Applying Time-Valued Knowledge for Public Health Outbreak Response

James Thomas Schlitt

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics, and Computational Biology

Dr. Bryan Lewis (Co-Chair)
Dr. Stephen Eubank (Co-Chair)
Dr. Kaja Abbas
Dr. Laura Hungerford

April 18th, 2019
Blacksburg, Virginia

Keywords: Epidemiology, outbreak science, open data, open science, infectious diseases, opioids, addiction, social media, twitter, agent-based modeling, influenza, influenza like illness, modeling and simulation, SEIR, preparedness, antiviral, vaccine

Applying Time-Valued Knowledge for Public Health Outbreak Response

James Thomas Schlitt

ABSTRACT

During the early stages of any epidemic, simple interventions such as quarantine and isolation may be sufficient to halt the spread of a novel pathogen. However, should this opportunity be missed, substantially more resource-intensive, complex, and societally intrusive interventions may be required to achieve an acceptable outcome. These disparities place a differential on the value of a given unit of knowledge across the time-domains of an epidemic. Within this dissertation we explore these value-differentials via extension of the business concept of the time-value of knowledge and propose the C4 Response Model for organizing the research response to novel pathogenic outbreaks.

First, we define the C4 Response Model as a progression from an initial data-hungry collect stage, iteration between open-science-centric connect stages and machine-learning centric calibrate stages, and a final visualization-centric convey stage. Secondly we analyze the trends in knowledge-building across the stages of epidemics with regard to open and closed access article publication, referencing, and citation. Thirdly, we demonstrate a Twitter message mapping application to assess the virality of tweets as a function of their source-profile category, message category, timing, urban context, tone, and use of bots. Finally, we apply an agent-based model of influenza transmission to explore the efficacy of combined antiviral, sequestration, and vaccination interventions in mitigating an outbreak of an influenza-like-illness (ILI) within a simulated military base population.

We find that while closed access outbreak response articles use more recent citations and see higher mean citation counts, open access articles are published and referenced in significantly greater numbers and are growing in proportion. We observe that tweet viralities showed distinct heterogeneities across message and profile type pairing, that tweets dissipated rapidly across time and space, and that tweets published before high-tweet-volume time periods showed higher virality. Finally, we saw that while timely responses and strong pharmaceutical interventions showed the greatest impact in mitigating ILI transmission within a military base, even optimistic scenarios failed to prevent the majority of new cases. This body of work offers significant methodological contributions for the practice of computational epidemiology as well as a theoretical grounding for the further use of the C4 Response Model.

Applying Time-Valued Knowledge for Public Health Outbreak Response

James Thomas Schlitt

GENERAL AUDIENCE ABSTRACT

During the early stages of an outbreak of disease, simple interventions such as isolating those infected may be sufficient to prevent further cases. However, should this opportunity be missed, substantially more complex interventions such as the development of novel pharmaceuticals may be required. This results in a differential value for specific knowledge across the early, middle, and late stages of epidemic. Within this dissertation we explore these differentials via extension of the business concept of the time-value of knowledge, whereby key findings may yield greater benefits during early epidemics. We propose the C4 Response Model for organizing research regarding this time-value.

First, we define the C4 Response Model as a progression from an initial knowledge collection stage, iteration between knowledge connection stages and machine-learning-centric calibration stages, and a final conveyance stage. Secondly we analyze the trends in knowledge-building across the stages of epidemics with regard to open and closed access scientific article publication, referencing, and citation. Thirdly, we demonstrate a Twitter application for improving public health messaging campaigns by identifying optimal combinations of source-profile categories, message categories, timing, urban origination, tone, and use of bots. Finally, we apply an agent-based model of influenza transmission to explore the efficacy of combined antiviral, isolation, and vaccination interventions in mitigating an outbreak of an influenza-like-illness (ILI) within a simulated military base population.

We find that while closed access outbreak response articles use more recent citations and see higher mean citation counts, open access articles are growing in use and are published and referenced in significantly greater numbers. We observe that tweet viralities showed distinct benefits to certain message and profile type pairings, that tweets faded rapidly across time and space, and that tweets published before high-tweet-volume time periods are retweeted more. Finally, we saw that while early responses and strong pharmaceuticals showed the greatest impact in preventing influenza transmission within military base populations, even optimistic scenarios failed to prevent the majority to new cases. This body of work offers significant methodological contributions for the practice of computational epidemiology as well as a theoretical grounding for the C4 Response Model.

Dedication

I dedicate this work to all of the brilliant and dedicated scientists, public health professionals, and volunteer first responders I have met whom sacrifice so much for the satisfaction of giving a little back; to my ornery ginger-cat, Rev. Dr. Baron Rufus T. Cat; to my family; to a better future for my beloved nieces; and to Jenn, who has been swell.

Acknowledgements

My time with the GBCB program, the Biocomplexity institute, and the NDSSL has erased any lines I have ever held between friends and colleagues. This place has been my home, a place where I've grown so much, and where I have been given so many opportunities to do good. I owe my endless gratitude to everyone here who was with me on this journey.

First and foremost I'd like to thank my advisors and committee members, Dr. Bryan Lewis, Dr. Stephen Eubank, Dr. Kaja Abbas, and Dr. Laura Hungerford for their wisdom and support throughout this project. Bryan, you're the best boss I've ever had, a fantastic mentor, and a true friend. I hope we never fall out of touch.

Secondly, I'd like to thank Dr. David Bevan and Dr. Christopher Lawrence of the GBCB program as well as Dr. Madhav Marathe of the NDSSL for taking me in and giving me a chance to be a scientist again.

I thank Paige Bordwine who provided the inspiration and first audience for ChatterGrabber years ago, from which so many great projects and collaborations have sprung.

I owe a great debt to Dennie Munson for being an amazing ally, for putting out so many fires throughout my time in the GBCB program, and for recognizing when I needed a friend and some sound advice during hard times.

I'd also like to acknowledge the students of the NDSSL Public Health Group. Meghana Cyanam, Alex Telionis, Gloria Kang, Daniel Chen, Gabrielle Smith, Arin Davis, and Arinjoy Basak. You all are the greatest group of friends I've ever had. The hardest thing about writing this dissertation is knowing that each chapter I complete is another chapter that brings me closer to leaving Blacksburg.

I thank Allison Reddy, you helped me write my application, my dissertation, and my slides. Your wisdom and friendship has seen me through the entire way, and I don't know if I would have ever attempted this without you.

Finally, I thank Becky Freemal, Alex Crookshanks, Kris Monger, and Dan Rosplock, you made our work shine, and you've brought so many opportunities to the students of this lab.

It has been such an honor knowing you all.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABM: agent-based model

ADM: active duty military

API: application-programming interface

APL: antiviral prophylaxis length

APTD: antiviral prophylaxis trigger delays

ATE: antiviral treatment efficacy

ATL: antiviral treatment length

C4 Response Model: collect, connect, calibrate, and convey

CBRN: chemical, biological, radiological, and nuclear agents

CDC: Centers for Disease Control and Prevention

CHIKV: Chikungunya virus

CHIKVD: Chikungunya virus disease

CNIMS: comprehensive national incident management system

CSV: comma separated values

CW: civilian workers

DARPA: Defense Advanced Research Project Agency

DOI: digital object identifier

EHF: Ebola hemorrhagic fever

EHV: Ebola hemorrhagic virus

GDI: Google Docs Interface

ID: identity number

ILI: influenza-like-illness

KDE: kernel density estimation

MERS: Middle East respiratory syndrome

MERS-CoV: Middle East respiratory syndrome coronavirus

SARS: Severe acute respiratory syndrome

SARS-CoV: Severe acute respiratory syndrome coronavirus

SQL: sequestration length

SQTD: sequestration trigger delay

t_0: time zero

VE: vaccine efficacy

VTD: vaccine trigger delay

WHO: World Health Organization

ZIKV: Zika virus

ZIKVD: Zika virus disease

Preface

The following chapters are co-authored manuscripts, which are in preparation for journal submission:

- **Chapter 2:** *Schlitt, J., Lewis, B., Eubank, S. (2019). The State of Open Access in Outbreak Science.*
- **Chapter 3:** *Schlitt, J., Truong, V., Lewis, B., Eubank, S. (2019). Opioid Social Media Message Mapping.*
- **Chapter 4:** *Schlitt, J., Lewis, B., Eubank. S. (2019). Planning Protection of Personnel: Optimizing Force Readiness in Response to Respiratory Infectious Diseases*

Chapter 1

Introduction

1.1 Dissertation Overview

Presented herein is a collection of works disparate in content yet connected in purpose. Each of these chapters represents a scholarly work pending submission and publication with a focus upon the development of analytical systems to facilitate decision-making across the early stages of an epidemic. Public health professionals are frequently challenged to make critical policy decisions in response to temporally limited knowledge. Such scenarios may include investigating heterogeneities in the case distribution of a novel and fatal pathogen as with the 2013 Middle East Respiratory Syndrome (MERS) outbreak (Coleman and Frieman 2013), correlating the sudden explosive growth of a known pathogen to cultural practices as with the 2014 West Africa Ebola outbreak (Curran et al. 2016), or modeling the human response to unprecedented mass casualty incidents involving chemical, biological, radiological, and nuclear agents (CBRN) (Parikh et al. 2013). In each case, careful consideration of the differences between the scarce observations of a novel crisis and historical data from a richly catalogued past crisis may yield valuable knowledge to improve outcomes and response efforts for poorly understood outbreaks.

The available knowledge regarding the population dynamics, pathogen dynamics, and the ground truth of a public health crisis will grow and mature significantly with time. In turn, the value of individual supportive efforts to the broader epidemiological community will change as academics, responding agencies, and communities react to each stage of a given crisis. Where early outbreak response efforts may be dominated by mass-data collection tasks and pathogen characterization through laboratory sciences, later efforts may draw upon insights connected from a broader range of disciplines as population-level measures become more essential. These insights in turn must be filtered through public health practicalities to be conveyed in an actionable way. Fortunately, the explosive growth of open data repositories (Gurin 2014), open-source analytical tools, open science journals, and social media application programming interfaces (APIs) have provided promising new platforms and methodologies with which such stakeholders may incrementally ***collect, connect, calibrate,*** and ***convey*** critical knowledge. This four-fold approach is known herein as the **C4 Response Model.**

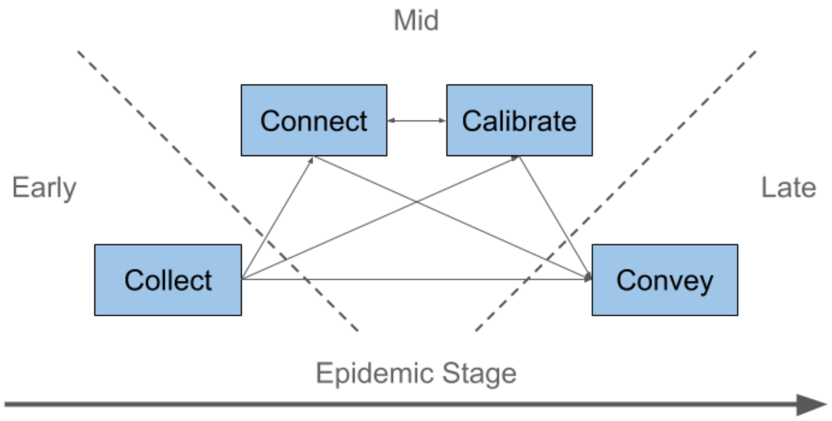
We begin this effort with a systematic literature review to identify key trends in the recent contributions of open science journal publications in academic outbreak response efforts across these stages of knowledge. Following this, we explore the utility of social media for improving the practice of public health message mapping through the development of an easy-to-deploy social media surveillance system and its application in a case study on the opioid epidemic. Lastly, we describe a modeling study built upon a pool of late-epidemic knowledge sources to build a protocol for pandemic influenza response for a military base.

1.2 The Time-Value of Epidemiological Knowledge

The concept of the time-value of money is well-supported in business philosophies, simply observing that as one's means change, the personal value of a fixed unit of wealth may change drastically in regards to one's own personal context. For example, humans have long understood that possessing a given sum of money in the present is preferable to waiting for an identical sum of money later. Having control of resources sooner rather than later removes uncertainty and allows for growth in value through, e.g., investment. This is a lodestar principle of economics and finance known as the time-value of money. A parallel but less-discussed concept exists in the time-value of knowledge (Dalmaris, Hall, and Philp 2006). Delmaris, Hall, and Philp defined the time-value of knowledge as a qualification to assist in evaluating the utility of knowledge in resolving challenges over diverse temporal domains (2006). While this initial coinage focused largely upon business competitive intelligence and the intractable consequences of past organizational decisions, there is unique potential to extend this concept to epidemiological practice. Indeed, epidemiological response efforts follow distinct temporal domains across the early, mid, and late stages of an epidemic. Novel pathogens provide a natural analogy to business competitors as well, as the strength, position, resiliency, and actions available to both an outbreak as a metaphorical entity and those responding to it may shift dramatically based upon previous actions. Knowledge as simple as recognizing the presence of a novel pathogen during the early stages of an epidemic may be sufficient to drive effective isolation and quarantine interventions, thereby crushing an epidemic before it starts. Should this window of opportunity be lost, public health professionals will require knowledge of far greater cost and complexity to curtail the growth of an epidemic through pharmaceutical interventions. Given the uncertainty across each stage of an outbreak response, it is vital that early knowledge-building efforts seek to bolster early, low-hanging fruit intervention targets while also laying the groundwork for later knowledge-building efforts.

The unifying objective of this body of work is to leverage novel information sources to assist public health decision-making with regards to this time-value of knowledge over the stages of a novel epidemic event. I posit that optimal academic responses to novel crises will follow the proposed **C4 Response Model** in which researchers follow a progression from an initial data-hungry **collect** stage, iteration between open-science-centric **connect** stages and machine-learning-centric **calibrate** stages, to a final visualization-centric **convey** stage. The research products of these stages may be substantially improved in both pace and thoroughness by the integration of open science tools and data sources.

Figure 1.1: The C4 Response Model



Many public health agencies have sought to prescribe a standardized sequence of actions in response to emergent threats. Notable examples of such include the CDC's *"Epidemiologic Steps of an Outbreak Investigation"* (Dicker et al. 2006) and the Wisconsin Department of Health's "*Disease Cluster Investigation and Analysis Protocol" (Fiore, Hanrahan, and Anderson 1990)*. These traditional programs may be broken down into the following four stages: (1) laboratory confirmation of the outbreak, (2) epidemiological hypothesis generation, (3) epidemiological hypothesis testing, and (4) institutional response (Murhekar et al. 2009). While these methods serve well to direct boots-on-the-ground responses by public health agencies and medical resources, they have reduced applicability within the realm of computational research.

The **C4 Response Model** is novel in that it specifically focuses on computational researchers' utilization of open or public data sources in support of larger response efforts. While previous studies have focused on the ability of Google Trends to predict epidemics (Cervellin G Comelli I Lippi 2017), trends in published epidemiological topics and sources (Ebrahim and Davey Smith 2018), and trends in epidemiological knowledge sources (Porta et al. 2013), very little attention has been given to underlying trends in outbreak response in open access publications. Likewise, while message mapping has frequently explored message effectiveness as a function of contents, delivery, and origination (Glik 2007), we innovate significantly over previous methods by using tweet virality as a metric of message engagement with regards to opioid messaging. In this case, we defined the virality of a tweet as the total number of potential exposures of a given tweet through the original tweet and all resulting retweets divided by the total number of potential exposures of the original tweet. Potential exposures were described as the sum of the number of followers and friends for a given profile. We applied analyses of virality to assess the success of current approaches as a function of timing, source, and message content. Finally, we **connected, collected,** and **calibrated** data within synthetic populations of statistically representative human agents via ABMs to **convey** detailed visualizations and granular analyses of human behaviors in response to biological and nuclear threats.

In this dissertation I offer a theoretical grounding of the C4 Response Model via analysis of the temporal availability of knowledge following outbreaks, a Twitter methodology for syndromic surveillance and message mapping, and visual analytical tools for multidimensional parameter sweeps of agent-based models of influenza-like-illness (ILI) transmission. The broad impact of these outcomes will be to demonstrate the utility of the C4 Response Model for directing academic research and the tools developed therein to provide policy support across the early, mid, and late stages of a crisis.

1.3 Outbreak Science

While every epidemic response is different, certain trends prove consistent over time. Disease histories, emergency declarations, media attention, and public interest play key roles in the degree and modes of engagement shown by the research community. Crude, early pathogen behavioral parameter estimates and bulk data **collection** may prove extremely valuable within the initial stages of a crisis yet only offer marginal utility as **connected** and **calibrated** knowledge percolates forth. Further, critical intervention targets may only be identified through the **connection** of later, mid-epidemic knowledge when large-scale heterogeneities become more easily observable. One prominent example of such an event occurred during the 2014 MERS epidemic, where knowledge of the gender and age trends amongst early human cases eventually led to the identification of the zoonotic component of the outbreak, as camel races and markets were disproportionately attended by older men (Azhar et al. 2014). However, one potential bottleneck to this distribution of knowledge is the gating and compartmentalization of knowledge, be it within individual research efforts or behind publisher paywalls. If critical early findings remain locked within closed access journals, the scientific community's use of such knowledge may be decreased or delayed when responding to urgent crises. Such delays could directly cost human lives.

Journal publication and citation analyses provide a means of assessing the impact of these bottlenecks. If we assume citations present a reasonable metric for publication utility, then we may utilize citation counts to measure the utility of open and closed access knowledge over time, both with regards to the calendar year of publication and the days elapsed since the start of a singular epidemic. Through this, we offer a direct means of comparing the utility of open and closed access journal articles for outbreak science. In **Chapter 2** we explore analyses to identify which past reference materials provided the most recent knowledge or the greatest repositories thereof at the start of epidemics. Further, we explore how quickly open and closed access journal articles were published in response to each epidemic, which article access-type was greater in number, and when each article access-type saw the greatest number of citations after the start of each epidemic. Finally, we explore communities of methods drawn from PubMed publication qualifiers and propose a sequence by which each community showed optimal citing utility starting from the first cases of an epidemic. Our findings suggest that closed access publications referenced newer knowledge with respect to both their time of publication and the outbreak itself. Closed access publications also saw higher mean citations during the early stages of most epidemics, as key findings see early, highly cited articles in prestigious journals. We also observe that open access journals ultimately constitute the vast bulk of outbreak science, both in terms of prior journal articles referenced by open and closed access outbreak response publications and in their sheer volume. Further, the number and proportion of open access journal articles has grown consistently over time in PubMed articles in general and outbreak response articles in particular.

Incentivizing research into highly citable subjects at key phases during future epidemics may allow research institutions and the open science community to organize more quickly to tackle future epidemics. This effort is greatly facilitated by novel data sources such as the OpenCite and CrossRef databases, as previous sources of citation data remain locked behind the same granular publisher paywalls we sought to explore (Pentz 2001; Peroni et al. 2015). Through the combined application of these entirely free public databases, the open source analytical tools developed herein, and free-to-use computational resources such as Google Colaboratory, we provide a generalizable means for researchers across disparate disciplines to explore the role of open access science within their own fields.

1.4 Social Media Surveillance

Social media surveillance provides a promising method of rapid data **collection** for public health crisis response. Traditional public health data collection methods are bound by delays in the recognition of symptoms, the seeking and scheduling of care, and the processing time and sensitivities of reporting mechanisms. Further, medically derived public health data sources may be strongly biased by matters of healthcare access, societal stigma for a given condition, and the fear of legal repercussions in certain cases. These problems are particularly relevant when studying the behaviors of people who use opioids. Given these research challenges and the urgent state of the opioid epidemic in the United States, there is a vital need for alternative data sources that may overcome these hurdles and provide actionable insights to drive policy interventions.

Numerous social media platforms and crowd data sources such as Google Trends are available to researchers. Of all of these, Twitter has shown the greatest utility for public health data collection due to its large market penetration, its open API, the default public access of tweets, and the simplicity of analyzing its short text format. Twitter posts, or tweets, provide an intimate human sensor network of real-time conditions. Tweets are instantaneous, publicly available, and often possess fine-level geolocation data. Tweets may also come with some perception of anonymity, whether by numbers, institutional disinterest, or the deliberate use of pseudonyms. In addition, each tweet contains clean, organized data regarding its profile of origination and the conversational context in which it resides. Original tweet content may see serial sharing via retweets from interested parties, and each original tweet or retweet carries data detailing its chain of custody as well as its point of origination. However, there are inherent challenges to consider regarding the application of social media analyses. Language is temporally fluid and bound by regional and societal contexts. Further, as with traditional medical data sources, many served populations may be inherently evasive, whether due to societal stigma or fear of legal repercussions, though this effect may be lessened by perceptions of anonymity.

During past research efforts we developed and deployed the ChatterGrabber syndromic surveillance toolkit to leverage these advantages of Twitter analyses and to reduce barriers to entry for public health agencies seeking to conduct social media surveillance (J. T. Schlitt, Lewis, and Eubank 2015). ChatterGrabber has proven its merits numerous times with applications in message mapping in response to toxic algal blooms (Skiles 2017), as a surveillance mechanism tracking suicide (CDC 2015) and norovirus clusters (J. S. Schlitt and Lewis 2017), and as a data collection system to explore anti-vaccination semantic networks (Kang et al. 2017). ChatterGrabber has also been cited as one of the innovative Healthcare Systems for the 21st century (Qudrat-Ullah and Tsasis 2017). ChatterGrabber was designed with open science philosophies in mind, uses no paid code or API services, and is publicly available for non-commercial use with attribution under the Creative Commons **CC BY-NC** license. Links to its source code, reference materials, and preconfigured virtual machines may be found in ***Appendix B.***

Given its history of use, ChatterGrabber presented a natural choice for a data collection methodology for **Chapter 3** of this work. In this chapter we set out to provide a quantitative analysis to assess the effectiveness of strategies for **conveying** health-positive messages regarding the US opioid epidemic. We selected message virality and message retweeting rates as key metrics of audience engagement and spatiotemporal dissipation respectively. Using these metrics, we sought to identify optimal combinations of message thematic content, message source category, and message timing that would maximize virality with the tweeting public. Likewise, by exploring the drop off of retweets over time and distance, we sought optimal strategies for saturating an at-risk population during the course of a public health crisis such as a contaminated batch of high-potency illicit opioids circulating through a health district.

