Workshop 1: Modeling Pandemic Potential for Disease Surveillance

Publication Note: This document is a "white paper" created as background for and a summary of discussion by participants at a virtual workshop held on November 4, 2022. The workshop was supported by the National Science Foundation Predictive Intelligence for Pandemic Prevention (PIPP) Initiative. The contents of this document are the opinions of the authors and participants.

Introduction and Background

Disease modeling in the early days of pathogen emergence is critically important in informing our understanding of the risk posed to the public and relies on robust data sources to improve model specificity and applicability. Traditional public health disease surveillance is the cornerstone of routine disease monitoring and outbreak investigation. However, in many situations, these systems are underfunded, resource-intensive, and some regional areas lack infrastructure for utilization at the onset of an impending outbreak[1–3]. Nontraditional data sources, such as internet-based disease detection and monitoring tools, can allow for rapid dissemination of data for use in real-time[3]. Digital disease detection tools combine multiple data streams—such as news media reports, official reports, social media, and eyewitness reports, among others—to monitor infectious diseases[3,4]. These nontraditional surveillance systems can be informative by highlighting events of interest more rapidly than traditional surveillance programs alone[3,4]. Moreover, these diverse systems are critical for modeling and forecasting efforts early on in the course of an outbreak, when clinical case data can be sparse[5,6].

The epidemic or pandemic potential of a pathogen—whether it be a novel, emerging pathogen (e.g., MERS-*Coronavirus*) or a new outbreak of a common, re-emerging pathogen (e.g., measles *morbillivirus*)—is often determined via estimation of the population-specific reproduction number. The basic reproduction number (R_0) is one of the most commonly cited metrics for describing infectious disease dynamics. R_0 can be used to describe the transmissibility of a pathogen during a given outbreak as the average number of secondary cases arising from a single infected individual in a completely susceptible population[7,8]. While R_0 has historically been modeled as a function of human contact rate, a pathogen's transmissibility, and the duration of the infection[8,9], additional factors can be included depending on the model used to calculate the metric [10]. If a population is not fully susceptible, either due to widespread infection or vaccination, the effective reproduction number (R_{Eff}) is estimated instead[8]. Though useful metrics, the basic and the effective reproduction numbers can lack generalizability due to the spatially heterogeneous nature of human behavior and pathogen characteristics[8,11]. In addition, many methods exist to calculate these metrics and no formal consensus on best practices exists. Moreover, communicating reproduction number estimates to policymakers and the general public poses additional challenges given their fluidity throughout the course of an outbreak and nuance required in their interpretation. Considering the biological, political, and socio-behavioral realities of present-day society may also be necessary to provide more contemporaneous reproduction number estimates in the setting of re-emerging infectious diseases[8].

This workshop explored best practices for the application of novel data sources for disease surveillance and standardization of reproduction number estimates in order to better inform future epidemic response efforts. Discussion among the group highlighted novel data sources for disease modeling that emerged amidst the COVID-19 pandemic, as well as regulatory and practical challenges associated with their use. In addition, workshop participants discussed

hurdles related to R_0 estimation at the onset of the COVID-19 pandemic, as well as communicating early findings to the general public in a nuanced yet approachable manner.

Data Sources for Disease Surveillance and Modeling

Novel data sources used during the COVID-19 pandemic

Several novel data sources were discussed during the workshop. Safegraph mobility data was one example of a dataset that was ultimately used to create thousands of models during the COVID-19 pandemic. However, participants were concerned about the accuracy of these types of data, their validity in different contexts, and subsequent models being built off the same potentially flawed datasets. Moreover, participants reported the co-optation of these types of data for commercialized products. Other data sources used for epidemic modeling included Google and Apple mobility data; however, participants noted that Google mobility data will soon cease to be consistently updated and voiced concerns related to the overall representativeness of this type of data. While these services provide novel data estimating human mobility and by extension its association with disease transmission, there are many regions in the United States (US) and globally where internet penetration remains low. As such, reasonable estimates related to disease modeling cannot always be reliably derived. While the relative availability of these types of data make them attractive for incorporating into models, their ultimate validity remains in question.

