Research Ideas and Outcomes 6: e54770
doi: 10.3897/rio.6.e54770

Grant Proposal

Quantifying the Impact of Data Sharing on Outbreak Dynamics (QIDSOD)

Daniel Mietchen‡, Jundong Li§

‡ School of Data Science, University of Virginia, Charlottesville, United States of America
§ University of Virginia, Charlottesville, Virginia, United States of America

Corresponding author: Daniel Mietchen (daniel.mietchen@virginia.edu), Jundong Li (jl6qk@virginia.edu)

Reviewable v1

Received: 27 May 2020 | Published: 27 May 2020

Citation: Mietchen D, Li J (2020) Quantifying the Impact of Data Sharing on Outbreak Dynamics (QIDSOD). Research Ideas and Outcomes 6: e54770. https://doi.org/10.3897/rio.6.e54770

Abstract

In this project, we will explore the range of data-related decisions made during public health emergencies like the ongoing COVID-19 pandemic and analyze the flow of information, data, and metadata within networks of such decisions.

Data sharing is now considered a key component of addressing present, future, and even past public health emergencies, from local to global levels. Researchers, research institutions, journals and others have taken steps towards increasing the sharing of data around the ongoing COVID-19 pandemic and in preparation for future pandemics.

We will quantify the effects of data flow modifications to identify parameter sets under which specific modes of sharing or withholding information have the largest effects on outbreak dynamics. For these high-impact parameter sets, we will then assess the current and past availability of corresponding data, metadata, and misinformation, and estimate the effects on outbreak mitigation and preparedness efforts.

Keywords

data sharing, public health emergencies, epidemiological modelling, network dynamics, decision making, data integration, outbreak management, disaster preparedness, misinformation

List of Investigators

This is a collaborative research project with research teams from two different schools at the University of Virginia (UVA), including the Co-PI Jundong Li from the School of Engineering and Applied Science (SEAS) and Co-PI Daniel Mietchen from the School of Data Science (SDS).

Significance of the research question to global infectious diseases

Public health emergencies require profound and swift action at scale with limited resources, often on the basis of incomplete information and frequently under rapidly evolving circumstances. The sharing of data and associated metadata is a relatively new flavor under this broader theme, but one that has been receiving steadily growing attention over the last few years, especially in the context of Public Health Emergencies of International Concern like the Zika outbreak[*1]. By now, we have reached a point where data sharing must be considered a key component of addressing present, future and even past public health emergencies, from local to global levels[*2]. In response, researchers, research institutions, journals, funders, and others have taken steps towards increasing the sharing of data around ongoing public health emergencies and in preparation for future ones[*3]. These measures range from the adoption of open lab notebooks to modifications of policies and funding lines, and they include conversations around infrastructure, cultural change, misinformation or data ethics (e.g. Ekins et al. 2016).

In this project, we will explore the range of data-related decisions made during example outbreaks and analyze the flow of information, data and metadata through the decision network with respect to different types of modifications of such flows. On that basis, we will quantify the effects of such flow modifications on outcome measures relevant to various stakeholders from local to global levels, so as to identify parameter sets under which specific modes of sharing or withholding information have the largest effects. For these high-impact parameter sets, we will then assess the current and past availability of corresponding data, metadata and misinformation, estimate the effects on outbreak mitigation and preparedness efforts and explore mechanisms through which that availability could be optimized.

Approach

The proposed research aims to bridge the knowledge gap between what we have (i.e., different formats of data sharing/withholding decisions and various stakeholders involved during an infectious disease outbreak) and what we need (i.e., quantify the impact of different data-related decisions during outbreaks). This project seeks to address the following three research aims.

Research Aim 1: Model the flow of data/information through a decision network

Disease outbreaks and other public health emergencies involve a potentially wide range of different stakeholders. These stakeholders are typically interconnected by various types of interactions within and between themselves as well as with the pathogen and the respective social, natural and built environments. We propose to model the interactions among different stakeholders as a heterogeneous decision network where nodes denote different stakeholders and edges denote different types of interactions among them, while the interactions can be characterized based on the availability and sharing of data pertaining to said interactions, e.g. as to whether the pathogen is known, how it can be transmitted, or whether vaccination or treatment is available and how much it costs. We refer to the sharing of data or metadata as data flow, which can be modulated in several ways as one interaction triggers the next. For instance, a student might decide to change their behavior towards others based on the outcome of a diagnostic test, and authorities might decide to temporarily close the affected school or not.