A pilot study was conducted to assess the feasibility of this approach, analyzing 5,403 tweets from the states of Virginia and Georgia following one such batch of contaminated drugs (J. Schlitt et al. 2019). As the pilot study yielded some promising signals, albeit limited by sample size, we conducted a full-scale study using a sample of 20,000 tweets from across the United States. This study incorporated additional source profile descriptors including urban location, non-negative messages, and non-suspected-bot profile status. This study showed that individuals drove the vast volume of messages whilst television news and other media profiles drove far greater views. Other findings included that law enforcement profiles showed the highest categorical virality for a single source-profile and message content pairing, that messages posted early in the day or week showed the best virality, and message non-negative tone had no consistent effect on virality. These results offer a promising glimpse into quantitatively derived best practices for opioid related messaging campaigns, practices that may extend across platforms. Further, this approach is trivially generalizable and its analytical pipeline could readily be applied to any public health crisis where there is an urgent need to **convey** health-positive messages on a larger scale.

1.5 Agent-Based Modeling

For crises with late-stage knowledge, agent-based models (ABM) provide the most flexible framework for incorporating detailed intervention strategies. Where more traditional, population-based epidemiological transmission models describe groups of humans in terms such as disease-state compartments or geographic patches, ABMs model the interactions of individual humans. By modeling each individual separately, ABMs have the potential to capture the heterogeneities of age, household structure, and roles in society. These resultant intricacies may then be used to guide the application of interventions in response to large-scale outbreaks of infectious disease. ABMs have been leveraged to describe everything from flu transmission within populated metropolises (K. R. Bisset et al. 2009) to the interplay between human agents and geographic concentrations of arbovirus-infected mosquitoes (Kuhlman et al. 2017). Agent-based models offer unique advantages in **connecting** and **conveying calibrated**, late-stage data, as the discrete nature of agents allows one to break the modeled population into ever more intricate subpopulations for comparison. Given a significantly larger and richer data set, a key challenge of ABMs lies in isolating subtle dynamics within the population for analysis to justify their added cost and development time.

Influenza is one such public health threat with a wealth of late-stage knowledge available, and as such the challenge of novel influenza strains to public health bears little introduction. One unique threat posed by influenza, however, is the effect it may have on defense readiness. As military and defense resources hold vital responsibilities in maintaining public order and function under existential threats, their protection is uniquely important with regards to pathogenic threats. Unfortunately, military base staff are also uniquely susceptible to threats such as pandemic influenza due to their frequency of travel, dense clustering of living quarters, and heavy use of common spaces. Fortunately, military populations also yield some unique advantages for epidemiological practice. Military staff are far more compelled to comply with the orderly collection of data and application of interventions. Where effective emergency isolation and compulsory vaccine administration may prove impossible within a civilian populace, such strong interventions are comparatively trivial to apply within the military.

To this end, in **Chapter 5** we employed the EpiFast ABM to explore a broad, multivariate sweep of intervention schemes in a plausible scenario where a brigade of soldiers returning from field training activities returned to base after being unknowingly exposed to pandemic influenza. We explored the combined application of sequestration, vaccination, and antiviral prophylaxis administration, each in turn hindered by realistic delays in implementation and efficacy. Likewise, we swept across variables describing the timing and effectiveness of each intervention, as stockpiles of pharmaceuticals may prove difficult to allocate en masse and in the midst of a pandemic.

Our results showed a strong benefit to rapid intervention and effective antivirals, as sequestration, however strong, served chiefly to buy time for the more lasting interventions. Likewise, we saw that no studied intervention scheme could prevent the majority of cases in such a highly connected population. This work highlights the importance of early detection and rapid response mechanisms for infrastructure critical to national security and the continuity of government. This study also provides a bleak but important view of the anticipated robustness of such systems against a truly novel threat. Further, this work demonstrates a full synthesis of the **C4 Response Model**, as the mass, underlying data **collected** regarding human mobility and co-occupancy networks has been **connected** with intricate models of influenza behaviors and **calibrated** to historic outbreak curves to **convey** actionable policy findings regarding the threat of novel pathogens to defense readiness.

Chapter 2

The State of Open Access in Outbreak Science

Attribution

The following chapter is based upon the manuscript in prep: *Schlitt, J., Lewis, B., Eubank, S. (2019). The State of Open Access in Outbreak Science.*

2.1 Abstract

Academic competition drives the compartmentalization of knowledge during critical periods of pandemic outbreaks. Fortunately, a growing open science community has arisen to combat this trend, democratizing access to knowledge, data, and code. In order to explore this trend, we sought to document the role of open access publications in outbreak response, and we propose that open access science is driving positive changes in epidemiology. We analyzed WHO-identified pandemic pathogens with outbreaks between 2000 and 2015 which were either new to science or which had limited recent human case history. From this list of pathogens we searched PubMed for articles published within the first two years of each outbreak and selected the pathogens that saw at least 50 cited publications with one or more references. This resulted in a final analysis set of 2,264 articles describing Severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), Chikungunya (CHIKVD), Ebola hemorrhagic fever, Bubonic plague, and Zika virus disease (ZIKVD). Each publication's citation and access rights data was queried via the OpenCitations COCI database. A second set of 10,000 randomly sampled PubMed articles published between 2000 and 2018 was similarly collected for comparison. Open access outbreak-response papers saw near-significant publishing lead with articles being published slightly earlier than closed access papers across epidemics *(p=0.06)*. Closed access outbreak response papers were found to cite references published closer to both their date of publication as well as the start of their respective epidemics. However, open access outbreak response papers and references were significantly greater in number, and increased from 40% of all references to 70% within the study window. Further, we observed that outbreak response papers were composed of significantly greater percentages of open science articles than PubMed biomedical publications at large. As open access journal articles grow in impact, further support of open access science may drive faster and better-informed outbreak response.

2.2 Introduction

Emerging infectious disease remains one of the foremost threats to global public health. As human behaviors continue to modify the environment and destroy global biodiversity, the rate of zoonotic spillover events can only be expected to accelerate (Blancou et al. 2005). Before the emergence of open science culture, traditional outbreak responses were often stymied by the institutional compartmentalization of knowledge. Fortunately, there is a growing open science community, and exciting efforts during the Zika virus disease (ZIKVD) and Ebola hemorrhagic fever (EHF) epidemics show that a promising change is occurring (Rodriguez et al. 2017; Rivers et al. 2014; Majumder et al. 2014; "Epidemics | The RAPIDD Ebola Forecasting Challenge | ScienceDirect.com" n.d.; Chretien, Rivers, and Johansson 2016). The institutional compartmentalization of early outbreak knowledge poses an inherent threat to the public. Linchpin findings regarding pathogens novel to science or otherwise exhibiting novel behaviors often present turning points in the effectiveness of institutional responses to pandemic threats. Cases such as this are abundant and include discoveries such as the role of West African funeral practices in the 2014 Ebola epidemic (Faye et al. 2015), the role of wet markets in the 2013 H7N9 epidemic (Y. Chen et al. 2013), and the role of camels in the 2013 Middle East Respiratory Syndrome (MERS) epidemic (Azhar et al. 2014).

Open access science presents an exciting opportunity to bring in a broader community of researchers earlier within future epidemics by disseminating key findings and data sources publicly. To explore the extent to which said opportunities have been realized in practice, we sought to document the role of open access science in outbreak response. We propose that open access publications are driving a marked improvement in the dissemination of knowledge and data in ways distinct from the greater bodies of biomedical research. In order to assess this theory, we compared the trends and patterns with which open and closed access outbreak response papers were published or referenced in the midst of epidemic response efforts relative to each other as well as a broader sample of biomedical research publications. Finally, we explored trends in citation rates over time as a function of both open and closed access status as well as broad research methodologies applied within each study. We propose that by studying such trends, we might identify effective windows of utility for open access publications to disseminate knowledge as well as productive trends by which this shift has accelerated the academic community’s response to novel pandemic threats.

2.3 Methods

2.3.1 Glossary

Citations: derivative works which cited the *primary papers.*

Epidemic time: days from *time zero (t_0),* or the first known cases of a given outbreak.

Notable outbreak: one cluster of human cases which drew significant press or academic attention and which saw significantly more total infections than would be considered seasonal.

Primary Papers: papers printed between zero and 730 days in epidemic time from the emergence of an outbreak with pandemic potential, whose abstracts possessed keywords related to the given outbreak, and which matched study inclusion and exclusion criteria.

References: previous works cited by the *primary papers.*

Time Zero (t_0): the time of first known case(s) of a given outbreak. Time zero for each epidemic was determined by the earliest known primary literature date tied to the suspected transmission or detection of cases from a given outbreak.

2.3.2 Study Parameters

Diseases were selected from the WHO’s list of pandemic pathogens and excluded if they did not drive a notable outbreak between 1/1/2000 and 12/31/2015 (“WHO | Disease Outbreaks” 2019). For Ebola virus and other pathogens where multiple outbreaks met these criteria, the largest recorded outbreak for a given illness within the study period was selected. Diseases were broken down into historical categories selecting for novelty which included *newly recognized illnesses* such as MERS or Severe Acute Respiratory Syndrome (SARS) and illnesses with *limited prior history* such as Ebola hemorrhagic fever (EHF) or Zika virus disease (ZIKVD).

Table 2.1: Disease categories

Disease Category	Category Description	Members
Newly recognized illnesses	Illness with no known human cases prior to the studied outbreak	MERS, SARS
Limited prior history illnesses	Illness with few known human cases, no prior outbreaks of comparable scale, or which otherwise exhibits significant novel epidemiological traits relative to prior epidemics	CHIKVD, Ebola, Marburg, Monkeypox, Nipah, Plague, ZIKVD

2.3.3 Data Preparation

Primary paper data was collected via PubMed queries submitted with the Entrez function of BioPython within a Google Colaboratory notebook (pubmeddev n.d.; Cock et al. 2009). Diseases were queried by their common name, scientific name, species, acronyms, and early names such as Coronavirus Erasmus Medical Center for MERS-CoV. Queries for acronyms such as *"NIV"* were omitted if they corresponded to other, unrelated conditions. For each disease, PubMed queries pulled data for articles containing the diseases queried names from the five years before and after the year of the outbreak. This was done to account for discrepancies in PubMed's date assignment across publication statuses, as COCI database citation dates were found to more consistently reflect the initial, journal publication date of each article. For comparison, an external set of 10,000 random PubMed articles was collected. It included articles from the year 2000 to present, randomly selected by their PubMed ID. For papers that lacked a digital object identifier (DOI), dates were taken from the first publication event recorded for a given paper. Primary papers with DOI information were queried by DOI against the OpenCitations COCI database for references, citations, and metadata (Peroni et al. 2015). From these COCI queries, journal publication dates, citation data, reference data, journal information, publication dates, and open access web links were collected into a database.

2.3.4 Selection Criteria

Primary articles were included only if they were published within the first two years following the start of their associated epidemic and if they matched the criteria of having one or more references and one or more citations available via OpenCitations. Primary Papers and comparison papers were excluded if they lacked a DOI, if they had a DOI but were not captured within the COCI database, if they were identified as books or book chapters, if they did not have at least one reference and one citation recorded, or if they did not contain one or more of the queried disease names within their abstract. Primary papers were further excluded if they were not published between days zero and 730 from the time zero of their respective epidemic. Whole outbreaks were excluded if they had fewer than 50 papers matching the inclusion criteria within this study window. References and citations were excluded if they or their referencing primary paper did not have at least month-level publication date resolution. This resulted in a final analysis set of 2,264 articles describing epidemics of Severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), Chikungunya (CHIKVD), Ebola hemorrhagic fever, Bubonic plague, and Zika virus disease (ZIKVD) following the exclusion of Marburg hemorrhagic fever, Monkeypox, and Nipah virus disease. For each primary and comparison paper, the title, abstract, DOI, date published, number of citations, number of references, keywords, qualifiers, descriptors, grant IDs, and funding agencies were also recorded for analysis. The names of qualifiers and descriptors describing broad and fine methodologies applied within a given paper were extracted from PubMed's mesh heading attribute.

Reference and citation data tables were constructed by querying OpenCite for the DOIs of references and cited papers for each primary paper. For each reference or citation made, the journal, open access web domain, open access status, time difference from the primary paper, and epidemic time of publication were recorded alongside the corresponding values from the primary paper.

Table 2.2: Disease parameters

Disease	Queried Names	Outbreak Name	Outbreak t_0
CHIKVD	"chikungunya virus" OR " CHIKV "	2013 Chikungunya Outbreak	12/22/2013 (Henry et al. 2017)
Ebola	"ebola virus" OR "ebola viral"	2014 West African Ebola Epidemic	12/26/2013 ("WHO | Origins of the 2014 Ebola Epidemic" 2015)
Marburg	"marburg virus" OR "ravn virus" OR "marburg hemorrhagic fever" OR " MARV " OR " RAVN "	2004 Angola Marburg Outbreak	10/02/2004 (Jeffs et al. 2007)
MERS	" mers virus" OR "mers-cov" OR "middle east respiratory syndrome" or "HCoV-EMC" or "Human Coronavirus-Erasmus Medical Center"	2013 Middle Eastern Respiratory Syndrome Outbreak	09/23/2012 ("WHO | Novel Coronavirus Infection in the United Kingdom" 2015)
Monkeypox	"monkeypox virus"	2003 Monkeypox Outbreak	05/15/2003 (Centers for Disease Control and Prevention (CDC) 2003)
Nipah	"nipah virus"	2001 Siliguri Nipah Outbreak	01/31/2001 (Chadha et al. 2006)
Plague	"bubonic plague" OR "pneumonic plague" OR "yersinia pestis"	2014 Madagascar Plague Outbreak	11/04/2014 ("WHO | Plague – Madagascar" 2015)
SARS	"sars virus" OR "severe acute respiratory syndrome"	2003 Asian SARS Outbreak	11/16/2002 ("CDC SARS Response Timeline | About | CDC" 2018)
ZIKVD	"zika virus" OR " ZIKV "	2015 Zika Virus Epidemic	03/29/2015 ("WHO | The History of Zika Virus" 2017)

2.3.5 Temporal Analyses

The temporal distribution of open and closed access papers over epidemic time following the start of each epidemic was visualized and compared via cumulative histogram and kernel density estimation (KDE) plots ***(figure 2.1)***. Analyses were conducted to compare the delay with which open and closed access articles published 25% and 50% of their outbreak response papers as well as the mean publication delay following the start of each epidemic. This metric was chosen to illuminate potential differences in the rapidity of emergent responses and whether open science journals were able to publish earlier following emergent threats.

The temporal distribution of references relative to both the publication date of the primary paper referencing them as well as the start of their respective epidemic was visualized ***(figures 2.2a, 2.2b, 2.2c, and 2.2d).*** References were excluded if they were published more than 1825 days from their referencing primary article or the start of their associated epidemic respectively. Using the same five-year subsets, the percentages of references published within one year prior to their citing primary paper and the ratios of references from the year before each outbreak and the year following each outbreak were derived. Finally, Two-Sample t-Tests were applied with pooled variance to evaluate whether open access references were published closer to the start of their given epidemic or primary paper. These metrics were chosen to explore whether open access references provided relatively newer data during the span of an academic outbreak response.

The percent and total counts of open and closed access citations by year were plotted with respect to the year of their citing primary publication ***(figure 2.3a)***. Likewise, the numbers of open and closed access citations were plotted with respect to the year of their own publication ***(figure 2.3b)***. In each of these plots the equivalent data from the comparator set was rendered in order to explore whether outbreak response publications differed from PubMed articles overall in their utilization of open science reference materials.

The mean number of articles citing open and closed access primary papers were plotted by time of primary paper publication in 30-day blocks starting from the beginning of each epidemic ***(figure 2.4)***. This metric was selected to explore whether primary papers showed consistent windows of high citation utility whereby earlier work could generate crucial findings that informed the later stages of academic response efforts. Finally, the top five journals most referenced by the primary papers and the top five primary paper journals by number of references made, respectively, were identified and plotted by their total number of references made by year of reference publication ***(figures 2.5a and 2.5b)***. This was done to explore whether notable open or closed access journals saw periods of high utility and whether journals publishing high numbers of captured articles were citing more recent open access reference materials.

2.3.6 Network Analysis

Due to the large number of unique research methodology qualifiers present within the primary publications, network methods were applied to aggregate closely related qualifiers into communities for analysis. A weighted, undirected graph was constructed in NetworkX using the complete set of primary paper publication qualifier data (Hagberg, Swart, and S Chult 2008). Each of these curated, categorical research disciplines was represented as a singular node. Edges were placed between nodes that were used within the same publications and weight was assigned as the total number of publications containing a given pair of qualifiers. The NetworkX implementation of the Clauset-Newman-Moore Greedy Modularity Maximization Algorithm was applied to separate the qualifier graph into communities (Hagberg, Swart, and S Chult 2008; Clauset, Newman, and Moore 2004). This resulted in one connected component with 33 distinct communities. However, due to the high connectivity of the graph, 22 of these communities represented singular, highly applied methods with high degree and cumulative edge weight. The five communities that yielded the greatest number of primary paper citations were selected for further analysis. The proportion of primary paper citations by epidemic time of primary paper publication was plotted for the five most-cited qualifier communities and subdivided by open access status ***(figure 2.6)***. The graph used to generate these communities has been included as ***Appendix A*** for further reference.

2.4 Results:

2.4.1 Data Description

The initial PubMed data pull resulted in 5,963 unique articles representing outbreak journal publications and excluding books. This resulted in a reduced set of 2,291 articles with 38.4% passing the exclusion criteria. Nipah, Marburg, and Monkeypox were excluded from further analysis as each had fewer than 50 acceptable articles. This left a final selection of six diseases and 2,264 articles for further analysis. A general trend was apparent of early 2000s primary articles being approximately 54% open access before increasing to the 70–80% range for each outbreak after 2010. This stands in stark contrast to the comparator sample, of which only 48% of articles were published as open access. Comparator articles were found to be somewhat less likely to match the exclusion criteria than the joined set of disease articles, with only 31% of articles being accepted. Comparator mean reference and citation counts were close to those of disease articles, utilizing 2.2 more references per article yet receiving 4.1 fewer citations per article.

