

Figure 2. Example of variability in results after multiple runs of a CODAP/NetLogo simulation of disease spread

DATA DETECTIVES CLUBS: A MODEL FOR STUDYING THE CONVERGENCE OF DATA SCIENCE AND EPIDEMIOLOGY

The "Data Detectives Clubs" program is an ambitious 20-hour out-of-school experience that integrates work with epidemiological data into a book club format in afterschool and summer camp programs. The unifying feature of each club is the adventure book *The Case of the COVID Crisis* (Noyce, 2021). The book follows the time travels of two young people as they investigate past outbreaks of measles, smallpox, the 1918 flu, Nipah, and Ebola. The protagonists also travel to different places in current times to learn about COVID-19. Each club session focuses on one chapter, which typically features a major data activity using CODAP (such as those depicted above) and an activity that involves social and emotional learning (SEL). We have also incorporated brief virtual visits from epidemiologists and others who work on infectious diseases, which provide youth with a chance to see actual scientists and interact with them. Initially, each club culminated with youth designing and presenting data-based Public Service Announcements (PSAs) to a community audience of their choosing. However, we quickly learned that youth considered any numbers to be "data" and often came up with artistic renderings of such messages as "Stay six feet apart." Therefore, we have revised the final project to be a "mission" in which youth examine trends in data for a specific US state to make recommendations to the state's school board about policies for school re-opening, masking policies, etc.

After pilot clubs were run virtually and in person, clubs were implemented on a larger scale with approximately 250 eleven-to-fourteen year-old youth during the summer of 2021. Clubs have been popular, and another 200 youth have joined for the fall of 2021. Imagine Science, which coordinates intensive STEM programming to underserved youth through major afterschool and summer programs in the US (Girls, Inc., Boys and Girls Clubs, 4-H, and the Y), is leading this work. It is notable that the Imagine Science partners already have a strong emphasis on STEM, and that they focus on underserved youth. Approximately 86% of the youth in the Imagine Science network are from low-income families, 93% are from racial minority groups, 49% are previously unserved youth, and 62% are girls. To prepare program leaders to implement the Data Detectives Clubs, project staff have designed and implemented a six-hour professional development program that introduces leaders to the epidemiology, data concepts, CODAP, youth activities, and career connections. We are paying particular attention to helping leaders learn to use CODAP because most of them have little or no experience with data science.

PRELIMINARY EVALUATION FINDINGS

We first administered surveys to youth after two pilot rounds of the clubs in 2020. This was a relatively small sample (n = 23), and evaluation questions focused on youth interest and engagement. The data show: 74% of youth became more interested in COVID-19 as a result of their experience; 48% became less anxious about the pandemic (while none became more anxious); and 65% became more confident in their ability to use and interpret graphs. Moreover, their interest in STEM careers grew, with 39% becoming more interested in immunology, 39% in data science, 30% in teaching about disease data, 30% in epidemiology, 22% in virology, and 17% in health policy.

Of the 122 youth who filled out surveys after participating in Data Detectives Clubs in the summer of 2021, 72% reported positive change in STEM engagement due to the program (versus 62% in a national comparison group of grade 4–8 students involved in informal STEM programs) and 53% reported an increase in STEM identity (versus 42% in the national comparison group). 40% of survey respondents reported increased ease in understanding charts and graphs.

Our observational notes show that youth quickly grasped the basic concept of rate when it was placed in the context of fair comparisons. Another observation focused on the use of "Mystery Graphs" showing total and daily COVID infections in various unnamed countries. Youth discovered how to match graphs of daily rates to those of cumulative case numbers as well as how use slope and interpret the shape of the graph to figure out the matching country. They also learned the value of having both actual numbers of infections as well as rates of infections to examine racial disparities in different states and to make comparisons about the COVID-mitigation measures being employed by various countries.

On surveys youth indicated that they appreciated having the opportunity to learn more about a pandemic that is affecting them deeply. This is significant, as some feedback we received in the spring of 2020 was that our project would not be relevant to youth either because the pandemic would soon pass or it would not affect youth personally since they are at less risk of getting seriously ill. We plan to continue keeping the clubs relevant to youth by helping them to better understand the science and data behind what they have already experienced, to appreciate the continued challenges of equitable global COVID-19 vaccine distribution and vaccine hesitancy as well as emerging variants, and to consider lessons learned for future epidemics.