Since the beginning of the COVID-19 pandemic, there has also been an increased effort to strengthen the national wastewater data collection infrastructure. Before this pandemic, there were many local and regional labs across the United States working independently to collect and analyze wastewater data for the purposes of public health surveillance. As these systems became increasingly recognized for their ability to identify circulating SARS-CoV-2 variants of concern, a national wastewater surveillance system (NWSS) was established. One workshop participant noted the relative success of wastewater surveillance for poliovirus in New York. Another participant stated that local wastewater surveillance on or near college campuses was successful in determining which dorms had high rates of COVID-19. Despite these successes, one of the ongoing limitations to national wastewater surveillance in the United States identified by workshop participants is the lack of lab standardization which can lead to faulty comparisons of data collected and tested at different locations.

Hospital, medical system, and large claims electronic medical record (EMR) data was discussed as another data source leveraged during the COVID-19 pandemic. While EMR systems have the advantage of pooling large amounts of data to draw inferences about epidemiological associations, they can also vary in quality. Worker turnover during the pandemic, particularly in the public health sector, resulted in a loss of hospital and public health data. It is important to consider the comprehensiveness and quality of data that is collected by healthcare workers, especially in situations where there may be a lack of manpower and worker fatigue. Rural areas in the United States in particular tended to be understaffed and lacked the support and infrastructure for robust data collection.

News media reports were also briefly discussed. News media outlets tend to be the first to report the onset of an outbreak, and academics and public health researchers have long used media reports to inform modeling efforts early in the course of an impending outbreak when traditional data is sparse. The COVID-19 pandemic expanded these opportunities, with additional data collection and data visualization conducted by many news media outlets. Workshop participants introduced the idea of promoting collaborations between academicians

and media outlets to support data collection and collation efforts for future emerging disease events.

Finally, workshop participants briefly discussed the use of genomic surveillance during the pandemic. While genomic surveillance presents unique opportunities to understand disease dynamics, transmission, and evolution, correct interpretation and the subsequent potential for misinformation circulation may be a barrier to its widespread use. Airflow and ventilation dynamics were also discussed as a data source that may inform the design of healthy buildings, particularly in congregate, high-risk settings such as hospitals.

Challenges associated with data access, acquisition, and sharing

Workshop participants discussed the need for the United States to establish a national system for communicable disease data collection and sharing. At present, there are certain federal policies that determine what data gets collected and how; for example, data for nationally reportable diseases are collected, but absence of data for other communicable diseases (i.e., those that are not currently required to be reported) create a fragmented system. Workshop participants reflected on the US experience during COVID-19 when national data systems changed collection and reporting requirements, causing comparability issues. The application of machine learning to parse through and collect data in EMR represents one potential avenue to improve data integrity; however, there are several limitations to this data to consider. For example, at the beginning of the COVID-19 pandemic, the absence of appropriate ICD or billing codes for COVID-19 prevented their use as an effective marker for surveillance. Moreover, variability across US hospital systems in documenting certain disease states as well as lack of interoperability between different EMR systems limit their utility. Participants discussed systems being established among countries in the European Union that allow electronic health data to travel with patients, potentially circumventing issues related to interoperability among health care data systems. There is currently no similar program in the United States which may be associated with a lack of political will.

Workshop participants also discussed the challenges with using news media data. In prior outbreaks, natural language processing (NLP) tools have been used to scrape relevant data such as preliminary case counts and fatalities from news reports. The workshop group discussed that existing NLP packages are not yet as useful, user-friendly, or generalizable as they could be. It would be ideal to be at a point where researchers can easily build their own classification models for novel topics they are interested in (i.e., for which existing tools are insufficient due to their novelty). The consensus was that this may not be realistic, however, as NLP classifiers often require high-volume human data curation at the training and validation phases.