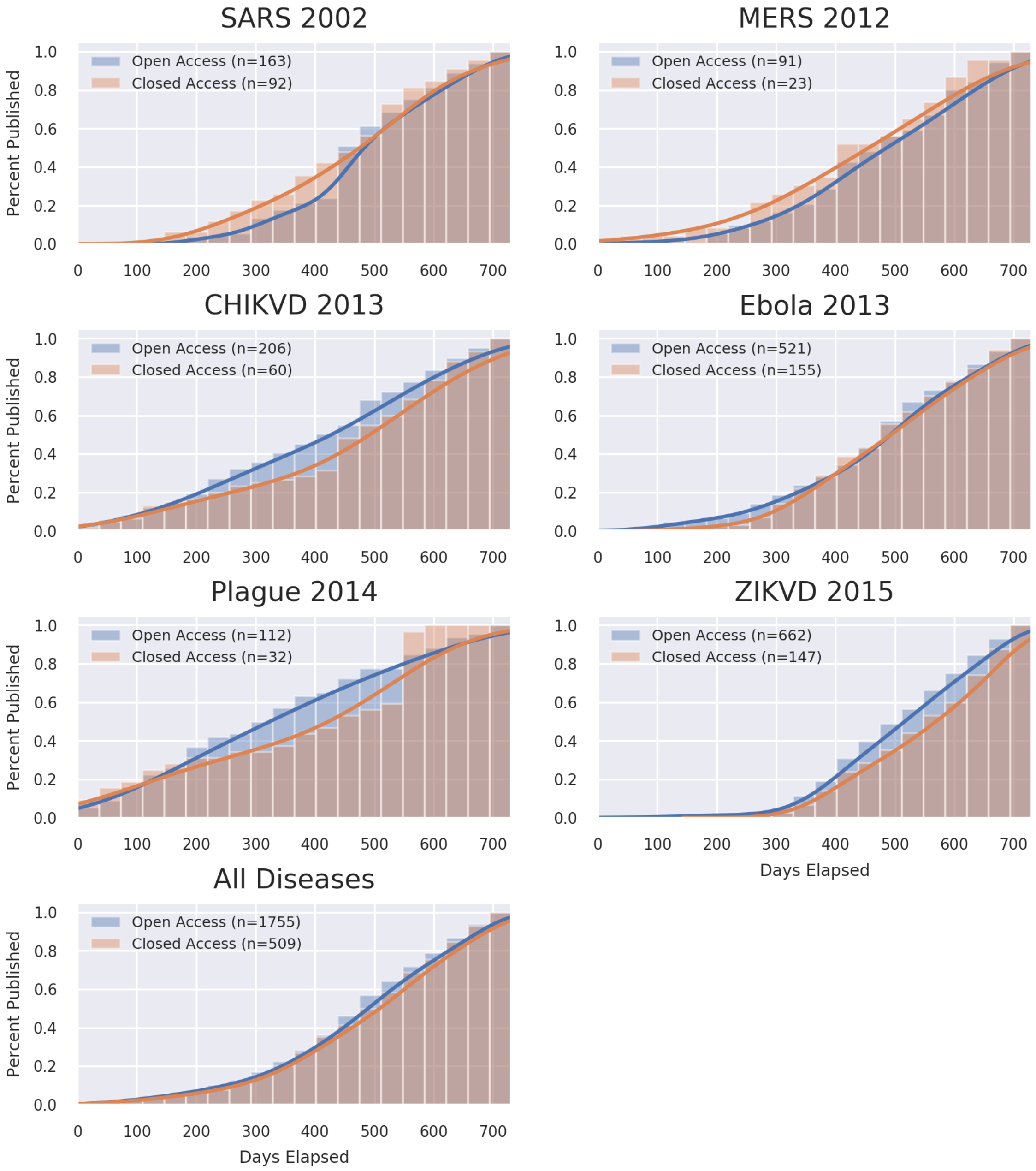
Table 2.3: Collection results by date of index case

Outbreak	PubMed Articles Pulled	Filtered Articles	Percent Open Access	Mean References per Article	Mean Citations per Article
Nipah 2001*	49	7	57%	16.7 (3.6)	34.7 (8.7)
SARS 2002	1,392	255	64%	15.6 (0.9)	34.5 (4.6)
Monkeypox 2003*	51	12	50%	11.8 (2.4)	31.5 (12.5)
Marburg 2004*	46	8	62%	28.0 (11.4)	40.0 (23.3)
MERS 2012	330	114	80%	28.6 (2.5)	41.5 (6.7)
CHIKVD 2013	494	266	77%	36.0 (1.8)	15.0 (1.6)
Ebola 2013	1,534	676	77%	28.7 (1.1)	15.2 (1.4)
Plague 2014	287	144	78%	43.9 (2.4)	5.5 (0.5)
ZIKVD 2015	1,780	809	82%	28.8 (1.1)	24.0 (1.7)
All Diseases*	5,963	2,291	77%	28.9 (0.6)	21.3 (1.0)
Comparator	10,000	3,098	48%	31.1 (0.5)	17.2 (0.9)

Nipah, Monkeypox, and Marburg were excluded from the following analyses

2.4.2 Similar epidemic timing of publications by access type

Figure 2.1: Cumulative publications vs. time from outbreak start



The number of primary publications by disease and open access status was plotted cumulatively as a function of days from the first known cases for each epidemic. While some heterogeneity appeared by pathogen, publications followed an approximately sigmoidal distribution with a long left tail during the early stages of outbreak response. For most diseases there was a notable surge in open and closed access publication rates between 400 and 500 days. The 2002 SARS and 2012 MERS epidemics showed the only cases in which closed access publications reached 25% and 50% of their total counts before open access publications. In the combination set of all diseases, open access papers reached the 25% and 50% benchmarks only 3 and 4 percentage points quicker than closed access papers. The CHIKVD response effort showed the most notable advantage for open access publications, with primary publications hitting the same 25% and 50% benchmarks in 66% and 90% of the time required by closed access publications. Open access publications remained more prolific than closed access across each epidemic. Open access publications also provided some of the only publication data available during the early stages of the ZIKVD and Ebola epidemics. However, the combined set of all diseases showed little difference in the proportionate rates of outbreak publications.

Table 2.4: Days until benchmarks met

	Days Elapsed Before 25% of Primary Articles Published			Days elapsed before 50% of Primary Articles Published		
Disease	Open Access 25%	Closed Access 25%	Relative Rate	Open Access	Closed Access	Relative Rate
SARS 2002	450	358	1.25	473	491	0.96
MERS 2012	381	340	1.12	492	506	0.97
CHIKVD 2013	233	352	0.66	437	486	0.90
Ebola 2013	379	379	1.0	499	499	1.0
Plague 2014	178	156	1.14	335	462	0.73
ZIKVD 2015	407	444	0.92	517	571	0.91
All Diseases	379	391	0.97	494	513	0.96

In order to test the hypothesis that open science papers are published earlier within an outbreak, we conducted a Two-Sample t-Test with pooled standard deviations comparing the epidemic times in which open and closed access primary papers were published. We sought to disprove the null hypothesis that the mean time elapsed for open access primary paper publications was greater than or equal to the mean time elapsed for closed access primary paper publications, written as such where *e* represents the time elapsed from the index cases of a given epidemic at the time of primary article publication:

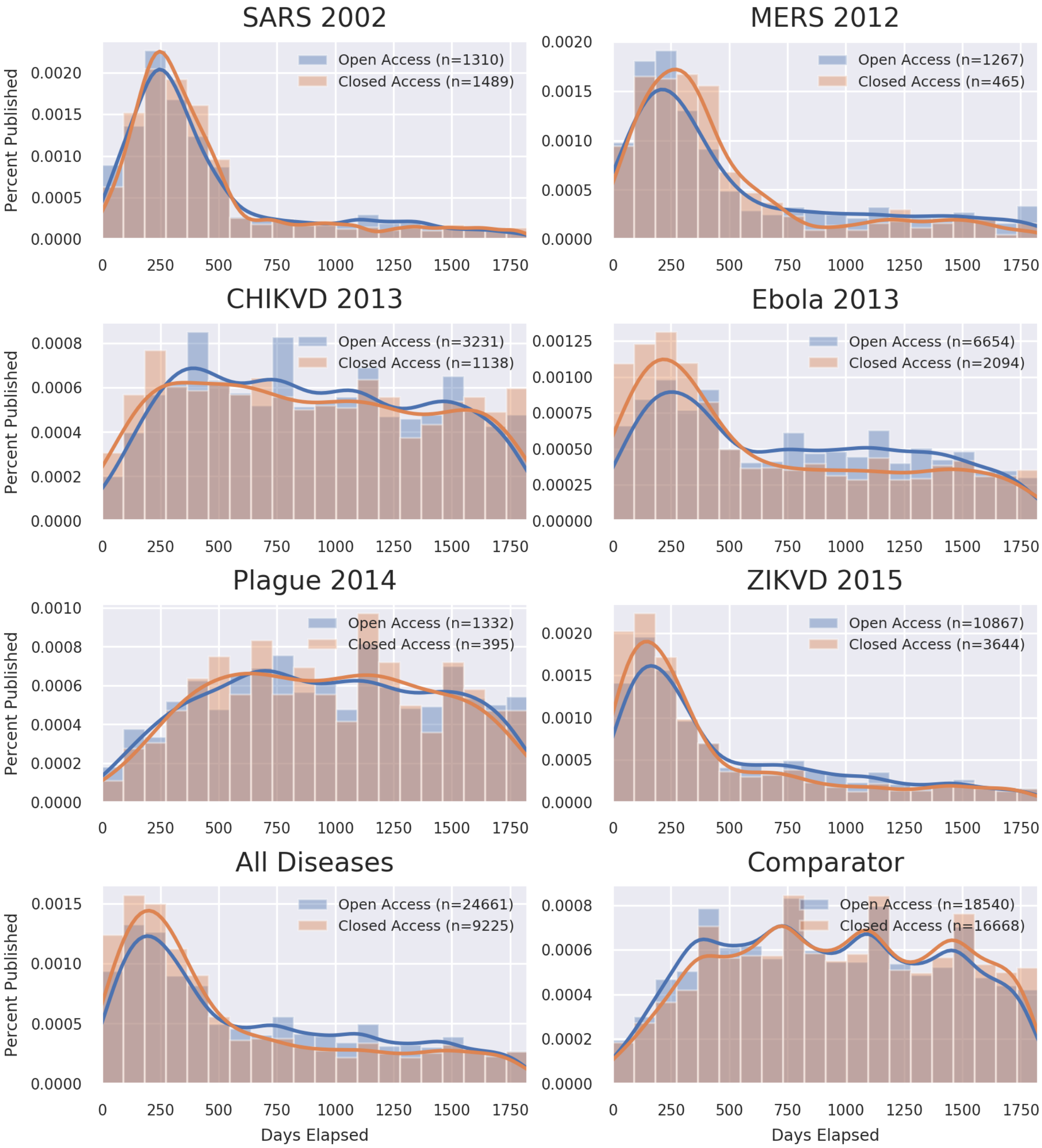
$$H_0 \colon \bar{e}_{open\ access\ primary\ publications} \geq \bar{e}_{closed\ access\ primary\ publications}$$

$$H_A \colon \bar{e}_{open\ access\ primary\ publications} < \bar{e}_{closed\ access\ primary\ publications}$$

For the selected outbreaks, there were 1,755 total open access primary publications, a mean delay of 474 days, and a delay standard deviation of 160 days. Likewise, there were 509 closed access primary publications with a mean delay of 490 days and a standard deviation of 158 days. This resulted in a test statistic of -1.9 with a p-value of 0.06, and thereby narrowly failed to reject the null hypothesis that open access primary papers were published with equal or greater delays than closed access papers.

2.4.3 Greater recency of closed access references relative to open access references

Figure 2.2a: Days between primary paper publication and reference paper publication by reference open access status



In order to determine which publication types used the most recent knowledge during outbreaks, we analyzed the time differences between primary papers and their references ***(figure 2.2a, table 2.5)***. We also explored the time differences between reference papers and the start of each outbreak by open access status ***(figure 2.2c, table 2.6)***. Time differences of greater than five years were excluded in order to select for recent medical knowledge as both Plague and ZIKVD notably had decades of prior references available. In each case, primary papers' references were categorized by the references' open access statuses, and the total numbers of references per given time period by reference type were plotted via normed histograms and KDE. Significantly more open access publications were referenced than closed access publications for every epidemic following SARS, with each disease's primary publications showing an approximately 2.7:1 preference for open access references with 73% of references coming from open access sources. There appears to be a general correlation between the novel, human threat posed by each illness to the shape of their publication distribution. SARS and MERS being completely new to science illnesses saw the bulk of their primary paper references being published less than 500 days before their citing primary paper. Of the remaining illnesses with limited human history, ZIKVD, Ebola, CHIKVD, and Plague each saw an increasing age of references in both their open access and closed access papers, drawing significantly fewer of their publications from the year prior ***(table 2.5)***. We also note that for every outbreak, with the exception of the 2014 Madagascar Plague outbreak, closed access references of less than one year of age at the time of citation represented a greater share of the total closed access references made than open access publications. This contrasted expectations from the earlier findings of significantly higher volume and marginally earlier open access publication epidemic times observed in ***figure 2.1.***

Inclusion of the comparator set showed a similar pattern to Plague, and CHIKVD, where a significantly lower proportion of references were less than a year old at time of comparator paper publication. The comparator set also showed a lower proportion of open access referencing than the primary papers, with 18,540 open access and 16,665 closed access references utilized.

In order to test the hypothesis that newer open science references are utilized earlier within an outbreak, we conducted a Two-Sample t-Test with pooled standard deviations comparing the age of references in days at time of primary paper publication. We tested this hypothesis by seeking to disprove the null hypothesis that the mean reference age of open access reference papers is greater than or equal to mean reference age of closed access reference papers. In this case, *a* represented the mean age of references at the time of primary paper publication. As in the ***figure 2.2a,*** papers published five or more years prior to the start of a given outbreak were excluded:

$$H_0 : \bar{a}_{open\ access\ ref\,erences} \geq \bar{a}_{closed\ access\ ref\,erences}$$

$$H_A : \bar{a}_{open\ access\ ref\,erences} < \bar{a}_{closed\ access\ ref\,erences}$$

This exclusion yielded 24,661 open access references with a mean age of 655 days at time of primary paper publication and a delay standard deviation of 513 days. Closed access references totaled 9,225 in number with a mean delay of 565 days and a standard deviation of 505 days. This resulted in a test statistic of 14.5 with a p-value of 2.4 e^{-47}. From this, not only do we fail to reject the null hypothesis, but also we note that open access references by volume were significantly older than closed access at the time of their primary paper's publication.

Figure 2.2b: Percent of references published within a year of their citing primary paper by reference open access status

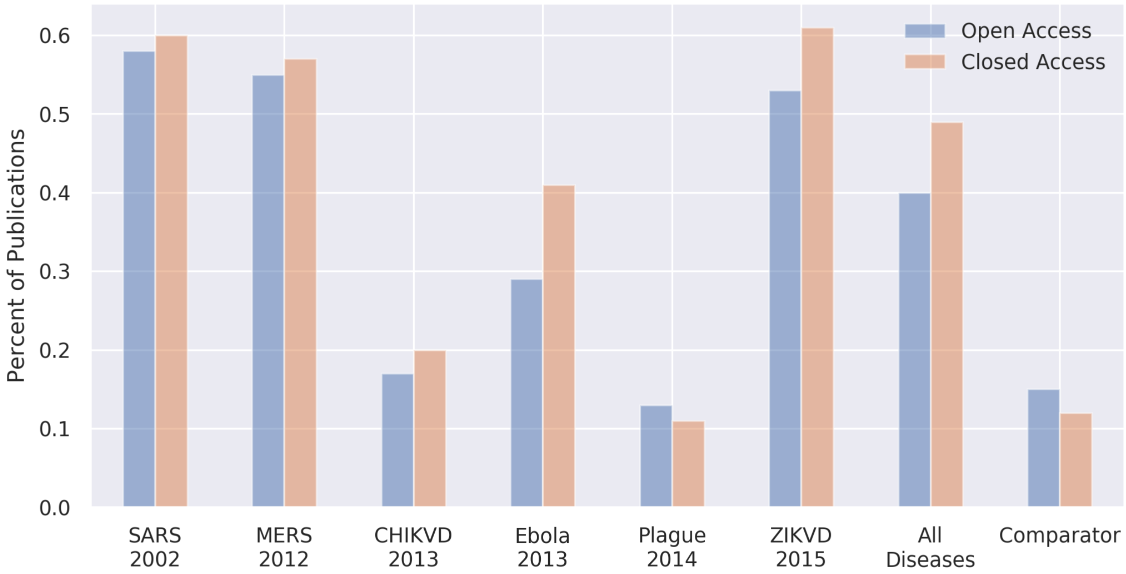
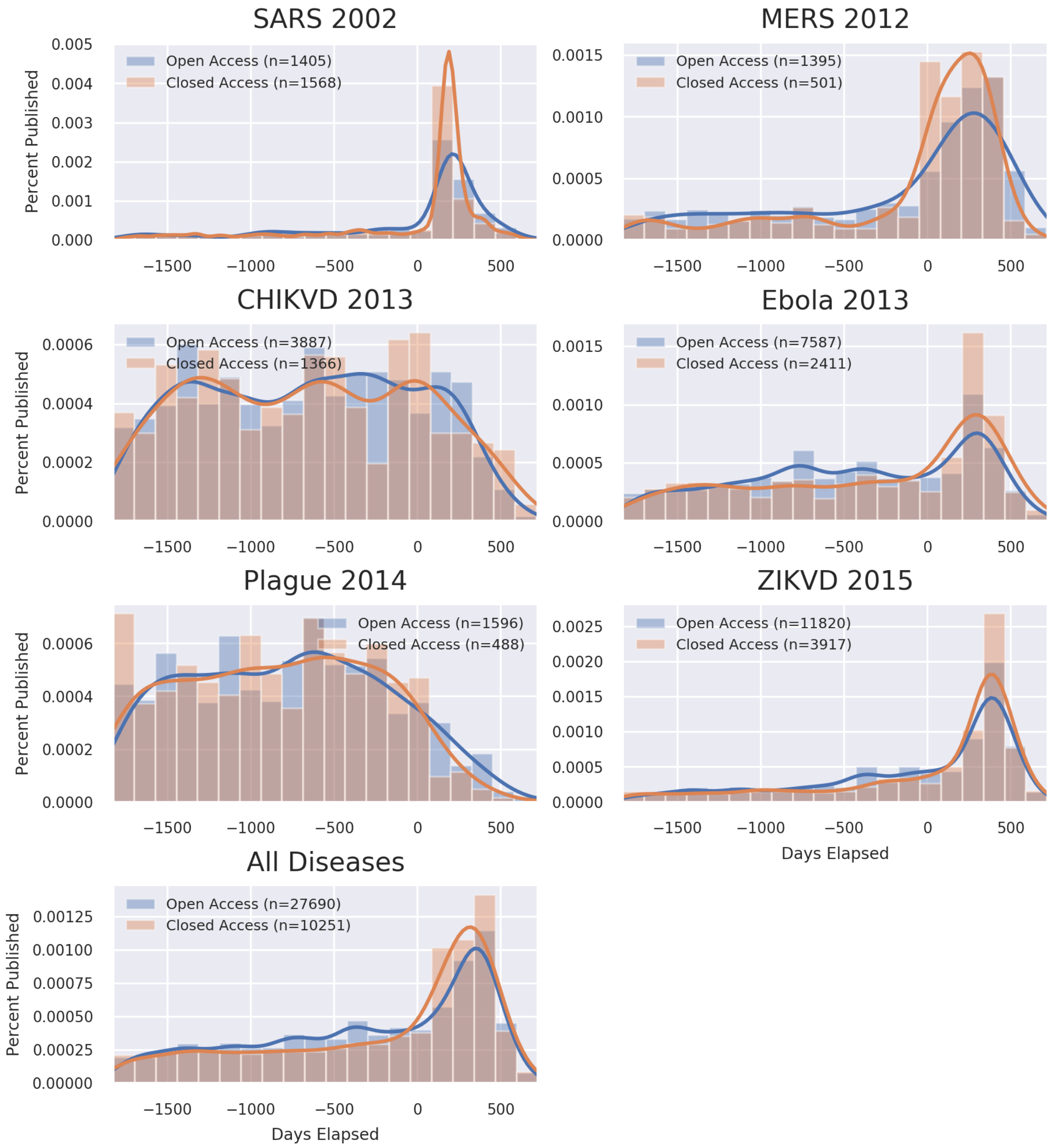


Table 2.5: Percent of references published within a year of their citing primary paper by reference open access status

	SARS 2002	MERS 2012	CHIKVD 2013	Ebola 2013	Plague 2014	ZIKVD 2015	All Diseases	Comparator
Open Access	0.58	0.55	0.17	0.29	0.13	0.53	0.40	0.15
Closed Access	0.60	0.57	0.20	0.41	0.11	0.61	0.49	0.12
Relative Ratio	0.97	0.96	0.85	0.71	1.18	0.87	0.82	1.25

Figure 2.2c: Days from reference paper publication to outbreak start by reference open access status



A similar pattern arises when comparing the time difference between reference publications and the start of each epidemic, as seen in ***figure 2.2c***. Here again references made five or more years from the start of each epidemic were excluded, as were references published two or more years after given the two-year limit for primary paper inclusion. As before, we see a notable trend where the academic responses to less novel or threatening pathogens drew from more dated literature sources. The ratios of knowledge drawn from the year after the outbreak to the year before ***(table 2.6)*** followed a similar pattern to the age of publication, where both novel pathogen responses heavily favored research published after the start of each epidemic, and the ZIKVD, Ebola, CHIKVD, and Plague response efforts showed progressively decreasing utilization of post-outbreak research. Likewise, the 2014 Madagascar Plague outbreak proved the only case where the open access references utilized were relatively newer. The comparator data set was excluded from this analysis, ***figure 2.2c, figure 2.2d,*** and ***table 2.6*** as there was no meaningful or consistent way to apply an epidemic start.