For various reasons (e.g., societal, political, technical, or ethical), much of the data relevant for a complete path through such a decision network may not be directly accessible, posing challenges to investigating its potential and real decision chains. For example, outbreak propagation could be better characterized, understood, predicted and communicated if precise demographic information and migration information were available for individuals near the epicenter (McDonald et al. 1992,Pastor-Satorras and Vespignani 2001). The existence of such data, e.g. in government databases, does not generally imply its availability to other stakeholders, though sometimes, similar insights can be gleaned from other sources, e.g. mobility data from providers of transport services or fitness apps [*4*5].

The modeling of the flow of data, metadata and related information through such a decision network enables us to create hypotheses to investigate the impact of the availability and quality of data on specific kinds of decisions, or with respect to specific stakeholders, locations or timing. In designing the model, we will initially focus on data from Public Health Emergencies of International Concern as well as the WHO's R&D Blueprint[*6], keeping in mind the applicability to other epidemiological contexts such as seasonal flu or the opioid crisis.

Research Aim 2: Quantify the impact of data sharing for decision making

Once the structure of the decision network and the nature of potential data flows has been captured in the model, we can study causal relations between individual or aggregated decisions (such as the closure of one school or multiple schools), the associated data flows, and the spread of the pathogen and the disease. Such decision-making processes are often based on randomized controlled experiments (Guo et al. 2020, Kallus and Zhou 2019). However, they could be expensive in multiple dimensions, including in terms of time and financial resources, which makes them challenging to perform in outbreak contexts. One way to address this is to consider observational data of different stakeholders to infer causal effects between a specific decision (e.g., closure of school) and an important outcome (e.g. spread or containment of disease; Pearl 2009,Rubin 1978).

Learning such treatment effects from observational data as in the mobility example above requires us to handle confounding bias, which are unobserved variables that influence both the treatment and the outcome. For example, an individual's poor socioeconomic status can affect their living conditions and may increase their chances to be infected, treated or cured. In addition to the observational data, we can include measures of the information flow between stakeholders (e.g. hygienic advice, or rumors) to infer the existence of hidden confounders (Hill 2011,Guo et al. 2019). To be able to adapt the decision network model to the nature of a given real or hypothetical outbreak scenario, data availability can be modeled generically as a systemic property, more granularly at the level of classes of stakeholders or decisions, and in a yet more fine-grained manner at the level of individual stakeholders or interactions if suitable data are available or can be inferred.

Research Aim 3: Leveraging the information for current and future outbreak management

For a given outbreak management context, we can fine-tune the model parameters for that context in order to use the model to address questions that might arise during outbreak management. While classical epidemiological modeling provides information on expected outbreak dynamics and recommendations on outbreak management like how much of what to stockpile, our model would provide additional insights into whether, how and when details about preparedness and response or associated research should be shared and with whom, or what data quality requirements should be aimed for at which junctions in the network.

These details might range in scope from individual patients to triage protocols used by health workers to institutional or international policies about sharing diagnostic kits, material samples or computational pipelines. In short, the decision network model would behave much like a machine-actionable data management plan for the outbreak in question, which would also allow, for instance, to notify specific stakeholders of data-related outbreak developments relevant to them (Miksa et al. 2019).

Potential impact of the expected outcomes of the research

The research is expected to yield best practice recommendations in terms of the sharing of data and metadata in the context of specific outbreak-related decisions by stakeholders ranging from individuals to groups and institutions to governments and international bodies. Besides identifying recommendable data sharing scenarios, we will also consider the effects of delays in data sharing, partial sharing as well as the spread of misinformation.

Availability of data and code

To the extent possible, we will follow best practices in sharing our code and data as well as associated documentation, as laid out by Barton et al. 2020. To this end, we have set up a GitHub organization at https://github.com/QIDSOD, which will be our default mode for sharing non-confidential aspects of the project. We welcome and encourage community participation throughout the project.

Funding program

The project is funded by a COVID-19 Rapid Response grant jointly provided by the Global Infectious Diseases Institute (GIDI) at the University of Virginia, in partnership with the Office of the Vice-President for Research of the University of Virginia. It is based on a proposal originally submitted to GIDI's Collaborative Seed Grants program on 2 March 2020.

Grant title

Quantifying the Impact of Data Sharing on Outbreak Dynamics (QIDSOD)

Hosting institution

The School of Engineering and Applied Sciences and the School of Data Science at the University of Virginia.

Ethics and security

At the time of writing, the project has not been assessed externally in terms of its ethics and security implications, but the ethical and security aspects of data-related decision-making during public health emergencies are within the scope of the project, and we will document them as well as our interactions with relevant oversight bodies as the project progresses.