RESEARCH QUESTIONS

In the current large-scale implementation of the project, our research specifically addresses questions about youth's ability to use epidemiological data. Questions include the following:

How do youth use datasets and data tools to study the spread and containment of infectious diseases? How do they ask questions of data, examine patterns, use models, and make predictions? To address these questions, we are collecting artifacts from youth at each site, including graphs, maps, and models made in CODAP and screen-captured and text boxes generated by youth that explain the graphs, models, or maps they generate. In addition to gathering these artifacts, we are asking program leaders to complete periodic brief diaries with prompts that relate to these research questions, such as: When using the CODAP simulations showing spread of infectious diseases, what did youth notice

about different R_0s? What did youth notice about variation between the different times they ran the models with the same R_0?

Through their activities, do youth develop an appreciation for and interest in epidemiological datasets and how to use them to communicate with others? To address this question, we are collecting youths' final presentations linking data to recommendations for school re-opening and masking policies. We will examine the extent to which youth: 1) Choose and compare data in forming their recommendations as opposed to relying on hunches or beliefs; 2) Make sound links between data and recommendations; and 3) Explain and demonstrate data effectively in crafting a persuasive message to a school board audience. Analysis of the presentations will focus on uncovering the strength of the connections youth make between data, message, and perceived audience needs in their own or other communities. In addition, we are addressing our research questions through focus groups with program leaders and youth.

CONCLUSION

Rapid response and fluidity are particularly important when studying data from the COVID-19 pandemic. These data unfold in real time, which makes it essential for us as developers to be nimble in providing curated, up-to-date datasets on COVID. The data on youth vaccinations, for example, is just beginning to be posted at the time of this writing, and we anticipate that youth will show great interest in the rate of vaccinations for their own age groups. "Just in time" learning is a hallmark of our data science education project. New outbreaks (of COVID as well as other diseases) present opportunities for data exploration, but we do not know in advance what data will be available. This also means that the club's data activities will change over time.

We have chosen out-of-school settings to introduce data science and epidemiology as arenas where youth can get their feet wet now and then plunge deeper as they move through their educational pathways and careers. Because data science plays such a pivotal role in epidemiology, the pandemic provides an ideal opportunity for learning how to use datasets and data tools.

ACKNOWLEDGEMENTS

The COVID-Inspired Data Science Education through Epidemiology project is funded by a grant from the US National Science Foundation (NSF), grant number DRL-2048463. We would also like to acknowledge the contributions of the following project partners: The Concord Consortium, Imagine Science, the Jackson Laboratory (JAX), Partnerships in Education and Resilience (PEAR), Strategic Learning Partners for Innovation (SLP4i), Science Education Solutions (SCIEDS), and Tumblehome.

REFERENCES

Bracken, M. B. (2014). Epidemiology as a liberal art: From graduate school to middle school, an unfulfilled agenda. *Annals of Epidemiology, 24*, 171–173. https://doi.org/10.1016/j.annepidem.2013.11.010

Centers for Disease Control and Prevention (CDC). (2021). CDC Science Ambassador educational activities. https://www.cdc.gov/careerpaths/scienceambassador/educational/index.html

Concord Consortium. (2021). Common Online Data Analysis Platform (CODAP). https://codap.concord.org/

Erickson, T. (2021). *Awash in data*. https://codap.xyz/awash/

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2). https://escholarship.org/uc/item/7gv0q9dc

Merrill, R. M. (2010). *Introduction to epidemiology (5th ed.). Jones and Bartlett.*

Neill, C. (2020). The rise of the armchair epidemiologist. HSC Public Health Agency. https://www.publichealth.hscni.net/node/5240

Noyce, P. (2020). *The case of the COVID crisis*. Tumblehome Books. https://tumblehomebooks.org/book/the-case-of-the-covid-crisis/

Osterholm, M. T., & Olshaker, M. (2021). The pandemic that won't end: COVID-19 variants and the peril of vaccine inequity. *Foreign Affairs*. https://www.foreignaffairs.com/articles/world/2021-03-08/pandemic-wont-end

Pea, R., & Cole, M. (2019). The living hand of the past: The role of technology in development. *Human Development, 62(*1–2), pp. 1–3. https://www.doi.org/10.1159/000499618