Figure 2.2d: Relative referencing rate of publications by reference open access status from the year after the epidemic to the year before

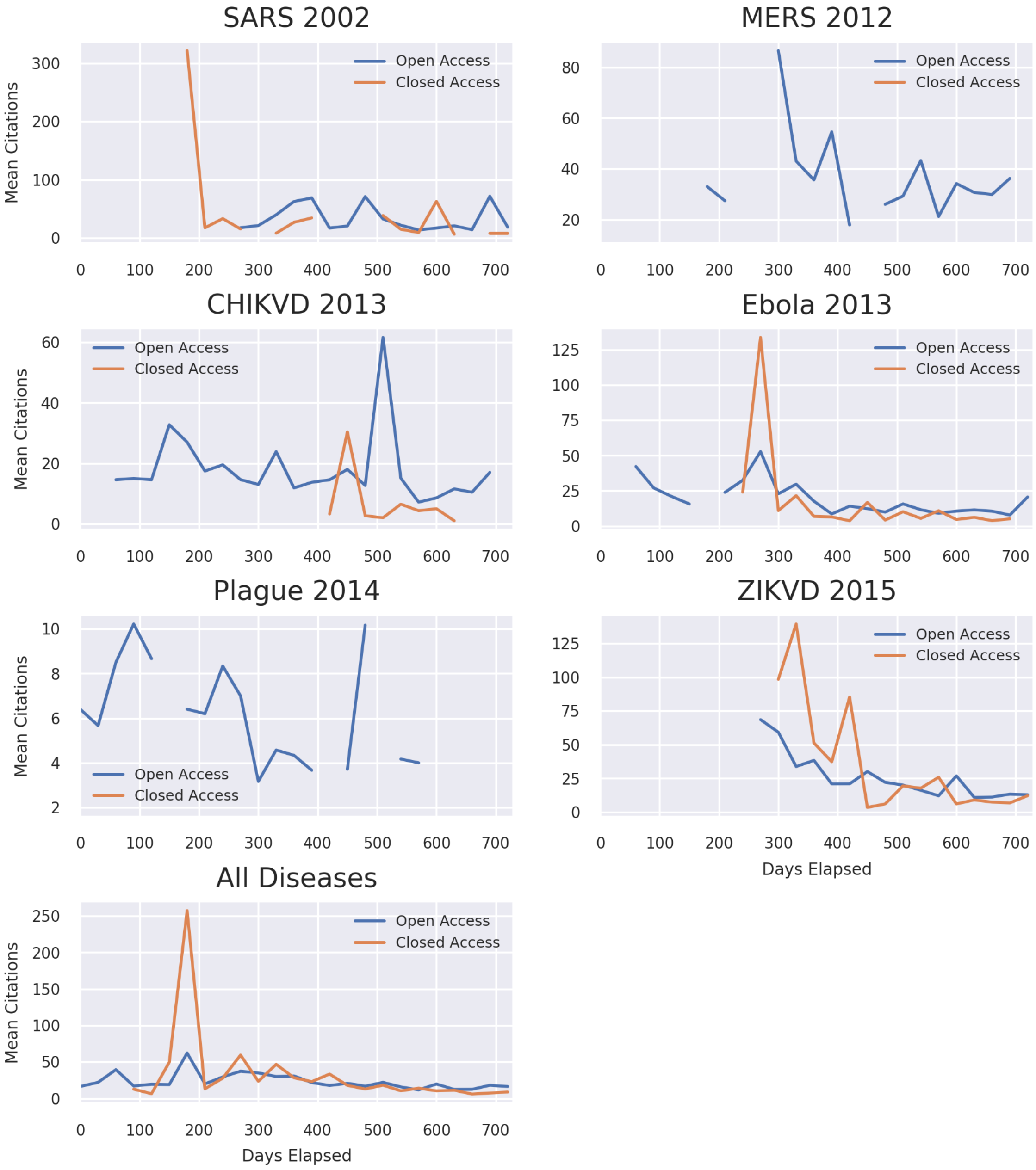


Table 6: Relative referencing rate of publications by reference open access status from one year after the epidemic to one year prior

	SARS 2002	MERS 2012	CHIKVD 2013	Ebola 2013	Plague 2014	ZIKVD 2015	All Diseases
Open Access	5.63	3.74	1.02	1.85	0.61	1.77	1.81
Closed Access	9.58	8.34	1.01	2.68	0.34	2.59	3.00
Relative Rate	0.59	0.45	1.01	0.69	1.79	0.68	0.60

In parallel to the previous comparison of reference age at time of primary paper publication, we conducted a Two-Sample t-Test with pooled standard deviations comparing the time of reference papers' publication relative to the start of their given epidemic. If newer open access references were being utilized relative to closed access, we might also expect to see a greater proportion of open access references being made after the start of each respective epidemic as well as higher mean epidemic times. We tested this hypothesis by seeking to disprove the null hypothesis that the mean epidemic time elapsed of open access reference papers is less than or equal to the mean epidemic time elapsed of closed access reference papers. In this case, *e* represents the epidemic time of references at the reference publication. As in the ***figure 2.2d*** papers published five or more years prior to the start of a given outbreak were excluded:

$$H_0: \bar{e}_{open\ access\ ref\,erences} \leq \bar{e}_{closed\ access\ ref\,erences}$$

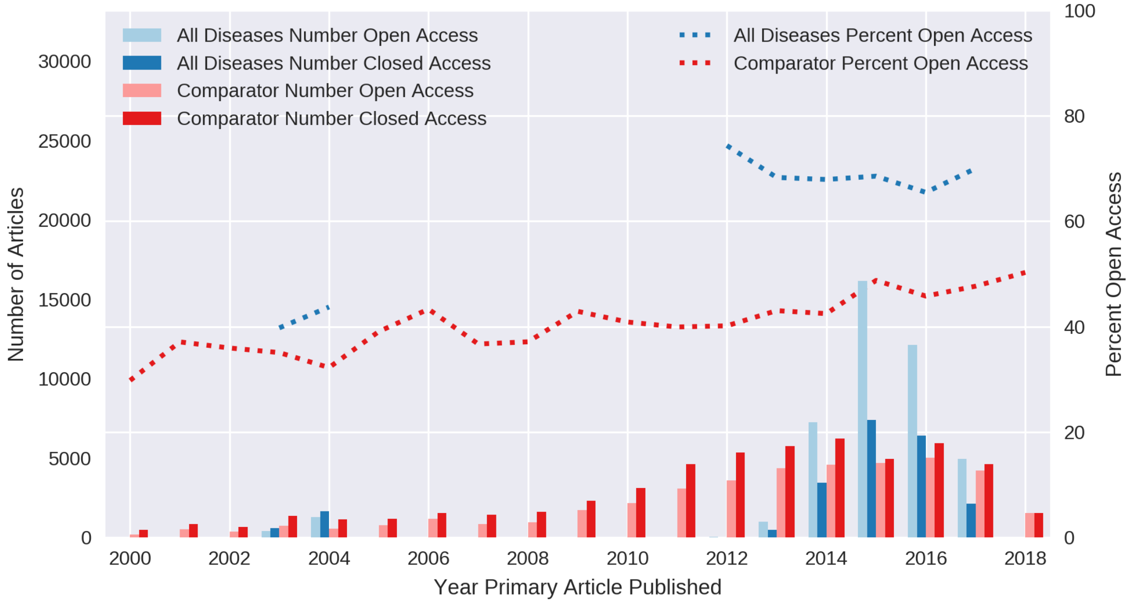
$$H_A: \bar{e}_{open\ access\ ref\,erences} > \bar{e}_{closed\ access\ ref\,erences}$$

This exclusion yielded 27,690 open access references with a mean epidemic time of -305 and a delay standard deviation of 683 days. The 10,251 closed access references showed a mean epidemic time of -216 days and a standard deviation of 683 days. This resulted in a test statistic of -11.3 with a p-value of 2.1 e^{-29}. This data suggests that not only were closed access references published closer in time to their citing primary papers, but that they were also published closer in time to the start of epidemics. Still, open access papers were referenced in far greater number.

2.4.4 Increasing utilization of open access references over time

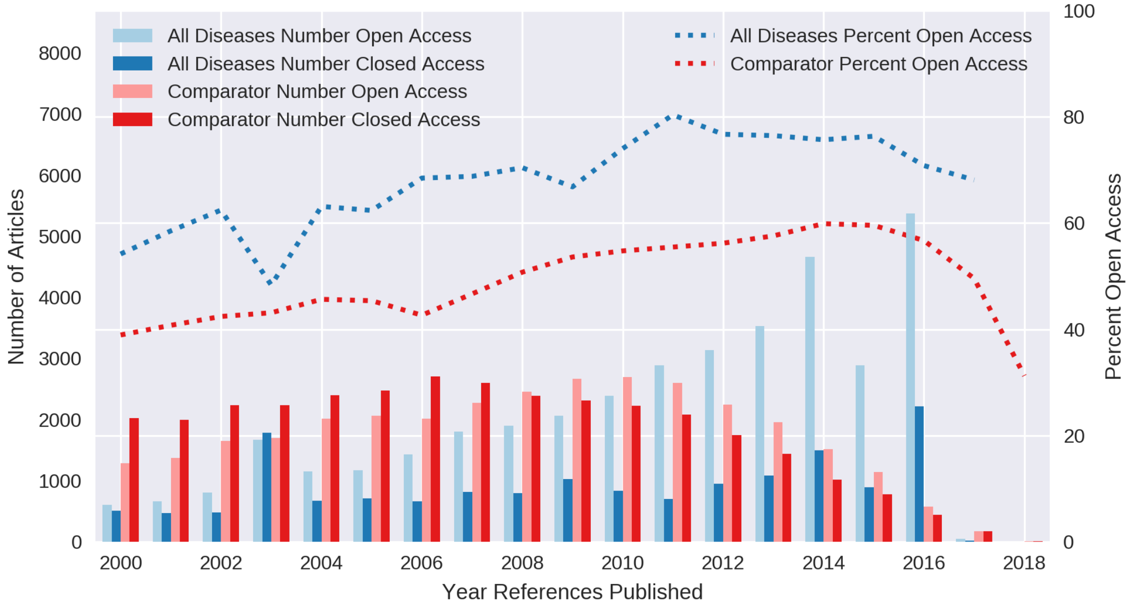
In ***figure 2.3a*** we compared the totality of all open and closed access references made by the year of the citing primary papers and comparator papers. This figure shows the notable gap in the disease data between 2004 and 2012 due to the lack of outbreaks matching the study inclusion criteria during that period. However, we still observe a marked increase in the percentage of open access reference utilization between 2004 and 2013, increasing from 44% to 74%. This is followed by an apparent leveling around 68% open access reference utilization between 2012 and 2017. The comparator set shows a similar trend of increasing open access citation utilization over time, starting at 30% in 2000 and increasing to 50% by 2018. While the gap in the data makes a firm conclusion difficult, it appears the utilization of open access reference material by outbreak response publications increased significantly in the years following the SARS epidemic before settling to a level which the comparator set has not yet reached. Likewise, while open access utilization in primary publications increased 30% between the 2002 SARS epidemic and the 2015 ZIKVD epidemic, an increase of only 20% was seen by the comparator set within the study window.

Fig 2.3a: Open access and closed access reference counts by year of primary paper publication



In ***figure 2.3b*** we compare the totality of all open and closed access references for disease papers and comparator papers once more. However, this time we compare each by the year in which their respective reference papers were made. Again we see a steady increase over time in utilization of open access disease papers. For each given year of reference publication, open access articles grew successively more likely to be referenced by outbreak response publications. Notably, in spite of a dip in 2003 during the SARS outbreak, open access reference articles were more likely to be cited on any given year. In contrast, closed access comparator articles saw greater citation rates than open access all the way until 2008 when open access secured a lead. Barring a sudden drop after 2016, comparator articles also saw consistent gains in the percentage of open access articles being cited for each given year, albeit at levels lower than those for outbreak related primary publications.

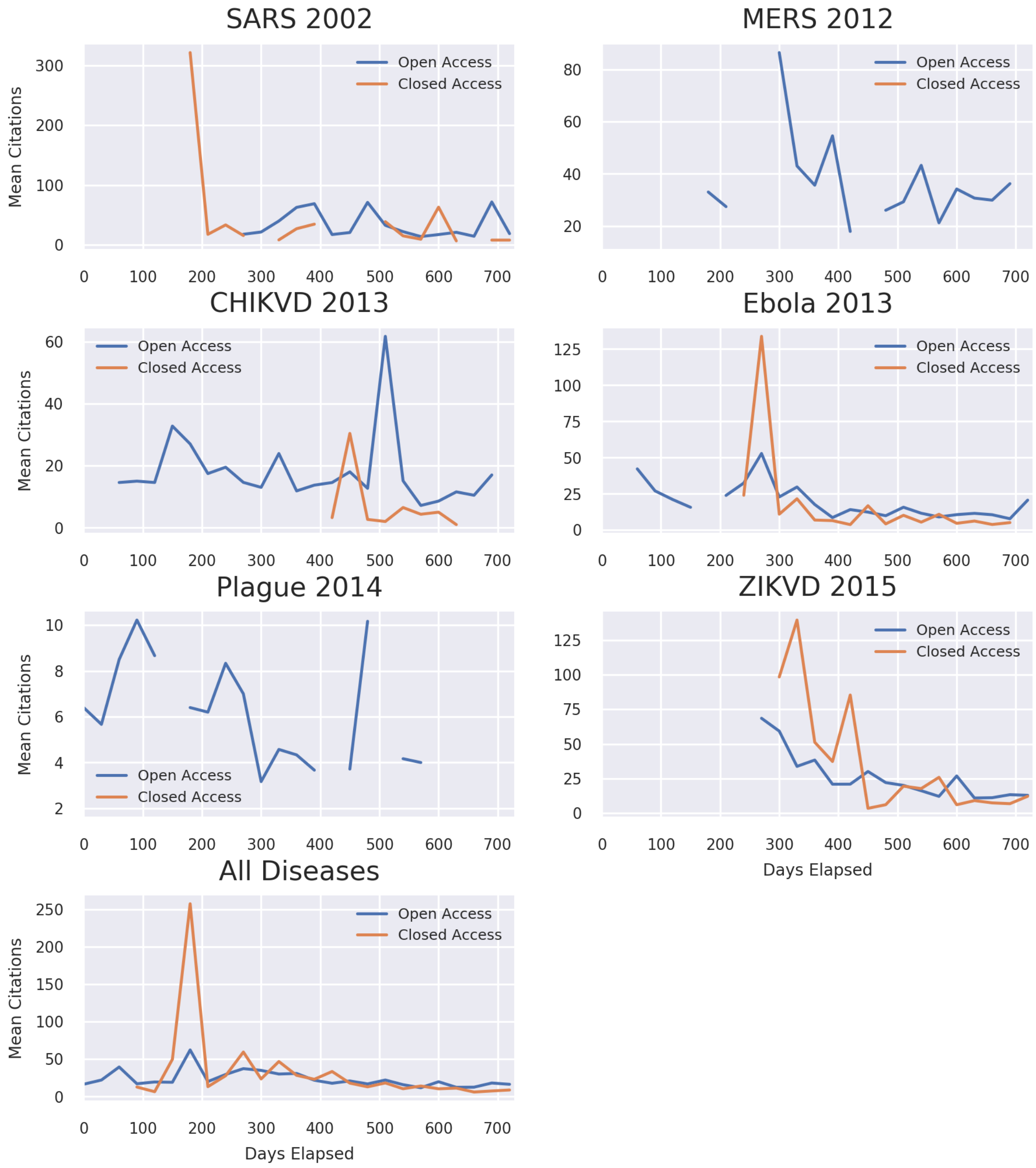
Fig 2.3b: Yearly counts of open access and closed access references made by year of reference paper publication



2.4.5 Divergent citation counts of open and closed access primary publications during early stage response

In ***figure 2.4*** we compare the mean citation count per primary article open access status over 30 day periods following the start of each epidemic. Periods were excluded if they had less than three articles published within their span. Mirroring the results of ***figures 2.2a and 2.2c***, we see some significant spikes in closed access citation rates near the initial stages of the SARS, Ebola, and ZIKVD epidemics. However, while closed access publications yielded some of the most impressive citation counts, they were notably less available for the early stages or even the complete spans of each epidemic barring SARS and ZIKVD. In the combined set of all diseases, it may be observed that open and closed access papers see very similar citation rates after 210 days from the start of the epidemic, with a gradual decline in citations as time progresses.

Figure 2.4: Mean citations per primary paper by open access status and time elapsed



2.4.6 Trends in open access references and referencing journals over time

In ***figure 2.5a,*** we compare the number of references by year made from the top five most referenced journals and subdivide these by the open access status of the referenced articles. Notably, we see an asymmetry in the presence of open and closed access articles, where journals such as the *Journal of Virology, Emerging Infectious Diseases,* and *Proceedings of The National Academy of Sciences* only provided open access references, whereas both *The Lancet* and the *New England Journal of Medicine* presented both open and closed access references. ***Figure 2.5b*** contrasts this by providing the volume of references by primary paper journal, reference year, and reference open access status. In this case we note that each of the five primary paper journals with the highest reference counts referenced both open and closed access articles. In ***figure 2.5a*** we see a somewhat equal albeit noisy distribution of open and closed access references made from the prestige journals. However, ***figure 2.5b*** shows a significant and growing volume advantage to open access references being driven most prominently by open access journals with the limited exception of the 2002 SARS epidemic.

Figure 2.5a: Yearly references made by reference journal, reference year, and reference open access status

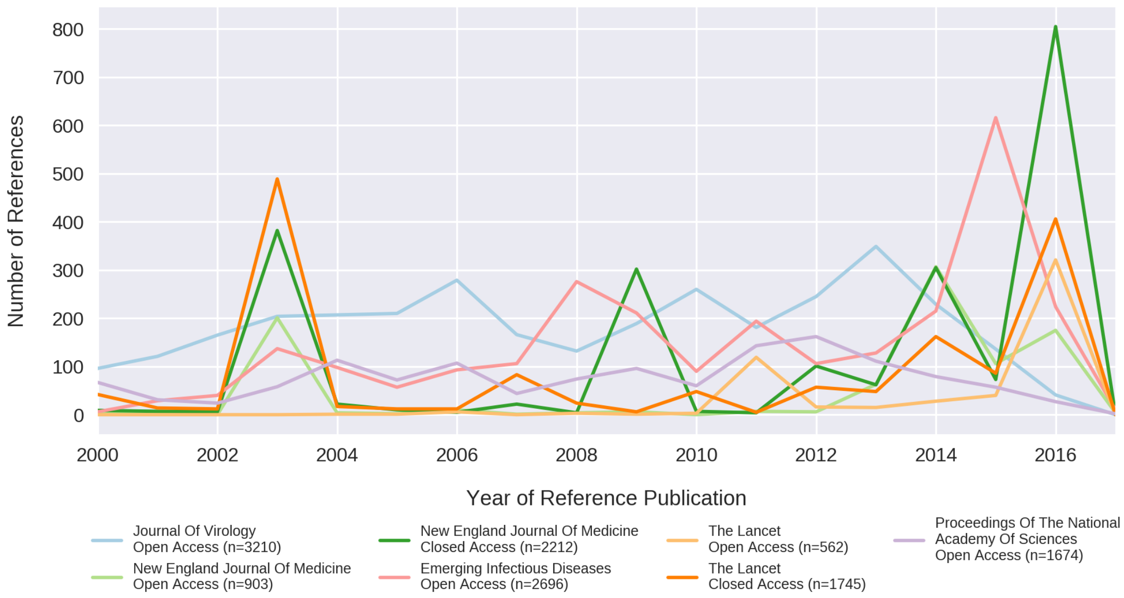
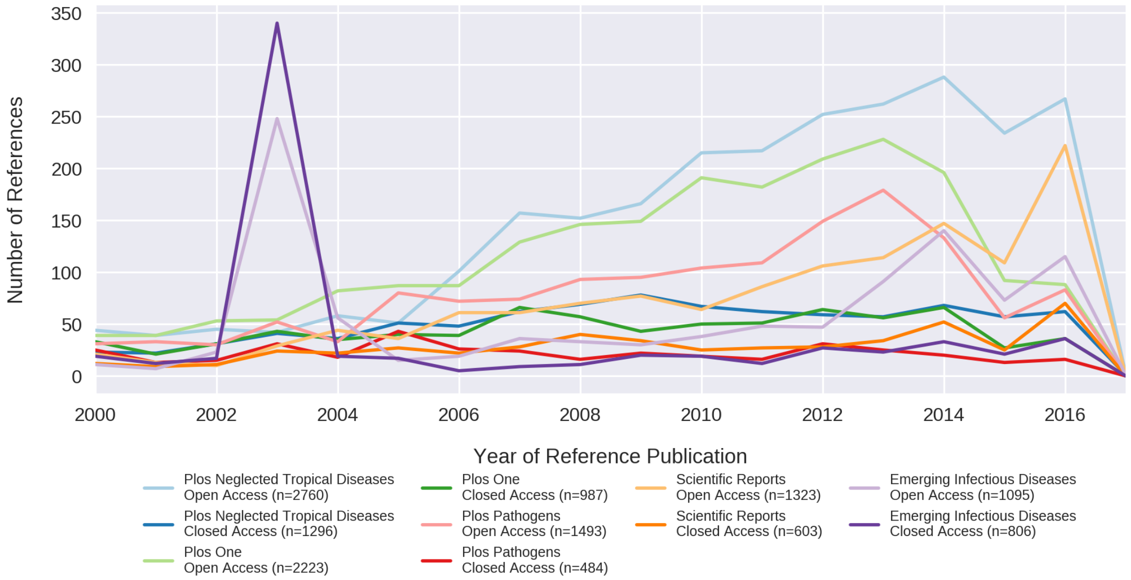


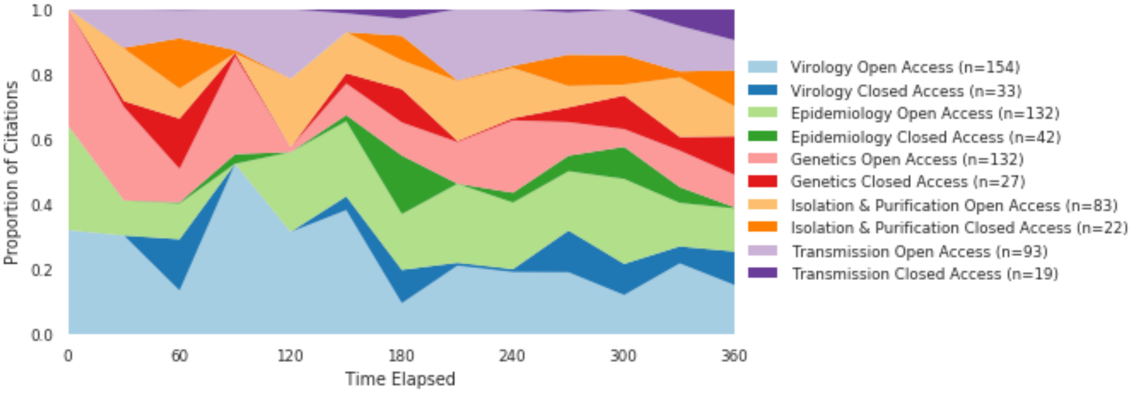
Figure 2.5b: Yearly references made by primary paper journal, reference year, and reference open access status



2.4.7 Trends of primary publication citations by research qualifier over time

In ***figure 2.6*** we show the proportion of all citations made for all diseases per 30-day period by the five most-cited qualifier communities. We constrain this figure to the first 365 days of the epidemic to better describe the emergent phase of each outbreak response. The most common single qualifiers included *virology, epidemiology, genetics, isolation & purification,* and *transmission* in order of decreasing citations. As qualifiers represented broad, standardized categories of research within the PubMed database, this shows us an approximation of the value of certain subsets of knowledge as a function of epidemic time. We see that laboratory methods of pathogen characterization dominated the early stages of the studied epidemics, with *virology* claiming a high percentage of the citations made on early epidemic publications and followed closely by *genetics* between the 30 and 60-day marks. Following this, *epidemiology* and *transmission* related articles grow somewhat in proportion between 120 and 180 days. Across each qualifier community we see a significant bulk of the citations going to open access. After about 240 days the relative citing rate for each qualifier community normalizes.