The topic of opportunism in epidemiological research was also discussed as a new challenge that emerged in the midst of the COVID-19 pandemic. Given the potential academic incentives for publication, participants discussed a significant perceived increase in publishing related to COVID-19 by individuals and institutions who had not historically studied communicable diseases, disease dynamics, or epidemic response. This posed a dilemma for the general public, as it became difficult to determine who was a legitimate voice. Workshop participants discussed potential interventions to combat this, including boosting local experts and amplifying their voices, particularly among those with public health communication experience. Opportunistic research can exist within groups as well, for example, an epidemiologist without

experience in mathematical modeling creating models. The FluScape survey was discussed as a fairly successful endeavor at connecting epidemiologists and modelers.

Future directions

Workshop participants discussed the ideal way to aggregate data geographically for future outbreaks. It is critically important that data collection in a specific community with limited existing data infrastructure is conducted with the involvement of stakeholders from that community in order to ensure future sustainability and equitable collection.

Social media data such as Yelp reviews to track foodborne illness or conduct syndromic surveillance are another potential source for data for the future. There are, however, concerns with the limitations of accurately using social media and other online data to accurately predict infectious disease outcomes[12]. There was also discussion regarding how datasets can be provided by private entities, for example, population-level over-the-counter medication purchase data or online shopping cart data. It may be possible to use data brokers to provide this information to researchers; however, requiring fees for data acquisition via private entities may create an unequal landscape for underfunded academics. There is also potential for conflicting or differing incentives for private entities who are conducting their own analyses compared to public health agencies.

Finally, workshop participants discussed an evolution of how "gold standard" data is scrutinized. Now more than ever, there is more attention on the imperfections of all data sets and models, shifting the focus from what is perfect to what is useful. The pandemic also highlighted very real, systematic consequences of data use and misuse at scale; for example, while social media offers a lens on human behavior at a large-scale, there is also the possibility of it being weaponized as a tool for the dissemination of misinformation and disinformation.

Developing Consensus Across Reproduction Number Estimates

R_0 during the COVID-19 pandemic

At the beginning of any epidemic, there is often limited and inconsistent data available. The mass-action compartmental models (e.g., SIR) offer one method to estimate early reproduction numbers; however, these can very easily over or underestimate R_0 due to assumptions surrounding parameterization. R_0 is also very context specific, making it hard to generalize outside of the geographic area where the data used to parameterize a given model was collected. More complex models that include social networks or movements may also be inaccurate in a data-sparse situation. These complexities raise the issue of what metrics *should* be highlighted early in the course of an outbreak—namely, how much weight should R_0 as an estimate of transmission risk be given in this context. Workshop participants discussed whether turning these considerations into a categorical question (i.e., $R_0 > \text{or} < 1$) may be helpful as an outbreak begins to unfold.

Since the pandemic began, R_0 has become a more commonly known term to the general public. Nevertheless, R_0 tends to be more helpful as a decision-making tool at a societal or policy level more than at an individual level. Workshop participants considered a non-numerical scale to communicate risk to the public; however, this becomes complicated if different organizations and institutional bodies employ different metrics. Workshop participants agreed that in the midst of the COVID-19 pandemic, communicating mathematical metrics like R_0 was confusing as the ever-evolving scientific process was visible to the public. It may be more effective to communicate in simpler terms when talking about risk of disease transmission to the public, similar to how weather forecasts are communicated. Statements such as "today is a good day to

wear a mask"—similar to "today is a good day to bring an umbrella"—may be clearer than communicating mathematical metrics.

Challenges associated with R_0 estimation

Data availability was a significant issue in R_0 estimation at the beginning of the COVID-19 pandemic. As in preceding outbreaks, data from related or similar pathogens were initially used to parameterize early models that aimed to estimate R_0. Workshop participants discussed starting with larger-scale (e.g., country-level) estimates in data sparse settings, followed by geographically specific (e.g., city-level) estimates as more granular data becomes available.

Achieving granularity in R_0 estimation, however, was a challenge throughout the COVID-19 pandemic. If estimates are initially derived from larger geographic areas, they will be less informative at a more granular level. The applicability of R_0 ultimately depends on its use and timing. At the beginning of an outbreak, R_0 is primarily employed as a metric to estimate the potential of a localized outbreak to achieve epidemic or pandemic potential, whereas later, there may be a focus on why certain populations or geographic locations have experienced different rates of transmission.