Author contributions

Both authors contributed equally to the design of the research and the writing of this proposal.

Conflicts of interest

None.

References

- Barton CM, Alberti M, Ames D, Atkinson J, Bales J, Burke E, Chen M, Diallo SY, Earn DD, Fath B, Feng Z, Gibbons C, Hammond R, Heffernan J, Houser H, Hovmand P, Kopainsky B, Mabry P, Mair C, Meier P, Niles R, Nosek B, Osgood N, Pierce S, Polhill JG, Prosser L, Robinson E, Rosenzweig C, Sankaran S, Stange K, Tucker G (2020) Call for transparency of COVID-19 models. Science 368 (6490): 2-483. https://doi.org/10.1126/science.abb8637
- Ekins S, Mietchen D, Coffee M, Stratton TP, Freundlich JS, Freitas-Junior L, Muratov E, Siqueira-Neto J, Williams AJ, Andrade C (2016) Open drug discovery for the Zika virus. F1000Research 5 https://doi.org/10.12688/f1000research.8013.1
- Guo R, Li J, Liu H (2019) Learning Individual Causal Effects from Networked Observational Data. arXiv:1906.03485 [cs].
- Guo R, Cheng L, Li J, Hahn PR, Liu H (2020) A Survey of Learning Causality with Data: Problems and Methods. arXiv:1809.09337 [cs, stat] https://doi.org/10.1145/3397269
- Hill J (2011) Bayesian Nonparametric Modeling for Causal Inference. Journal of Computational and Graphical Statistics 20 (1): 217-240. https://doi.org/10.1198/jcgs.2010.08162
- Kallus N, Zhou A (2019) Confounding-Robust Policy Improvement. arXiv:1805.08593 [cs, stat].
- McDonald CJ, Hui SL, Tierney WM (1992) Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. M.D. computing : computers in medical practice 9 (5): 304-12.
- Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. PLOS Computational Biology 15 (3). https://doi.org/10.1371/journal.pcbi.1006750
- Pastor-Satorras R, Vespignani A (2001) Epidemic Spreading in Scale-Free Networks. Physical Review Letters 86 (14): 3200-3203. https://doi.org/10.1103/physrevlett.86.3200
- Pearl J (2009) Causal inference in statistics: An overview. Statistics Surveys 3: 96-146. https://doi.org/10.1214/09-ss057
- Rubin D (1978) Bayesian Inference for Causal Effects: The Role of Randomization. The Annals of Statistics 6 (1): 34-58.

Endnotes

*1 For an overview, see the Scholia profile for Zika virus and data sharing, accessible via https://tools.wmflabs.org/scholia/topics/Q5227350,Q202864 (archived at http://web.archive.org/web/20200526015751/https://tools.wmflabs.org/scholia/topics/Q5227350,Q202864).

*2 For an introductory overview, see https://github.com/Daniel-Mietchen/events/blob/master/Environmental-Health-Seminar-Env-H-580-on-2019-01-31.md (archived at http://web.archive.org/web/20200526023807/https://github.com/Daniel-Mietchen/events/blob/master/Environmental-Health-Seminar-Env-H-580-on-2019-01-31.md).

*3 For instance, see "Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak" at https://wellcome.ac.uk/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-covid-19-outbreak (archived at http://web.archive.org/web/20200216100815/https://wellcome.ac.uk/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-covid-19-outbreak).

*4 Hern A (2014) New York taxi details can be extracted from anonymised data, researchers say. *The Guardian*. Accessible via https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn. Archived at http://web.archive.org/web/20200302195328/https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn.

*5 Sly L (2018) U.S. soldiers are revealing sensitive and dangerous information by jogging. *The Washington Post*. Accessible at https://www.washingtonpost.com/world/a-map-showing-the-users-of-fitness-devices-lets-the-world-see-where-us-soldiers-are-and-what-they-are-doing/2018/01/28/86915662-0441-11e8-aa61-f3391373867e_story.html. Archived at http://web.archive.org/web/20200302192229/https://www.washingtonpost.com/world/a-map-showing-the-users-of-fitness-devices-lets-the-world-see-where-us-soldiers-are-and-what-they-are-doing/2018/01/28/86915662-0441-11e8-aa61-f3391373867e_story.html.

*6 World Health Organization (2020) R&D Blueprint. Accessible at https://www.who.int/blueprint/en/. Archived at https://web.archive.org/web/20200213011758/https://www.who.int/blueprint/en/.