Figure 2.6: The proportion of primary paper citations made by primary paper epidemic time and qualifier community



2.5 Discussion

During this study, several trends were observed regarding the utilization of open access journals in outbreak response publications. When exploring the relationship between open access status and the rates in which primary articles were published, we did not find open access articles to be published sooner than closed access articles, though the trend was quite close to significance in that direction (p=0.06). This was also observable in ***figure 2.1,*** where no notable trend distinction could be seen between the open access and closed access cumulative curves in the merged set of all diseases' primary papers. However, open access articles still significantly outnumbered closed access articles, comprising 73% of the total post-exclusion papers in this study. Together, these findings suggest that there is no significant difference over time in the manner and delivery with which these articles are published and submitted, and that equal rigor and delays may be applied to each category. However, if one were to propose a different benchmark comparing the time in which a given quantity of articles were published, every included outbreak would give a significant advantage to open access publications due to the sheer volume of articles published.

A similar trend was noticed when comparing the age of references relative to both the referencing primary papers and epidemic time. Contrary to expectations of open science articles driving the early stages of outbreak response, we noted in ***section 2.4.3*** and observed in ***figures 2.2a*** and ***2.2b*** that closed access reference articles were generally published much closer to both the time of their referencing primary article's publication and the start of their respective epidemic. However, there was again a significant numeric advantage to the volume of open access references made. While open access references did not see a greater proportionate use within the early stages of the epidemic, they still significantly outnumbered closed access references for this same time period. Part of this may have been due to the impact of prestige journals, whereby the most exciting or useful early findings within an epidemic may have found their way to higher impact, closed access journals.

Considering the statistical tests applied in ***sections 2.4.2*** and ***2.43,*** we see that open access primary publications likely do not come out earlier after outbreak initiations, and that open access references do not come out closer to either the start of their given outbreak or the publication date of their referencing papers as each test failed to disprove the null hypothesis. It's worth noting the challenges in applying a Two-Sample t-Test in this scenario, as all PubMed entries matching the given query and exclusion criteria were incorporated into the analysis. Likewise, PubMed by no means presents a conclusive record of all publications within outbreak science, and some inherent biases may be present due to difficulties in capturing and accurately filtering this volume of data. Further, article inclusion biases are likely inherent to the OpenCitations COCI database due to its data ingestions from community-curated sources such as CrossRef (Pentz 2001). However, the inclusion of the comparator set provides at least a secondary indicator of trends and/or biases in the methods by which PubMed and COCI data were combined for this analysis, as strongly divergent trends were seen between the primary publications' relationships to open access science and those of the comparator set. Finally, while the OpenCitations data was largely constrained in use to qualitative comparisons, it is worth noting that only a single sample was taken due to challenges in the slow process of querying citation data. This presents a risk of type I error, as novel comparator samples were not independently collected for each analysis.

An analysis of open access reference utilization over time shows a more promising and optimistic trend for open access publications. In ***figures 2.3a*** and ***2.3b,*** we observe that the percentage of open access references by year of primary article and reference article publication respectively are both continuously increasing and are consistently greater than the percentage of open access references used within the comparator random sample of PubMed articles taken for same time period. In ***figure 2.3a*** there's a notable downwards trend towards article publication rates and open access publication percentages between 2016 and 2018. Likewise, ***figure 2.3b*** shows an even longer downturn starting around 2015. This may be due to an inherent bias in the databases used, where later papers might not be simultaneously represented by DOI in both the PubMed and OpenCitations database. As COCI data is based upon periodic ingestions from a range of community curated sources (Peroni et al. 2015), this may represent an inherent lag period to inclusion and render data for later years suspect.

We further explored the temporal utility of primary publications by time from outbreak start using mean citations per primary article as a metric. Here we again saw that some of the first closed access articles were the most highly cited. We also observed that open access articles became available much earlier within each epidemic, and that following this initial period, both saw extremely similar mean citation rates with a gradual decline in mean citations per article. This would suggest that open access science provided some of the first useable pieces of intelligence during the academic outbreak response, yet that closed access synthesis of these findings presented linchpin moments in characterizing each epidemic before they were both treated more or less equally.

By comparing the five most referenced journals' total citations by year, ***figure 2.5a*** shows few strong patterns beyond that open and closed access seem both equally represented. The *New England Journal of Medicine* closed access articles reached the highest annual reference total in 2016, followed by *Emerging Infectious Disease* open access articles in 2015 and *The Lancet* closed access articles in 2003. Contrasting this to ***figure 2.5b****,* we see an impressive showing by open access journals, with consistent increases in the rates of open access references made by each over time, albeit with the same downwards trend towards the later years seen in ***figures 2.3a*** and ***2.3b.*** This would again suggest that while the more prestigious journals retain their status as the early, high-impact publications, open access primary publications and references are growing to drive the bulk of the academic outbreak response. As with prior analyses, 2003 once more presented an outlier year in the midst of the SARS epidemic. It's difficult to say whether the SARS epidemic marked a turning point in publication behaviors due to the lack of notable outbreaks of limited prior history and new-to-science pathogens in the intervening years before the 2012 MERS epidemic. It's entirely possible that larger trends in open access, open data, and open-source science had more of an impact during this time. However, as ***figure 2.3a*** showed a notable upswing in open access reference utilization that directly coincides with the 2002 SARS epidemic, we also see a similar spike in the number of open access publications 400 days from the start of the SARS outbreak in ***figure 2.1,*** further suggesting a change in publication behavior around this period.

Finally, in ***figure 2.6****,* we observed the trends in subject referencing across a merged set of diseases as a function of time from the start of each epidemic. Notably, the laboratory sciences focusing on more granular pathogen characterization strongly drove the early response to each epidemic. *Genetics* lead by a substantial margin before *epidemiology* and *isolation & purification* followed. *Epidemiology* studies defined the later stages of outbreak responses. This provides a rough conceptual approximation of the value of knowledge over time, and suggests the orderly sequence by which outbreak responses may respond to wholly novel threats.

2.6 Conclusion

This study provides a novel method for characterizing the value of open access publications within the realm of outbreak science. We saw that open science papers enjoy neither shorter paths to publication nor referencing within an outbreak in spite of some reduced barriers to access. However, we also observed that open access sciences grew significantly within the study period, claiming a majority of the publications and referenced materials within the academic response to pandemic pathogens. There appeared to be a broad trend of prestigious, closed access journals publishing novel, high value findings within the early epidemic response periods. However, short of these findings, we see that open access science provided some of the earliest accessible data sources for outbreak response. It also saw similar performance albeit drastically greater volume outside of select high-value closed access publications. Finally, we observed a notable sequence of research categories over time, whereby *genetics* lead *virology, isolation & purification,* and *epidemiology* in turn in delivering citable knowledge within the early stages of the epidemic. In each of these subjects, open access publications provided both the bulk of the cited knowledge and the earliest knowledge within a given category.

This study faced inherent limitations due to the potential for latent biases within its utilized databases. Identifying and sussing out those biases may prove intractable, however, due to the challenges in obtaining equally accessible, complete, and free or cheap to query data sources for comparison. This study shows an important trend occurring within the realm of outbreak science, a trend that may be happening organically and at a rate greater than biomedical sciences at large. As open access journals continue to grow in impact factor and volume of articles published, this work may help inform funding and support schemes to promote open access science's continued acceleration of the academic community's response to outbreaks.

Chapter 3

Opioid Social Media Message Mapping

Attribution

The following chapter is based upon the manuscript in prep: *Schlitt, J., Truong, V., Lewis, B., Eubank, S. (2019). Opioid Social Media Message Mapping.*

3.1 Abstract

The Opioid Epidemic presents one of the most challenging threats to public health in the United States, with an average of 130 deaths involving opioid use every day (Scholl 2019). In response, there have been numerous examples of social media campaigns for opioid harm reduction, disseminating messages regarding awareness, Narcan distribution, and the presence of contaminated or adulterated opioids in communities. This yields a substantial opportunity to improve the function of such campaigns via quantitative message mapping. In pursuit of this goal, a ChatterGrabber instance was deployed to collect a master set of 132,521 non-retweet tweets from within the continental United States. From this set, 20,000 opioid and benzodiazepine drug-related tweets were sampled at random, paired with their retweets, and classified by their profile and message types. Profiles were classified singularly as individual, agency, news, organizational, social, or law enforcement pages. Tweets were classified by the presence of one or more messages including generic avoidance, health consequences, legal consequences, contamination/adulteration of drugs, public interventions, and drug use witnessed. Tweets containing no messages or originating from banned or deleted profiles were discarded from detailed analysis. Tweets were analyzed to identify optimal messaging strategies for maximizing virality with regards to timing, message and profile pairings, message tone, urban context, and the application of automation methods. Tweets showed a greater virality if posted before periods of heavy Twitter usage, with Mondays identified as the optimal day for message dissemination and 5AM & 3PM EST identified as the optimal times. Law enforcement and news profiles saw the highest overall viralities by category, with law enforcement performing particularly well with messages regarding public interventions and health consequences whereas news profiles saw higher virality with messages regarding legal consequences. However, news profiles commanded far greater message exposures due to their larger follower bases. Government agencies posting messages of avoidance saw the greatest single exposure rate per category, but generic avoidance messages showed otherwise poor mean viralities or exposures across profile types. Most profile types showed significantly reduced virality if originating from suspected bot profiles. Urban news sources saw strong

virality, and no consistent, statistically significant difference was noted between the virality of tweets with or without a negative tone. These findings may be applied to facilitate public health outreach efforts and to better inform institutional message mapping strategies. This work presents a novel synthesis of social media analytics to pair message content to source profile types with regards to regional cultural contexts.

3.2 Introduction

The Opioid Epidemic presents one of the most challenging and compelling threats to public health in the United States in recent memory. In 2017 alone, 70,237 US residents died from drug overdoses with 47,600 of those deaths involving the use of opioids (Scholl 2019). This marked a 9.6% increase in total drug overdose mortality from the previous year with an average of 130 deaths per day due to opioids in specific (Scholl 2019). Early origins to this epidemic may be observed via prescribing rates, as the number of new opioid prescriptions roughly quadrupled between 1999 and 2010 (Centers for Disease Control and Prevention (CDC) 2011). A 2013 study revealed that roughly 80% of people who use heroin abused prescription opioids before trying heroin (Jones 2013). While the prescribing rate of opioids has steadily dropped since 2012, prescribing rates and resulting addiction remains significantly higher in rural communities with greater percentages of white residents, greater prevalence of chronic illness and disability, and greater per capita numbers of dentists and primary care physicians (Hoots et al. 2018). Another notable threat posed by this epidemic is the risk of contaminated or adulterated street drugs (Mohr et al. 2016). A prominent case in Atlanta saw 24 overdoses and a single death stemming from a batch of counterfeit Percocets (Rhonda Cook 2017).

The severity and scale of the crisis provides both a need and an opportunity to apply quantitative message mapping strategies. There have been numerous examples of use-agnostic social media messaging campaigns for opioid harm reduction. Campaigns such as *DanceSafe* and *Face the Fentanyl* have achieved significant followership by disseminating messages focused upon awareness and Narcan distribution and by alerting the public to the presence of contaminated or adulterated opioids within communities ("DanceSafe" n.d., "Face the Fentanyl" n.d.). Under the Health Belief Model, such campaigns may fulfill the role of cues to action, thereby increasing the perceived threat of a disease to effect positive behavioral change within individuals (Rosenstock 1974). While no social media platform will perfectly represent the populace, social media analyses provide distinct opportunities to analyze the manner in which target audiences engage with messaging campaigns. This in turn offers generalizable findings from message mapping that may be applied on alternate social media platforms or via more classical health messaging channels.

In order to identify strong, health-positive messaging strategies with regards to the opioid epidemic, we deployed ChatterGrabber to collect a set of 20,000 tweets for manual classification (J. T. Schlitt, Lewis, and Eubank 2015; J. S. Schlitt and Lewis 2017). Virality was selected as a key target metric for message success, herein defined as the ratio of the number of total potential exposures from a tweet's source profile and subsequent retweets to the number of a tweet exposures from the source profile alone. Tweets were analyzed to identify optimal messaging strategies for maximizing virality with regards to timing, message content, source profile type, message tone, urban context, and the presence of suspected automation.

3.3 Methods

3.3.1 Data collection

A ChatterGrabber instance was run from January 28th, 2018 to August 8th, 2018 to collect a superset of public tweets related to the opioid crisis in the United States. These tweets were queried for origination within the continental United States, the presence of opioid and/or benzodiazepine related search-words, the absence of exclusion keywords, and the absence of retweet status. From this superset, a subset of 20,000 tweets, heretofore referred to as *original tweets*, was randomly sampled for further analysis.

3.3.2 Inclusion criteria

For each keyword set, inclusion was determined by the case-insensitive presence of a given substring with tweets. Substrings such as *"oxy"* which were frequently found within larger, unrelated words were appended with spaces to improve keyword specificity. Search keywords were derived from a DEA list of opioid and benzodiazepine related keywords ("2018 Slang Terms and Code Words" n.d.). While efforts were primarily focused on opioids, benzodiazepines such as Xanax were included due to reports of counterfeit benzodiazepines contaminated with fentanyl and other high potency opioids (Pergolizzi et al. 2018; Allen, n.d.; Abuse and Overdose 2016). A reduced subset of keywords ***(Table 3.1)*** was manually curated from the initial DEA list during a pilot study phase ("2018 Slang Terms and Code Words" n.d.). Keywords were retained if they were found frequently within tweets that appeared related to opioid or benzodiazepine drugs and if they did not yield excessive false positives due to their presence in non-drug-related conversation. Exclusion keywords ***(Table 3.1)*** were selected for words commonly found in original tweets that contained search keywords but did not appear related to real and present drug use. Exclusion keywords generally referenced musicians, musical lyrics, and political discourse.

Table 3.1: Keywords and exclusions

Keywords	Exclusions
['actiq ', 'astramorph ', 'benzo', 'black tar', 'buprenorphine', 'carfentanil ', 'codeine ', 'dance fever ', 'demerol ', 'dilaudid ', 'drug', 'fake drugs', 'fentanil', 'fentanyl ', 'fentenel', 'fentenil', 'fentora ', 'fiorcet ', 'gray death', 'grey death', 'heroin ', 'hydro ', 'hydrocodone ', 'hydromorphine ', 'lean', 'lorcet ', 'lortab ', 'meperidine ', 'methadone ', 'morphine ', 'muffin man', 'naloxone ', 'narcan', 'norc', 'norco ', 'opioid', 'oxy ', 'oxy blues ', 'oxycodone ', 'oxymorphine ', 'perc ', 'percocet ', 'poison pill', 'prescription', 'purple drank', 'roxicet ', 'scag ', 'speedball', 'suboxone ', 'vicodin ', 'xan', 'yellow pill ', 'zan', 'zohydro ']	['Blac Youngsta', 'Cedric Gervais ', 'Future ', 'Lil Dicky ', 'Lil Durk', 'Lil Uzi Vert ', 'Lil pump', 'Mally Mall ', 'Mask Off', 'Migos', 'Post Malone', 'Quavo', 'Schoolboy Q', 'Stripper Joint', 'The Weeknd ', 'Tyga ', 'Wiz Khalifa ', 'bad and boujee', 'conservatives', 'cookin up dope in the hotpot', 'cookin up dope with the uzi ', 'death penalty', 'duterte', 'jazz', 'liberals', 'music ', 'song', 'trump']

3.3.3 Geocoding methodology

The original tweets were geocoded using the highest resolution location data available for a given tweet. Tweet latitude and longitude coordinates were given the highest priority, followed by the centers of geographic bounding boxes, and using named locations as the lowest value. Tweets with only named locations were geocoded to coordinates using the GeoPy Python wrapper for the Google Maps geolocation API. Tweets that lacked coordinates originating within a bounding box of the lower 48 states of the continental United States were excluded.

3.3.4 Manual labeling of message and profile types

Original tweet text contents were manually labeled by a human operator for the presence of one or more drug use related messages, detailed in ***table 3.2***. Tweets that did not possess one or more messages were marked as irrelevant. Messages were considered present if the tweet contained a direct reference to behavior, consequences, or sentiments regarding real and present opioid or benzodiazepine drug use. Exceptions to this criteria generally followed the form of using drug use as a conversational scapegoat (e.g. *"We need the moat because these people are bringing their drugs, diseases, and crime."*) or hyperbole (e.g. *"The group of people I don't like is comprised entirely of pill-popping fentanyl users, someone should drug test the lot.")*. Public message threads were reviewed for additional clarity for tweets in which the original messages were not fully interpretable absent the context of the conversational thread in which they originated. Each tweet was also separately labeled as either negative or not negative in tone via the Vader sentiment analyzer from the Natural Language ToolKit (Hutto and Gilbert 2014).

Table 3.2: Message categories

Message Type	Definition	Examples
Contamination	Tweet discussed drugs with dangerous potency or unknown ingredients	"That stuff isn't even Percocet anymore, what are you taking?"
Public Intervention	Tweet discussed public actions to decrease the public burdens of opioid abuse	"New law requires doctors to limit all Xanax refills"
Legal Consequences	Tweet discussed legal consequences of opioid abuse	"My plug just got arrested"
Avoidance	Tweet discussed avoidance or aversion to opioid abuse	"Why are you guys wasting your lives on those pills?"
Health Consequences	Tweet discussed negative health consequences of opioid abuse	"I lost my best friend to percs and molly"
Use Witnessed	Tweet discussed the recent abuse of opioids	"I'm off the lean and feeling the feels"

Irrelevant	Tweet did not match any of the above categories	“Finally got my prescription for Xanax”

The set of user profiles that created one or more tweets with relevant messages was identified and classified to a single profile type from ***table 3.3,*** heretofore referred to as *original posters*. Original posters were automatically labeled as banned or deleted via the HTML contents of their respective profile pages if their accounts were suspended or removed by the time of scoring. Original poster profiles which were neither banned nor deleted were manually classified by a human operator by assessing their Twitter profile page for indicators of a given profile type. Profiles were generally classified by their apparent use, and individuals acting as representatives of larger organizations were classified by the profile type of their given organization. For example, individuals stating *“opinions my own”* in their biographies would be classified to the profile type of their organization if their recent posting history was clearly and primarily furthering the goals of the larger organization.

Table 3.3: Profile categories

Message Type	Definition	Examples
Law enforcement	Profile representing a law enforcement agency or public figurehead thereof acting in an official capacity	Police departments, DEA, local sheriffs
Agency	Profile representing a government body or public figurehead thereof acting in an official capacity	POTUS, local politicians, city government
Organization	Profile representing a school or private organization or public figurehead thereof acting in an official capacity	Companies, schools, non-profits
News/media	Profile representing a media agency or public figurehead thereof acting in an official capacity	TV Stations, newscasters, news bloggers

Social	Profile employed for entertainment or the advertisement thereof	Musicians, comedians, meme pages
Individual	Non-professional profile strictly representing an individual or an individual's interests	Generic class for people who use opioids, personal profiles of public figures
Banned/deleted	Profile deleted, suspended, or otherwise no longer accessible	Banned individuals, cancelled profiles

3.3.5 Automated labeling of profile and message attributes

Additional categorical labels were applied to describe whether or not a given original poster profile appeared to be a bot and whether or not a given profile was located within an urban area. An analysis of recent profile posts by the Botometer API was used to label profile suspected-bot-status (Davis et al. 2016). Following Pew Research’s practices, profiles were labeled as a suspected bot if they received a Botometer bot-score of greater than or equal to 0.37 (Gramlich n.d.). Original poster profiles and retweeting poster profiles suspected of being bot accounts were retained in analyses as they likely influenced the discourse and as many may have represented legitimate automated notification systems. Profiles were labeled as urban if their geocoded location intersected with the 2017 urban areas shapefile from the US Census Bureau (US Census Bureau Geography 2012).

Following classification tasks, retweets of each original tweet were queried via the Python Tweepy wrapper for the Twitter statuses/retweets API. While Twitter will return no more than the 100 most recent retweets for a given post, in practice the highest number of retweets found to a given original tweet was 94, so it did not appear that any retweets were lost due to this limitation. It’s important to note however that this method only captures retweets made by a user directly clicking the retweet button of a previous tweet, assigning them to the original creator of the tree of retweets. Therefore, alternate methods of retweeting such as quoting, screen capping, and copy/pasting original tweets or retweets are inherently not captured by this method.