Human judgment models may be a useful adjunct in this context, as work during COVID-19 demonstrated their ability to accurately estimate R_0 early on. Moreover, crowdsourcing may be a useful tool to integrate lived experiences in a local demographic that computational approaches are unable to include. Human judgment is impacted by how an individual is situated in the world; for example, a frontline worker's understanding of the pandemic will most likely be different than someone who is able to stay home. Epidemiologists and other experts are also able to use their judgment when an estimation of R_0 looks unusual given their experiences with past outbreaks. While it may be challenging to convince more computationally minded researchers that there is value in human judgment models, combining human judgment and computational models may ultimately yield more accurate estimates for R_0 in the future.

Future directions

Reproduction number estimates can be challenging to interpret due to spatial heterogeneity. Improving access to, as well as quality and representativeness of, human behavior data to add context to reproduction number estimation is critical. Contact matrices and surveys are good examples of behavioral data that can be collected to inform our understanding of the susceptibility of a given community to a given pathogen. There is a need for interdisciplinary teams—including social scientists, anthropologists, epidemiologists, and others—to be involved in estimating metrics related to disease transmission.

Workshop participants agreed that R_0 should be viewed as a relative value rather than an absolute value. As a metric, it is most useful for policymakers and public health practitioners to take action on disease control and resource allocation measures. Despite this consensus, participants also acknowledged that R_0 has become popularized in the mainstream media and to the general public. Now that the public is aware of this metric, it is important to emphasize that it is not the same for every individual or context. In the future it may be helpful to demonstrate to the public how interventions may change reproduction number estimates (i.e., R_{Eff}) over time.

Summary

Recent innovations in computing, data collection, and tools for disease surveillance offer myriad scientific opportunities to inform communicable disease modeling efforts early in the course of

an epidemic. However, existing logistical and regulatory barriers continue to present challenges for leveraging novel data sources in the future. Perhaps more importantly, accurate and responsible communication of the nuances around disease dynamics—and the models used to estimate them—will be paramount to responding to future epidemics and pandemics.

References

1. Desai A, Nouvellet P, Bhatia S, Cori A, Lassmann B. Data journalism and the COVID-19 pandemic: opportunities and challenges. *The Lancet Digital health*. 2021;3(10):e619-e621.

2. Desai AN, Kraemer MUG, Bhatia S, et al. Real-time epidemic forecasting: challenges and opportunities. *Health Security*. 2019;17(4):268-275.

3. Bhatia S, Lassmann B, Cohn E, et al. Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread. *npj Digital Medicine*. 2021;4(1):1-10.

4. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection--harnessing the Web for public health surveillance. *National institutes of health*. 2009;360(21):2153-2157.

5. Majumder MS, Nguyen CM, Cohn EL, Hswen Y, Mekaru SR, Brownstein JS. Vaccine compliance and the 2016 Arkansas mumps outbreak. *Lancet Infect Dis*. 2017;17(4):361-362.

6. Majumder MS, Kluberg S, Santillana M, Mekaru S, Brownstein JS. 2014 ebola outbreak: media events track changes in observed reproductive number. *PLoS Curr*. 2015;7. doi:10.1371/currents.outbreaks.e6659013c1d7f11bdab6a20705d1e865

7. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am J Epidemiol*. 2013;178(9):1505-1512.

8. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the basic reproduction number (R0). *Emerg Infect Dis*. 2019;25(1):1-4.

9. Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. OUP Oxford; 1992.

10. Dietz K. The estimation of the basic reproduction number for infectious diseases. *Stat Methods Med Res*. 1993;2(1):23-41.

11. Ridenhour B, Kowalik JM, Shay DK. Unraveling R0: considerations for public health applications. *Am J Public Health*. 2014;104(2):e32-e41.

12. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science*. 2014;343(6176):1203-1205.