3.4.6 Units of comparison

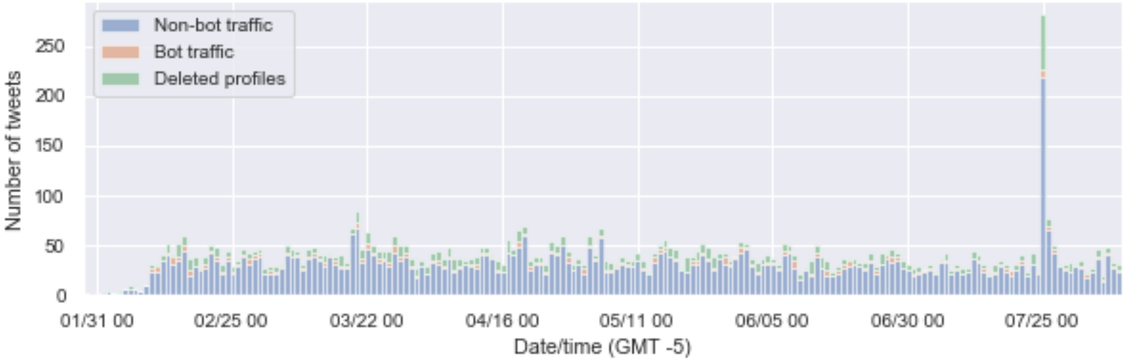
Tweets were assessed via the proposed metrics of ***volume, exposures, virality,*** and ***relative virality.*** Volume represented the direct count of captured original tweets for a given profile type, message type, etc. Exposures represented the maximum potential viewership of a given tweet via the sum of all friends and followers of the originating poster's profile as well as that of all profiles retweeting the original tweet. Virality was defined as the sum of all exposures for a given tweet divided by the number of direct followers and friends of a given profile. By this metric, a user with 25 friends and followers being retweeted once by user with 75 friends and followers would achieve a virality of four, as said tweet was able to reach four times its initial audience through retweeting. Relative virality represents the mean virality for a given subset of tweets divided by the mean virality for a paired subset of tweets, such as urban media tweets regarding contamination vs. non-urban media tweets regarding contamination.

Analyses were conducted to evaluate exposures and virality as a function of message types by profile types and message and profile type by day of week and hour of day. The distribution of retweets by time elapsed and distance travelled was plotted for review. The relative virality of urban vs. non-urban message sources, negative vs. non-negatively toned messages, and suspected-bot vs. non-suspected-bot original posters was evaluated by profile and message type pairings to further identify optimal messaging strategies. The relative virality analyses were also assessed for statistically significant difference in means between conditional subsets via Welch's t-Tests under the assumption of unequal variance due to the multitude of category subsets. For each analysis by message category, messages were subsetted to a given category by presence of a given label within the manual scoring data. As each tweet may have one or more labels, complex messages containing multiple labels were represented once for each category with which they were labeled.

3.4 Results

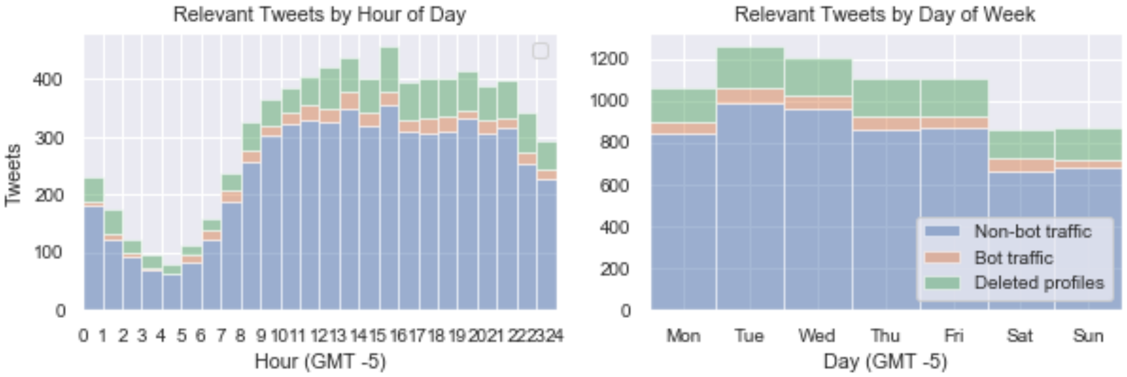
3.4.1 Data collected

Figure 3.1: Relevant tweets by day of year



132,520 original (non-retweet) tweets with locations within the continental United States were collected between January 28th, 2018 and August 8th, 2018. From this superset, 20,000 tweets were randomly selected for analysis, 7,475 (37.5%) of which were manually labeled as relevant. 6,298 unique Twitter users created relevant tweets, 322 of which were suspected bots and 1,028 of which had been deleted since posting. Tweet volume remained relatively consistent throughout the collection period barring one notable spike in late July following a celebrity overdose ***(figure 3.1)***. Likewise, tweets followed relatively typical posting patterns with regards to hour of day and day of week, showing the a slightly bimodal distribution of hourly non-bot traffic around 12PM and 8PM Eastern Time and daily non-bot traffic on Tuesdays ***(figure 3.2)***.

Figure 3.2: Relevant tweets by hour of day and day of week



5,078 publicly available retweets from 4,751 unique user profiles were collected. These profiles included 492 suspected bots and 49 deleted profiles. 3,571 (70.3%) of the retweets contained locatable geographic information and 3,282 of these retweets (64.6% of total retweets, 91.9% of locatable) originated from within the United States.

Individuals' profiles comprised the vast majority of the original tweets captured, with 55.1% of all tweets coming from individuals, followed by 12.0% from news profiles, 10.0% from deleted profiles, and 9.1% percent from social profiles ***(figure 3.3)***. Drug use-witnessed messages were the most prevalent by volume with notable leads amongst the individual, social, and deleted profile categories. However, this trend was significantly diminished when accounting for the actual sum of exposures due to each profile type ***(figures 3.4, 3.5)***. News profiles saw the highest number of total message exposures and mean exposures per tweet, commanding 31.3% of all total tweet exposures. This was followed by individual profiles with 17.3% of the total message exposures, deleted profiles with 16.3%, and organization profiles with 14.4%. However, in contrast to the total exposures, news, agency, and social profiles were able to command the highest mean exposures per message. The 13 captured agency profile tweets with avoidance messages averaged 68,805 exposures per tweet, the greatest single mean exposures per profile and message category pairing. Following this were news profiles posting messages regarding legal consequences, averaging 40,499 message exposures across 357 tweets.

3.4.2 Message volume and virality by profile and message categories

Figure 3.3: Total message volume by profile category

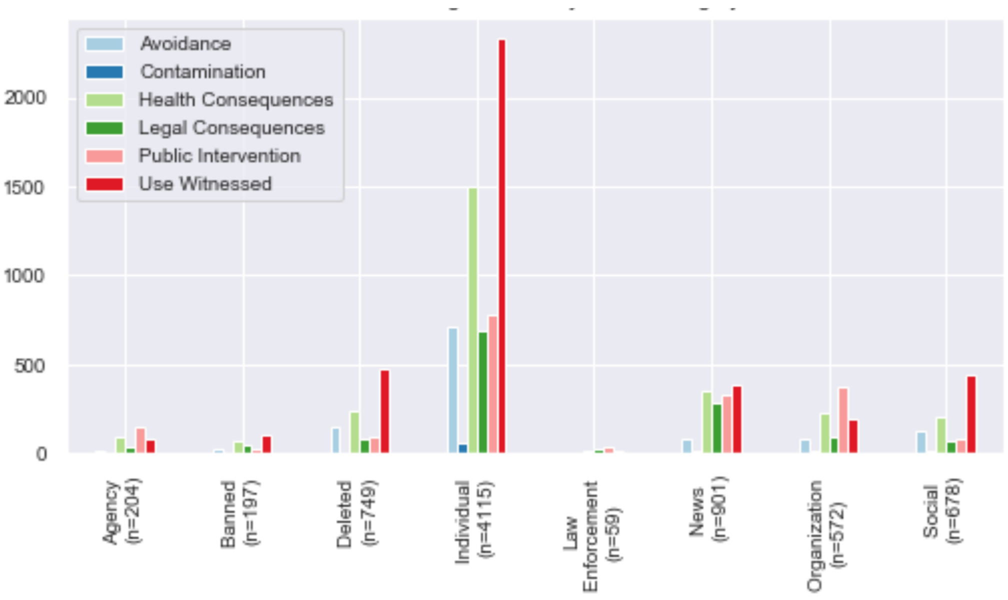
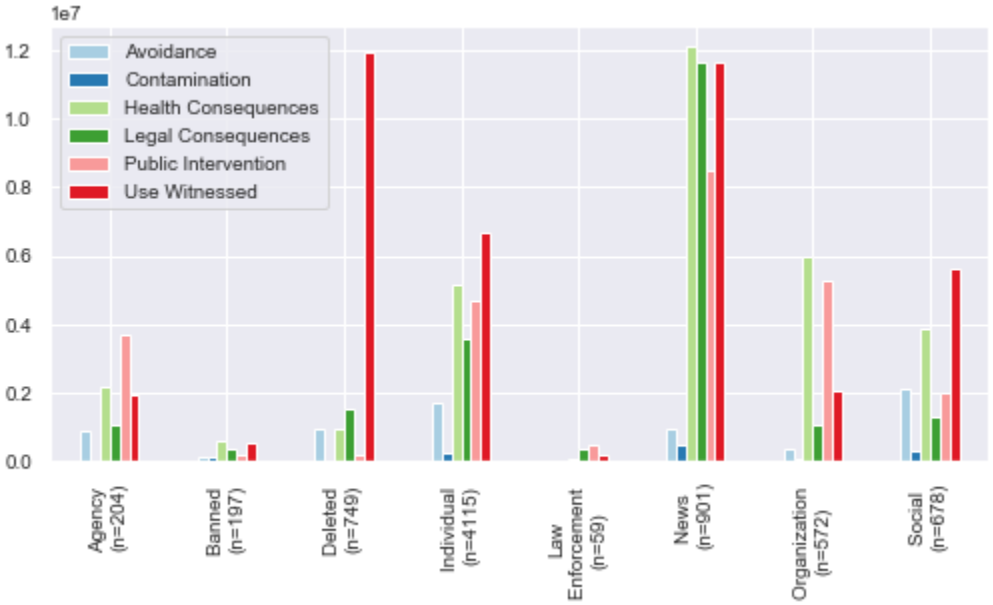
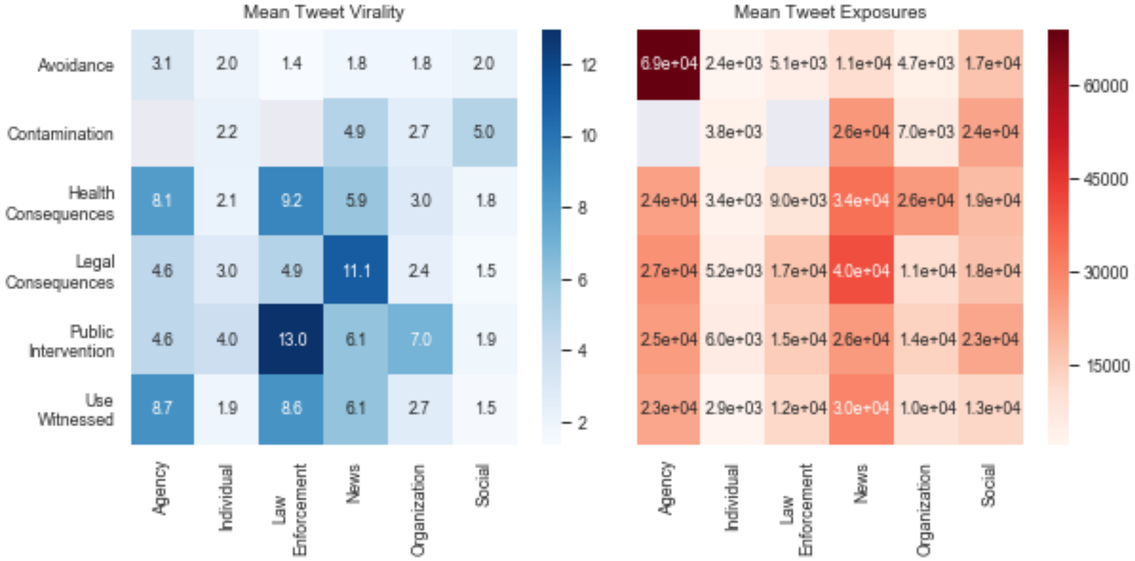


Figure 3.4: Total message exposures by profile category



Deleted profiles showed very strong amplification of use-witnessed message volume to exposures. This may have been the result of profiles promoting illicit activity or internally retweeting bot networks. Law enforcement profiles commanded a very slim share of the discourse, delivering only 0.8% of the volume of captured tweets and 1.0% of the exposures. However, while they were used sparingly, law enforcement pages actually carried public intervention and health consequence messages the best relative to their follower counts ***(figure 3.5)***. Many of these posts promoted community events and prescription take-back programs. Message virality tended to line up logically across user categories, where agencies saw excellent virality for health consequences whilst organizations saw excellent virality for public interventions. Contamination messages were best conveyed by local news or social profiles. Individual profiles, social profiles, and avoidance messages saw generally poor virality across categories with the exception of social pages carrying messages regarding contamination.

Figure 3.5: Message virality and exposures by message type and source



An analysis of mean message virality and exposures by hour of source tweet creation ***(figure 3.6)*** showed significant variance across the day. Exposure volume was largely concentrated during the workday, ranging from 8AM–4PM Eastern Time. Mean virality by hour was much more concentrated, showing its greatest peak at 5AM Eastern Time and a steady increase to a second peak at 3PM, albeit with each showing wide variance. Repeating the same analysis with regards to the day of the week, daily exposures and virality were both observed to peak on Mondays. There was also significantly less hourly variance in virality by day than exposure by day.

3.4.3 Weekly and hourly dynamics of message virality

Figure 3.6: Mean tweet exposures & virality by hour and day

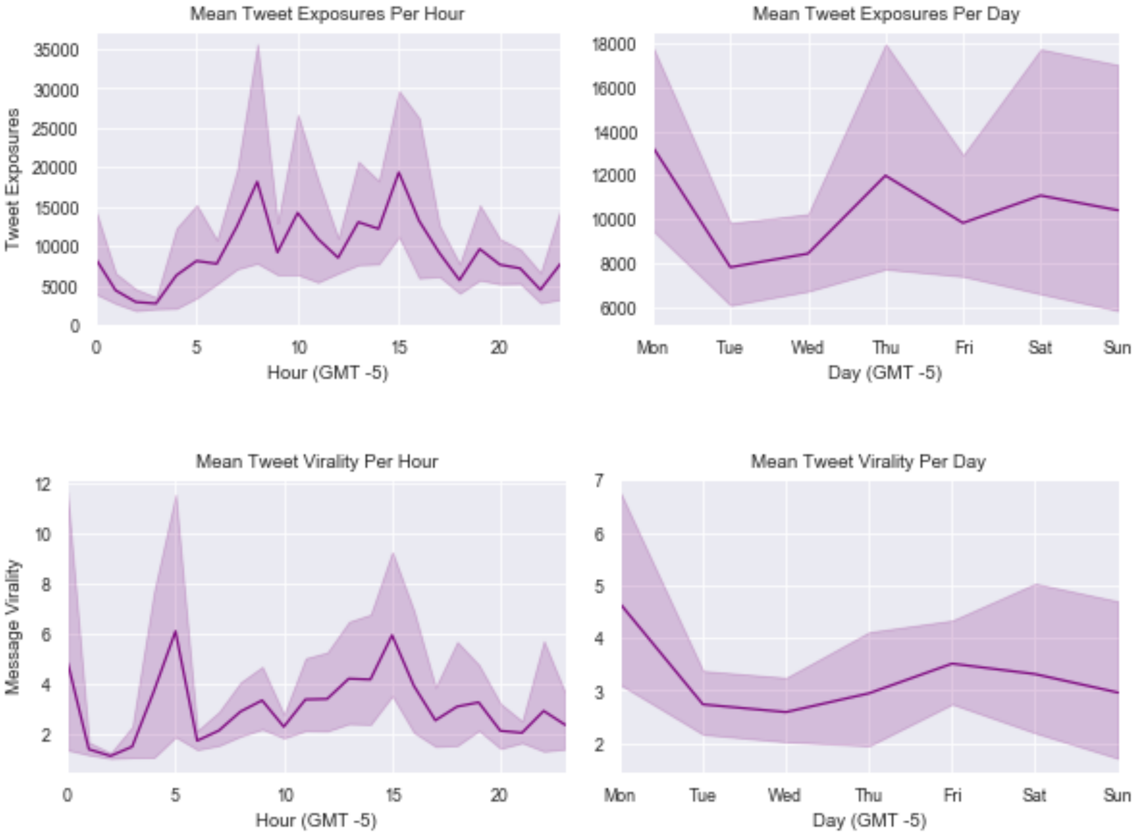
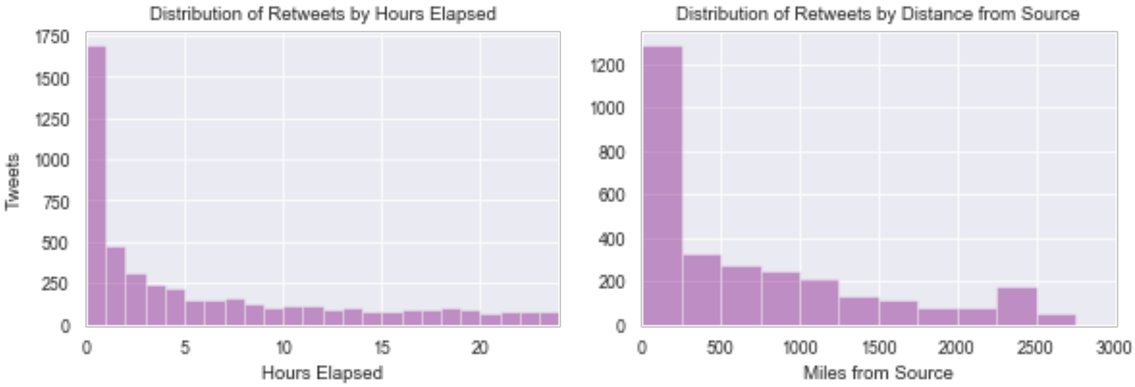


Figure 3.7 further explores these dynamics, using retweeting by hour and distance as a proxy to audience views and engagement. A histogram of retweets by hours elapsed from original post creation shows that optimal engagement occurs largely within the first two hours following tweet creation. However, a long right tail is observed, and a near stable rate of retweeting with a steady drop off took hold approximately ten hours following original tweet creation. Retweeting by distance from source showed a less prominent drop off with a notable secondary peak at the right tail. This was likely due to retweeting between densely populated coastal cities. Only 18% of retweets were created within 50 miles of their original tweet's profile location, with most likely exceeding the expected bounds of a city or county-level health agency. The distribution of tweets by both distance and time elapsed showed no notable or consistent trends with regards to source profile or tweet message types.

3.4.4 Spatiotemporal message dissipation

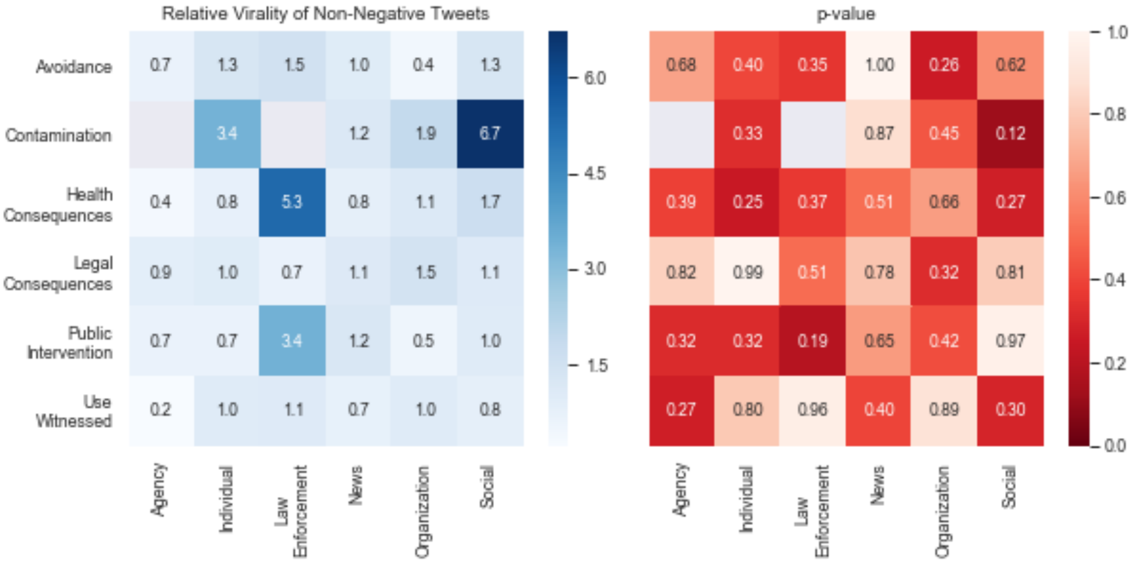
Figure 3.7: Message persistence and travel



An analysis of relative message viralities by message tone showed no statistically significant ($p<0.05$) advantage or disadvantage to positivity in tweets for any profile and message type pairing. Social profiles tweeting messages of contamination showed the greatest relative virality, with the average non-negative tweet reaching 6.7 times the virality of negative tweets (p=0.12). Law enforcement profiles' two highest virality profile pairings, public intervention and health consequences, also showed greater relative viralities for non-negative tweets, seeing increases of 5.3 (p=0.19) and 3.4 (p=0.37) over their base viralities of 9.2 and 13.0 respectively. Conversely, agency profile messages appeared to benefit from negatively toned tweets across categories, though the effect was again too weak to firmly establish a benefit.

3.4.5 Relative viralities by profile tone, location, and bot labeling

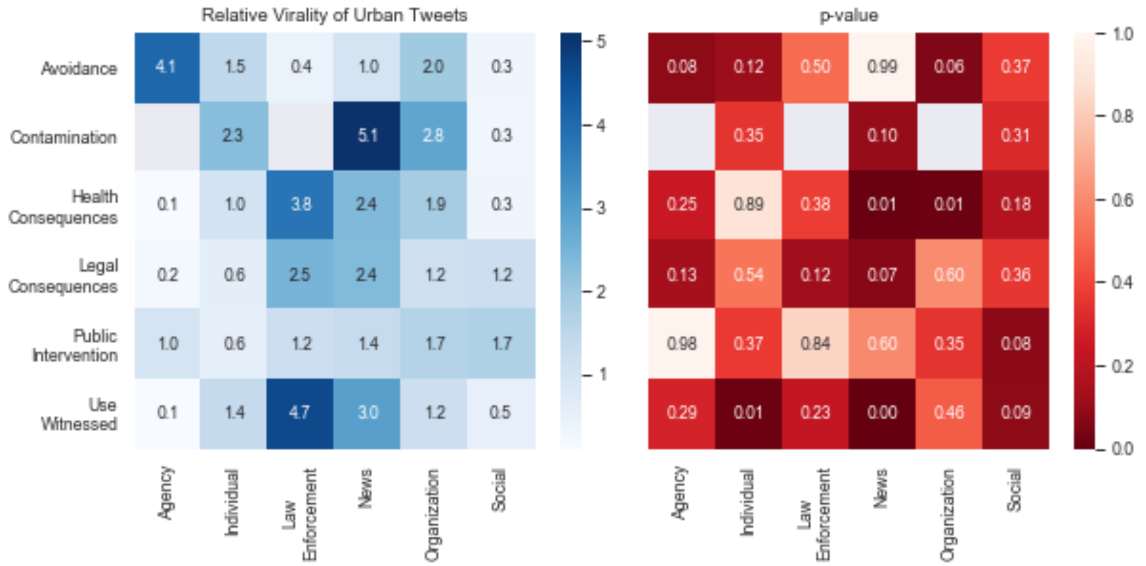
Figure 3.8: Relative viralities by category as a function of message negativity



When comparing relative viralities by urban message sources a much stronger effect was observed, with four profile and message pairings yielding statistically significant effects ***(figure 3.9)***. News and organizational profiles tweeting health consequence messages both saw 2.4 (p=0.01) and 1.9 (p=0.01) times higher virality for urban-located profiles relative to non-urban profiles. News and individual profiles sharing messages regarding drug use witnessed also saw statistically significant differences between urban and non-urban accounts, with urban accounts achieving 3.0

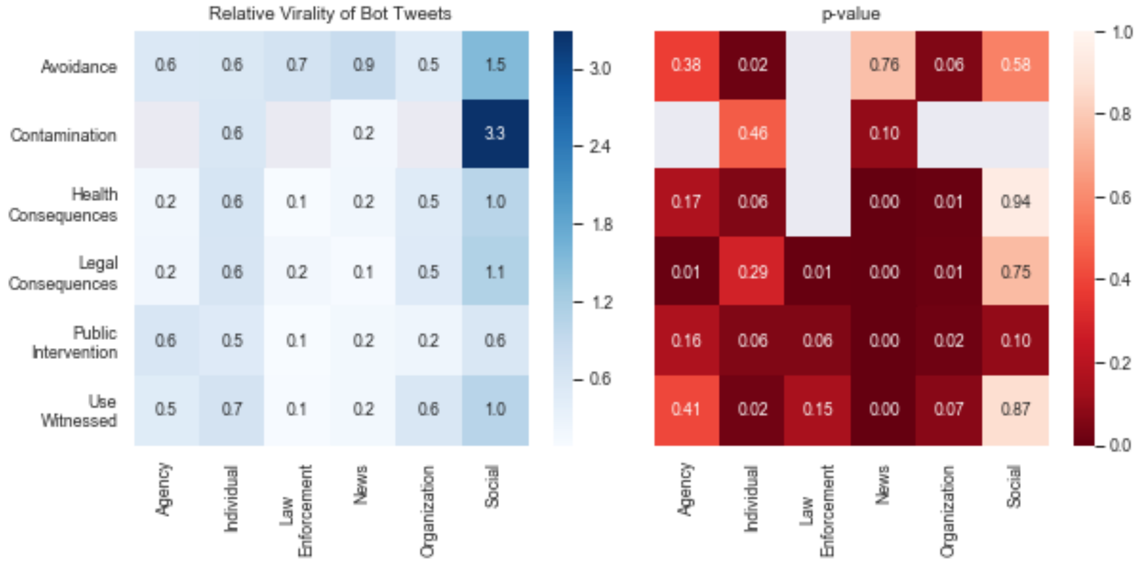
(p<0.01) and 1.4 (p=0.01) times greater virality than non-urban accounts. Agency and social pages showed somewhat greater virality for non-urban tweet sources, though the effect was not statistically significant.

Figure 3.9: Relative viralities by category as a function of original poster urban location



For the final analysis, a comparison of virality as a function of the suspected bot status of profiles was conducted. It's important to note that bots in this context were not necessarily malicious or manipulative in nature. Such tweets could also constitute automated retweeting or posting of procedurally generated reports serving the public interest. Here, a striking negative effect to automated profile use was observed. All profile types with the exception of social pages saw reduced virality across all message types. Likewise, 12 of the 25 non-social page pairings available within the data set showed statistically significant differences between the viralities of tweets coming from suspected bot profiles whence compared to those coming from profiles not suspected of being bots.

Figure 3.10: Relative Viralities by category as a function of suspected bot status



3.5 Discussion

The study identifies several strategies for more effective message mapping with regards to source, contents, time, tone, and locational context. Messages were notably ephemeral in space and time, with only 17% of retweets coming from a 50 miles radius of the original poster ***(figure** 3.7)*. Likewise, retweets were observed to drop off rapidly after two hours elapsed from the time of the original post, though a long lingering tail was observed ***(figure** 3.7)*. Both of these effects were consistent across user profile and message types categories. This suggests that human interactions with the Twitter timeline algorithm were the key drivers in spatiotemporal message dispersal, rather than local or contextual heterogeneities.

While this rapid dispersal may limit one's ability to directly target a given health jurisdiction, it raises two notable best practices. Firstly, as retweeting rates and engagement seem to rapidly decline after two hours, in the face of an urgent crisis such as a contaminated batch of drugs, agencies should consider repeated posting to ensure messages do not disappear into the timeline. However, there are likely to be practical limits to re-posting strategies, as ***figure 3.10*** shows that non-social suspected bot profiles saw significantly lower virality than profiles not suspected of being bots. This might suggest either a general audience rejection of automation, or a general resistance to repetitive messaging strategies. Therefore, while frequent tweeting appears important to maintaining audience exposures throughout a crisis as novel Twitter users log in, it might induce diminishing gains if messages become repetitive or do not appear appropriately human in origin. The exclusion of social profiles from this trend raises interesting questions. Social profiles often broadly resembled individual pages in tweeting behaviors by message category distribution, message contents, and viralities by topic. This lack of a clear automation penalty for suspected profiles may be due to factors such as greater skill in message crafting, different goals for audience engagement, or simply challenges in the algorithmic identification of such pages. A question remains whether these profiles were attached to astroturfing campaigns to widen pre-existing social tensions, as many repeatedly posted use-witnessed messages that appeared political in nature. Secondly, there seems to be a broad benefit to getting messages out ahead of high exposure periods. As seen in ***figure 3.6,*** tweets posted at 5AM or 3PM EST achieved higher mean virality than tweets posted at other times. Given the two-hour optimal engagement window, such tweets likely had the advantage of preceding the beginning and end of the working day where social media use may have increased. ***Figure 3.2*** somewhat corroborates this, as such windows loosely preceded the daily periods of higher frequency tweeting. Combining a weekly analysis, we see a similar virality benefit to earlier tweets, where the greatest weekly volume of tweets was shared on Tuesdays and Wednesdays, but the greatest mean virality was observed on Mondays.

Beyond the timing of messages, there is significantly more potential for nuance in the pairing of messages to source profiles. Analyses were conducted with the assumption of an implicit economy of posting, whereby pages with higher viewership may pose proportionately greater barriers to the health practitioner working to circulate pro-health messages with community partners. In ***figure 3.5*** we observe government agency and news sources surpassing other profile types by orders of magnitude in many cases.

However law enforcement profiles yielded some of the highest mean viralities in spite of their low volume and number. While generic avoidance messages performed poorly across all profile categories, we see the second lowest captured mean virality for law enforcement profiles was messages regarding legal consequences of drug use. Such messages often detailed investigations, stings, drug busts, and related police actions. These messages were better carried by news profiles, where they reached the second highest virality behind law enforcement posts regarding public interventions.

The analysis of positive tone showed no consistent, statistically significant difference between tweets which did and did not use a negative tone. With this in mind, social and individual profile tweets regarding contamination did yield some of the stronger measured benefits to non-negative tone. Likewise, law enforcement posts regarding health consequences and public intervention also derived some benefit from positive tone. An interesting trend arises in combination with the excellent virality of these messages from law enforcement and the weaker virality for law enforcement legal consequences messages. Twitter communities appeared broadly receptive to law enforcement messages aligned positively with concern and assistance, but significantly less receptive to law enforcement messages regarding legal consequences.

Taking into account the urban context of tweet sources, we see in ***figure 3.9*** that all four statistically significant comparisons of mean virality by profile and use category favored urban contexts. Agency and social pages appeared to favor a non-urban context, though the effect was not statistically significant. One notable omission across all categories is the presence of agency and law enforcement tweets regarding contamination. As many of the tweets came from a use agnostic, pro-health perspective, this may have been due to positioning, whereby such profiles may have been bound to post from a use-negative perspective or otherwise use different messages in pursuit of their goals. However, as contamination messages were the least common message category, and as contaminated street drug batches may be a rare or otherwise difficult-to-discuss event, this omission may have been due to other factors. Still, this finding highlights the importance of individual and social profiles for self-protection amongst people who use opioids.

The first practice for any public health professional should be to utilize the available network of partner profiles with the greatest numbers of followers to distribute and disseminate health-positive messages. However, as such profiles may be inherently more difficult to leverage, an understanding of virality by source may prove invaluable in making the best use of available resources. While individuals drove the greatest volume of messages shared, other message sources were able to secure comparable if not greater shares of the total message exposures. Therefore, even when combating an entrenched health-negative culture, it is still possible to secure a presence in the discourse by utilizing available community resources. Analysis of optimal combinations of message, source, tone, and urban context might yield the best messaging strategies for a given message. For example, when dealing with a contaminated batch of drugs in town, one might consider a cautionary or neutrally toned news source for urban regions or a positive, use-agnostic tone for rural social influencers.

This study was conducted with several inherent limitations. First, opioid tweets and profiles are particularly difficult to classify quickly. Manual scoring efforts averaged 500 tweets or 250 profiles per hour, and each tweet was scored only once in the hopes of capturing a greater sample of rare messaging events. Pooling or automating this task would have had a significant impact on turnaround times. Due to the time required for labeling a sufficiently large sample, it may prove prohibitively slow to manually classify tweets or profiles under the threat of a looming crisis, particularly if drawing one’s data from a narrow geography or uncommonly discussed subject. Therefore, this approach is generally better suited to planning and preparation tasks for anticipated or otherwise long-term health threats.

Secondly, as manual classification relies upon intuitive judgments made regarding an inherently evasive culture, one’s manual classification efficacy will drift with time as familiarity with the vernacular and memes surrounding a given subject develops. Therefore, it crucial to consider sampling strategies during manual labeling tasks to account for such drift. For many of these profile and topic pairings a low tweet volume was observed, all the while retweets showed limited geographic bounding to their source profile’s location. Therefore, targeting messages to a given region may prove challenging if not impractical, though the impact of locally trusted information sources for risk messaging is not to be ignored.

3.6 Conclusion

This study demonstrates a viable method of assessing strategies in message creation, alignment, and distribution in order to provide behavioral cues to modify population risk perceptions. This method has an inherent trade-off between labor costs and sample size and may serve best as a means of preparing messaging strategies for potential crises rather than responding to imminent ones. Tweets will neither perfectly represent nor change the minds of the public, but they do provide a ready and organized data source to assess the general pulse of public perceptions and engagement. Knowledge of such may yield applications and generalizable findings for health practitioners working across diverse platforms or modes of communication to push health-positive messages to a target population. While this study used only publicly available tweets, the ethical collection, storage, and dissemination of such data must always be considered in the application of social media research. The sensitivity of novel surveillance tools may exceed the assumed perception of anonymity by numbers within a given population, and it is important that such technologies be narrowly deployed within the scope of preventive and supportive capacities.

3.7 Funding

ChatterGrabber was supported by the National Institutes of General Medical Sciences of the National Institutes of Health under award number 5U01GM070694-11 and the Defense Threat Reduction Agency under DTRA CNIMS Contract HDTRA1-11-D-0016-0001.

Chapter 4

Planning Protection of Personnel: Optimizing Force Readiness in Response to Respiratory Infectious Diseases

Attribution

The following chapter is based upon the manuscript in prep: *Schlitt, J., Lewis, B., Eubank. S. (2019). Planning Protection of Personnel: Optimizing Force Readiness in Response to Respiratory Infectious Diseases.*

4.1 Abstract

The threat of novel pathological infection to defense readiness is an ancient problem faced by modern militaries. The goal of this study is to explore a multivariate intervention parameter space in order to determine efficient intervention programs. Knowledge of combined intervention efficacies may be used to inform policies that maximize military readiness during outbreaks. Simulated interventions include antiviral treatment, antiviral prophylaxis, vaccination, and sequestration. This study builds upon prior work in agent-based modeling of infectious disease by simulating the spread of a respiratory illness, influenza, through a synthetic population of a military base. The study found that rapid initiation of the intervention sequence is the most important parameter, subtly dictating the strength of each intervention and compensating for the weaknesses of other pharmaceutical interventions. While one would ideally vaccinate immediately upon the detection of an outbreak and prophylax and sequester at-risk groups for the duration of said outbreak, one's agency may be limited by the availability of effective pharmaceuticals and by external factors. This research illustrates the need for a multifaceted intervention program for managing novel influenza outbreaks. An outbreak within a military base population presents unique challenges, as the base needs to maintain readiness and cannot be shut down. Further, given that the base population is highly connected, the impact of delays in action and/or of weak pharmaceuticals are pronounced.

4.2 Introduction

The threat of novel pathological infection to defense readiness is an ancient problem faced by modern militaries. Sequestration has proved a valuable and time-tested means of protecting the world's militaries from outbreaks (Markel et al. 2006). However, in an increasingly connected world, infections may travel across the globe in a matter of hours. Therefore, the importance of a multi-tiered infection strategy becomes increasingly apparent. Generalized antivirals and strain specific vaccines may protect vulnerable individuals from getting infected (Medlock and Galvani 2009; Meltzer, Cox, and Fukuda 1999). Ring prophylaxis, or a combination of antiviral prophylaxis and sequestration of the exposed, shows further promise in reducing morbidity within highly connected groups (Marathe, Lewis, Barrett, et al. 2011; Lee et al. 2010). However, each of these methods depends upon the timely availability of pharmaceuticals and their effectiveness, the two most important factors that cannot be guaranteed when facing a novel influenza strain.

The goal of the study is to explore a vast, multivariate intervention parameter space in order to determine resource-efficient intervention programs, allowing for informed policy decisions that maximize military readiness under strained logistics and epidemiological challenges. Simulated interventions include antiviral treatment, antiviral prophylaxis, vaccination, and sequestration ***(Table 4.1)***. This study builds upon prior work in agent-based modeling of infectious disease by simulating the spread of influenza through a synthetic population of a US military base, using the simulation platform, EpiFast (K. R. Bisset et al. 2009). Examples of prior works that use synthetic populations and the EpiFast platform can be found in J. Chen, Marathe, and Marathe (2010); K. Bisset and Marathe (2009b); Eubank et al. (2004); Marathe, Lewis, Chen, et al. (2011b); and Halloran et al. (2008). Synthetic populations are a statistical representation of the individuals that are built using a combination of data sources such as the US Census, activity surveys, trip diaries, land-use information, Dun and Bradstreet data, and other public and private data sources. The disease diffusion takes place on the synthetic social network of the military base. To construct the social network, for each activity performed by a household, a preliminary assignment of a location is made based on the gravity model using the home location as an anchor. The locations of households are based on observed land-use patterns and tax data. A co-location based social network is formed when agents carry out their daily activities and mix with other agents who are present at the same locations. For details on how synthetic populations are constructed please see Beckman, Baggerly, and McKay (1996); Christopher Barrett et al. (2013); and K. Bisset and Marathe (2009).

Table 4.1: Intervention programs simulated in the experiment

Intervention	Definition
Antiviral Treatment	The delivery of antiviral drugs to self-reporting infected individuals to reduce symptoms and the odds of transmission to new individuals.
Antiviral Prophylaxis	The delivery of antiviral drugs to individuals of unknown exposure status in order to reduce the rates of transmitted and received infection.
Sequestration	Collective isolation for a group suspected to have been exposed to a contagious illness.
Vaccination	The delivery of an intravenous vaccine to uninfected individuals to promote lasting individual and herd immunity.

Below we describe a user-defined scenario in a military setting. Our goal is to find the most effective way to control an outbreak of a novel influenza strain among the military personnel. This is done via simulation of a broad set of intervention parameters to explore realistic scenarios in which the best interventions cannot be executed in a timely manner. Further, scenarios are explored based on the premise of delayed notification and prolonged rollout. This study seeks to provide both an illustrative comparison of the roles of each intervention as well as a guide for the infection control staff to make informed decisions when choosing between strategies in the face of an ongoing epidemic.

4.3 Methods

4.3.1 Simulated base population

The baseline scenario sought to model the spread of a novel, highly infectious influenza like illness (ILI) within a military base population for 60 days following the return of a brigade battalion unknowingly exposed during field exercises. To this end, we first constructed a synthetic military base population of 27,633 active duty military staff and 15,912 civilian workers on post stationed adjacent to Tacoma, Washington USA. The brigade population is a 4,000-member subset of the active duty military staff comprising 14.5% of its total number. We simulate the brigade's exposure to ILI during field exercises by seeding infections in 3% (i.e. 120) randomly selected individuals within the brigade on day one of the simulation, as suggested in the user-defined scenario.

Using EpiFast (K. R. Bisset et al. 2009) we simulate a respiratory illness, parameterized the same as influenza (Halloran et al. 2008) spreading between the active duty military, civilian workers on post, and the general civilian populace. The disease is represented by the four basic SEIR states: susceptible, exposed, infectious and removed. The duration in these states and relative levels of infectiousness and susceptibility are modified by the pharmaceuticals applied to the individuals (J. Chen, Marathe, and Marathe 2010) (Chris Barrett et al. 2011) (Yi and Marathe 2015) and their contacts are modified by changes to the individual's social network (e.g. sequestration).

The modeled scenario starts with the affected brigade involved in field exercises from days one to three where they are exposed on day one. Upon returning to the base some of the infected start exhibiting symptoms and are infectious as they resume their normal activities. On day six the extent of the outbreak has been more properly realized, and public health officers on the base begin taking action. On day seven the affected brigade is sequestered, and antiviral treatment and prophylaxis as well as vaccination campaigns follow along the dimensions described in ***table 4.2.***

4.3.2 Simulation Interventions

From day zero in the simulation, all base staff exhibiting ILI symptoms were prescribed a five-day course of antiviral treatment with a 90% compliance rate. The general intervention sequence began with the immediate sequestration of the brigade in isolation from the remaining active duty military staff. This prevented all transmission to and from the brigade staff. At the same time, antiviral prophylaxis was administered to the brigade for a period of zero or ten days with a compliance rate of 90%. Antiviral treatment and prophylaxis used the same regimens and were identical in effectiveness, reducing the transmission to and from an individual by the efficacy of the antiviral (30% or 70%). The entire base was vaccinated on day 7, 14, 21, or 28 to account for possible delays in obtaining the vaccine. In addition, the possible delay in the efficacy of the vaccine was considered to be 14 days to account for antigen recognition by the host immune system and a compliance rate of 90% was assumed. Vaccine efficacy values of 30% and 70% were simulated.

Due to the logistics of mass pharmaceutical rollout and the limited number of medical staff, the antiviral prophylaxis of the brigade and vaccination of the active duty military staff were both set to require five days to complete. Finally, as the civilian population was left to manage their outbreak without outside assistance, no interventions were triggered external to the base until day 50 when public schools were shut down.

Table 4.2: Simulated intervention program parameters

Intervention	Trigger Delay (days)	Days to Apply	Compliance	Efficacy	Duration (days)	Population
Antiviral Treatment	0	1	90%	30%, 70%	5	Active Duty Military
Antiviral Prophylaxis	7	5	90%	30%, 70%	0,10	Brigade
Sequestration	7	1	100%	100%	0,7,14	Brigade
Vaccination	7,14,21, 28	5	90%	(30%,70%)+ 14 days	Permanent	Active Duty Military

Table 2 legend: *Each intervention is modeled with variations in efficacy, duration, and/or delay to explore effects of changes in each parameter. Interventions were split between the prophylaxis and sequestration of the exposed brigade and the larger treatment and vaccination programs administered to active duty military staff.*

4.3.3 Sensitivity Analysis

Results were analyzed with respect to the count of cumulative infections by day within active duty military personnel. The peak daily infection count was used as a measure of the strain on the base healthcare system. In total, 2,400 simulations were run covering eight combinations of vaccine efficacy and delay, four combinations of antiviral efficacy and prophylaxis duration, and three sequestration durations for a total of 96 experimental cells. For each cell 25 replicates were run and their mean numbers are reported here.

4.4 Results

4.4.1 Infections by intervention program

The mean count of daily infections within the active duty military population for 25 iterations was selected as a metric for analysis. ***Table 4.3*** provides a summary of interventions, their description, abbreviations, and associated parameters. ***Figure 4.1*** provides a heat map of total and peak infections under the parameters stated in ***table 4.3.***

Figure 4.1: Total and Peak Daily Infections by Intervention Program and Pharmaceutical Efficacy

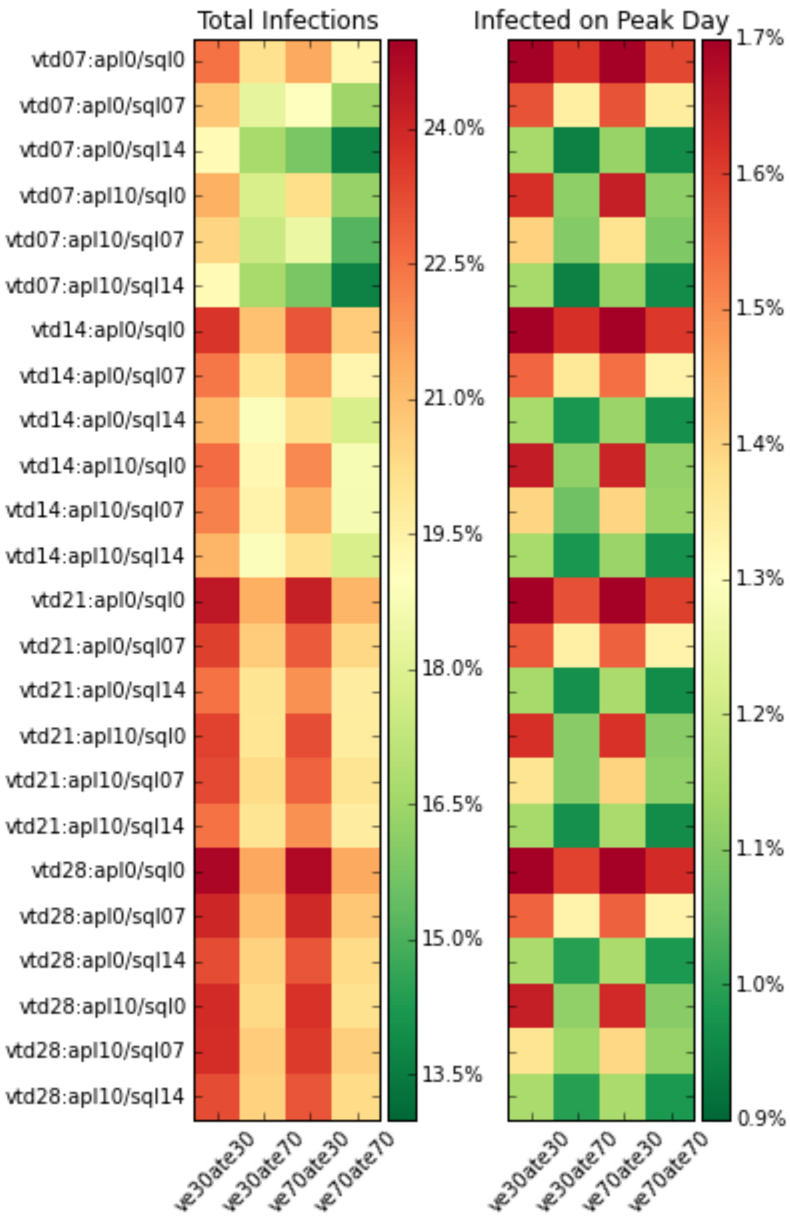


Figure 4.1 legend: *Total and peak infections by different interventions and their respective parameters: in the above figure, the rows are used to describe directly controlled, logistical parameters such as delays and program durations, whereas pharmaceutical efficacy parameters are described in columns. Combinations of high vaccine efficacy and short delays were useful in suppressing the total number of*

infections. However, they had no effect upon peak infections, as the vaccines did not confer resistance until after the height of the epidemic

Table 4.3: Interventions used in Figure 4.1

Parameter	Abbreviation	Value(s)	Description
Antiviral Prophylaxis Length	apl	(0,10) days	Duration for which antiviral prophylaxis is maintained
Antiviral Prophylaxis Trigger Delay	aptd	7 days	Day on which antiviral prophylaxis begins
Antiviral Treatment Efficacy	ate	30%, 70%	Percent reduction in the rate of received and transmitted infections for individuals taking antivirals
Antiviral Treatment Length	atl	5 days	Duration for which antiviral treatment is given
Sequestration Trigger Delay	sqtd	7 days	Day on which sequestration begins
Sequestration Length	sql	(0,7,14) days	Duration for which the brigade is sequestered
Vaccination Efficacy	ve	(30%,70%)+14 days	Percent reduction in the rate of received infections for vaccinated individuals
Vaccination Trigger Delay	vtd	(7,14,21,28) days	Day on which the vaccination is administered

4.4.2 Short Sequestration Analysis

As a secondary scenario, a subset of the experiment was selected where sequestration length was held at one and two weeks, as shown in ***table 4.4***. This was done to provide a reasonable scenario in which the base staff could maintain a sufficient standard of readiness while accepting some setbacks due to caseload. Next a sequestration length of one week was selected as the one most likely to represent a practical scenario given the logistical burdens imposed by sequestering a large percentage of the military base staff. Within the cells where sequestration length equaled seven days (from simulation day seven until simulation day 14), scenarios with the lowest active duty military attack rate per a given constraint were selected in order to explore the best strategies where low vaccine and/or antiviral pharmaceutical efficacies could not adequately reduce transmission among individuals. ***Figure 4.2*** displays the infection results for a one-week sequestration under different scenarios.

Table 4.4: Mean number of active duty military infected for seven and 14 day sequestration scenarios

	7 Day Sequestration			14 Day Sequestration		
Scenario	Number Infected	Attack Rate	Percent Reduction	Number Infected	Attack Rate	Percent Reduction
Control or Base Case	7,792	28.2%	0.0%	7,792	28.2%	0.0%
Worst Possible Outcome	6,652	24.1%	14.6%	6,425	23.2%	17.5%
Best Possible Outcome	4,190	15.2%	46.2%	3,779	13.7%	51.5%
Best Outcome Under Weak Vaccine	4,831	17.5%	38.0%	4,617	16.7%	40.7%
Best Outcome Under Weak Antiviral	5,075	18.4%	34.9%	4,371	15.8%	43.9%
Best Outcome Under Weak Vaccine and AV	5,652	20.4%	27.5%	5,308	19.2%	31.9%

Table 4.4 legend: *In the control scenario no interventions were implemented on the base, though the outside community closed schools on day 50. The best outcomes were able to reduce roughly half of the total infections from the control scenario.*

Figure 4.2: Best Daily Infection Curves for Weak Pharmaceutical Efficacy

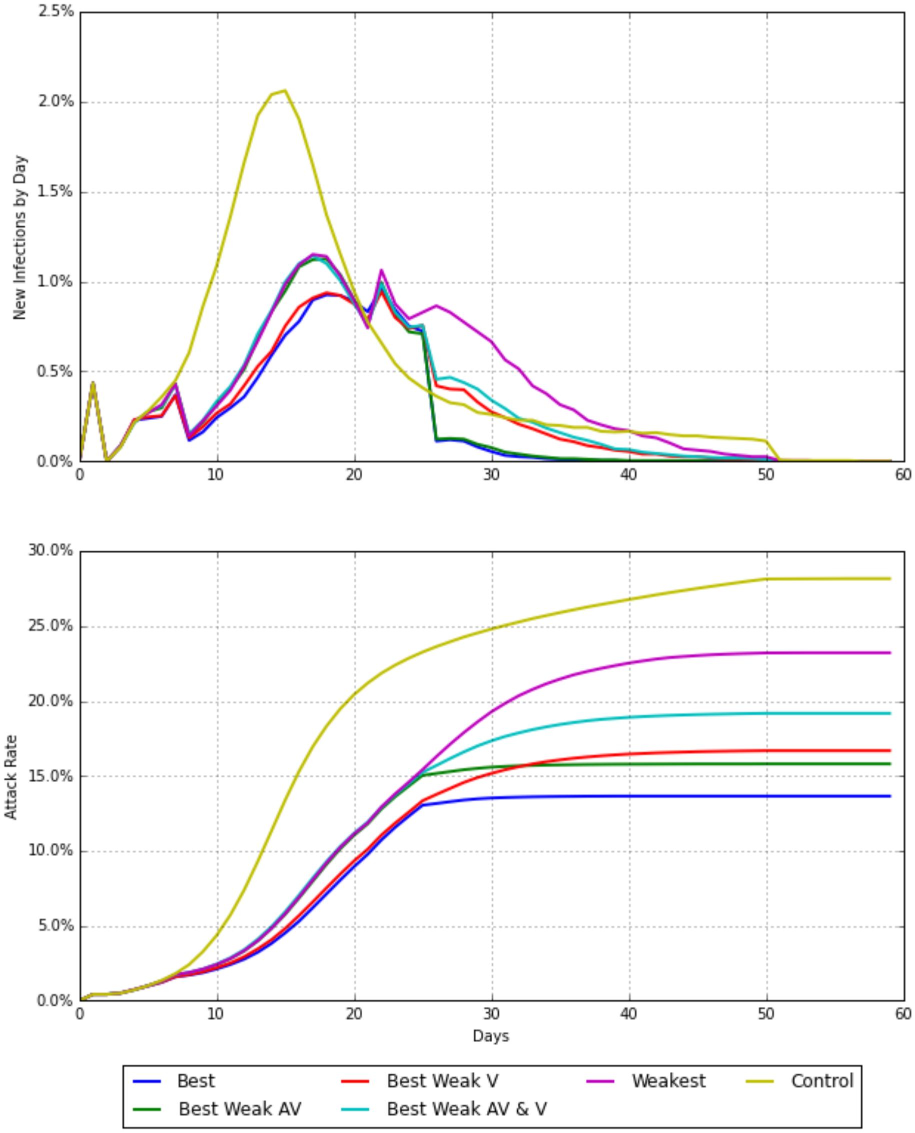


Figure 4.2 legend: *Number of active duty military infections for a seven day sequestration program: AV refers to antiviral and V to vaccination. Best implies the best possible scenario among all the scenarios considered. Best Weak AV represents the best scenario given a poorly efficacious antiviral; Best Weak V is the best scenario given a poorly efficacious vaccine; Best Weak AV & V is the best scenario given a poorly efficacious vaccine and antiviral; and Weakest is the weakest scenario among all scenarios considered. Notable inflections in daily infections follow the initial seeded infections on day one, the five day rollouts of vaccination and antivirals beginning on day*

seven, and the resumption of susceptibility among the brigade following the completion of their antiviral regimen on day 22.

4.4.3 Prolonged Sequestration Analysis

While this paper focused primarily upon short sequestration length interventions, it is important to consider the benefits of prolonged sequestration when situations allow. To further explore the effects of sequestration length, a comparison of attack rates by sequestration length and intervention strengths was conducted. Cells were selected for the strongest and weakest antiviral and vaccine interventions for each possible sequestration length. In the best program, the vaccine was 70% effective from day 21, the antivirals were 70% effective in stopping infection and transmission by treated individuals, and ten-day antiviral prophylaxis was given. In the worst program, the vaccine was 30% effective from day 42, the antivirals were 30% effective in stopping infection and transmission by treated individuals, and no prophylaxis was given.

Figure 4.3: Best and Worst Outcomes by Length of Sequestration

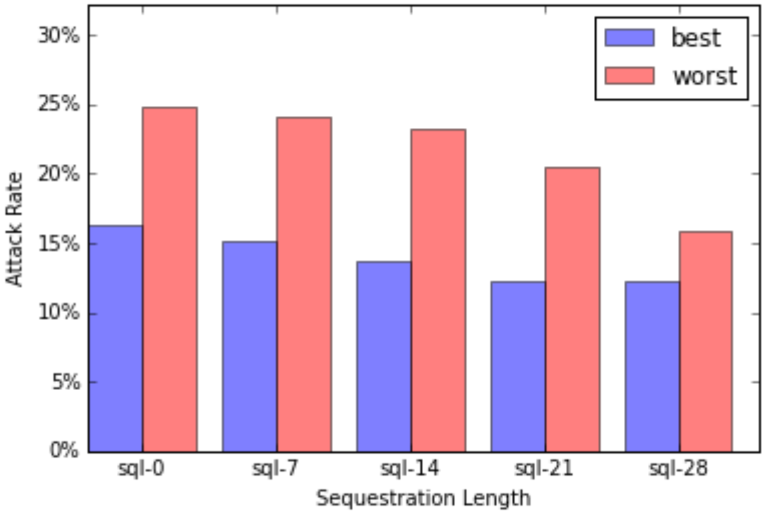


Figure 3: *Effects of sequestration length upon number infected with best and worst intervention: best and worst interventions were selected for antiviral prophylaxis,*

antiviral treatment, and vaccination. Strong interventions were selected by run parameters, where strong interventions held the highest pharmaceutical efficacies, the earliest vaccination, and the longest prophylaxis. In contrast, weak interventions were selected such that they held the weakest pharmaceutical efficacies, severely delayed vaccine rollouts, and short prophylaxis. The best and worst programs show a decrease in the attack rates with long sequestration of 25.3% and 35.8% respectively.

4.5 Discussion

4.5.1 Infections by intervention program trends

From ***figure 4.1***, several trends become immediately apparent. First, each intervention helps in reducing the cumulative infection count. Effective, earlier, and longer interventions result in beneficial outcomes with no one intervention able to function well in isolation. Among these, antiviral efficacy shows the greatest effect in reducing the attack rate due to the high level of compliance and immediate availability of treatment. Next, the number infected on the peak day showed limited variability with all observable effects coming from the sequestration length and the antiviral efficacy. This is intuitive, as given the two-week delay in vaccine efficacy and one-week delay in administration, there is no scenario where the vaccine becomes effective before the outbreak peaks between days 17–19. Further, the sequestration helps the traveling brigade and not the active duty military at large because, upon returning to the base, the brigade has four days to seed infections within the non-brigade active duty military staff before the sequestration is initiated.

Increased vaccine efficacy was only shown to significantly decrease the total number of active duty military infections if the vaccine was administered with a seven or 14-day trigger delay. Increasing the vaccine efficacy from 30% to 70% resulted in approximately three and one percentage point decreases in the total number of active duty military infected for the seven and fourteen-day sequestration scenarios respectively. This outcome is largely a product of the assumptions of delayed vaccination, delayed vaccine efficacy, and a leaky vaccine that reduced transmission rates rather than preventing transmission entirely. An alternate study scenario with an earlier vaccine rollout, earlier vaccine efficacy, or a previously administered and efficacious vaccine would show a much stronger impact of vaccine efficacy on the number of active duty military infections. The reduction in active duty military infections would also have been greater had we assumed a completely effective vaccination. Contrarily, as the 21 and 28 day vaccine trigger delay scenarios showed little difference in outcomes between the 30% and 70% effective vaccine scenarios, the assumption of a leaky vaccine had limited effect on simulation outcomes for longer vaccine trigger delays as the epidemic had largely run its course before the vaccine became effective. Finally, we see that the late, 28 day vaccine trigger delay yielded the highest of all attack rates among active duty military, reaching 14.1% in a best intervention scenario with 28 day sequestration length, 70% vaccine efficacy, 70% antiviral efficacy, and ten-day antiviral prophylaxis and 24.8% in a worst-intervention scenario with no sequestration, 30% vaccine efficacy, 30% antiviral efficacy, and no antiviral prophylaxis. In this scenario, prolonging antiviral prophylaxis for 42 days until the vaccine is administered and takes effect may mitigate some of the increased infections from delayed vaccination. Ultimately, the number infected will be greater given the practical limitations of antiviral supplies and compliance.

4.5.2 Short Sequestration Analysis

Comparing the lowest attack rate scenarios for weak, 30% effective vaccinations and antivirals in ***figure 4.2*** showed that strong antivirals have a greater effect in reducing the number of infected active duty military. It was observed that the vaccine and antiviral prophylaxis did not significantly alter the course of the epidemic until the completion of their respective five-day rollouts. Intervention strategies were algorithmically selected from the pool of scenarios for the lowest active duty attack rates given 30% antiviral efficacy, 30% vaccine efficacy, or 30% vaccine and antiviral efficacy. The selected strategies followed expected patterns in that strong, early vaccination was favored.

The best scenario given low antiviral efficacy yielded a cumulative and peak infection count of 5,075 and 379 active duty military cases respectively. The best scenario given low vaccine efficacy yielded 4,831 cumulative and 304 peak daily infections among active duty military. The peak infection count was found to range from 429 for the weakest intervention sequence to 302 for the best, showing a reduction of 29.6%. Likewise the cumulative infection rates were found to range from 6,652 infections for the weakest intervention to 4,190 for the best intervention with a 37.0% reduction in infections. This suggests that, while the total burden over time may be reduced by the best intervention, strategies will still be needed to manage peak burden as no strategy could effectively shift the peak burden by more than five to six days as compared to the base case. In each simulation, the mean peak day was found to fall between 17 and 22 days of the initial exposure.

Looking closely at the selected epidemic curves, it may be observed that sequestration serves chiefly to delay the onset of the epidemic, buying time before lasting pharmaceutical interventions could be administered. However, sequestration alone cannot be expected to halt a large-scale outbreak as delays in the detection and diagnosis of cases may allow for significant spread of infection outside of the sequestered group before the intervention sequence is initiated. The highest, 70% antiviral efficacy reduces the transmission rate during the early epidemic, significantly decreasing the cumulative infections and peak daily infections. Given immediate vaccination from an existing stockpile on day seven, a 14 day delay in vaccine efficacy, and a five day rollout of the vaccination program, vaccination is never able to show a beneficial effect before the peak epidemic has passed. Though the first vaccinations become effective on day 21, the courses of each strong and weak vaccine epidemic for a given antiviral efficacy do not diverge until day 25. This follows the course of the vaccine rollout program where all vaccines act as the key guard against further infection following the termination of antiviral prophylaxis.

4.5.3 Prolonged Sequestration Analysis

For long, three to four week sequestrations, vaccine efficacy delays become significant, yielding a substantial increase in the number of infected for delays of two weeks or longer. For sequestrations of one to two weeks, the previous pattern is maintained where antiviral prophylaxis length held the greatest significance, with effective antivirals most important for long prophylaxis, and early vaccination most important for short prophylaxis. As seen in ***figure 4.3***, moderate duration sequestration is of limited benefit to weak interventions. Poor pharmaceutical effectiveness against a given strain or logistical limitations significantly reduced the efficacy of pharmaceutical interventions. However, longer durations lead to significant improvement in curtailing infections. For the best intervention, the benefits of longer sequestration length plateaued after 21 days. The best and worst programs showed a decrease in the attack rates from zero to 28-day sequestration of 25.3% and 35.8% respectively.

4.6 Conclusions

This research illustrates the need for a multifaceted intervention program for managing the outbreak of a novel influenza strain. An outbreak within a military base population presents unique challenges, as the base needs to maintain a standard of readiness and cannot simply be shut down as one may with a school or office place. Further, given that the base population is highly connected, it exacerbates the impact of delays in action and/or of poor quality pharmaceuticals.

Across all experiments, timely action is the most important parameter, subtly dictating the relative strength of each intervention as well as compensating for the weaknesses of other interventions. One important aspect of this experiment is the consideration of a range of parameters, some of which can be directly controlled by the regulators and policy makers such as the decision to sequester and its duration, and others that are outside their control, such as the delay in vaccine efficacy. While one would ideally vaccinate early, prophylax maximally, and sequester long, one’s agency may be limited by the availability of pharmaceuticals and the presence of ongoing conflicts or emergencies. Further, existing pharmaceuticals may have little or no effect upon the transmission of a novel ILI, necessitating an alternate strategy and a realistic appraisal of likely outcomes. This experiment has been able to successfully demonstrate the outcomes of the epidemic under a variety of scenarios that the user considered realistic.

4.7 Funding

This work has been partially supported by the NIH’s National Institute of General Medical Sciences grants 5U01GM070694-13 and 1R01GM109718 as well as Defense Threat Reduction Agency Comprehensive National Incident Management System Contract HDTRA1-11-D-0016-0001.

Chapter 5

Conclusions

5.1 Summary

In **Chapter 1** we discussed the role of emerging trends and novel resources in outbreak response as the open data, open source, and open science movements have together enabled new research paradigms leveraging the broader skills of the academic community. We extended the business concept of the Time Value of Knowledge to epidemiology to capitalize on the opportunity these resources provide. This extension was enabled by the natural analogy of competition between the spread of an emerging infectious disease and the labors of public health professionals involved in outbreak response to business competition. Bringing these together, we proposed the **C4 Response Model** for guiding the academic response to outbreaks of infectious disease as stakeholders incrementally ***collect, connect, calibrate,*** and ***convey*** critical knowledge across the early, mid, and late time-domains of an epidemic. Under the **C4 Response Model,** collaborating researchers first seek to **collect** cleaned, organized data into open repositories for public access across disciplines. This collection effort feeds directly into the **connection** and **calibration** tasks. During the **connection** task, researchers seek trends and correlations to better describe a novel epidemic in contrast to similar, prior epidemics in search of intervention targets. In parallel, the **calibration** task seeks to fit pre-existing models to the knowledge **collected** from a novel outbreak in order to predict the course of epidemics and find key points of differentiation from prior epidemics. Finally, these efforts culminate in the **convey** task, whereby **collected, connected,** and **calibrated** knowledge is **conveyed** to the appropriate audience at the appropriate level to drive policy interventions and health-positive behavioral changes. This **convey** task is the most critical of all within the **C4 Response Model** as without proper advocacy and message crafting, even the most exciting findings may flounder in obscurity.

In **Chapter 2** we saw that open science papers are neither published nor referenced earlier than closed access papers during the course of an epidemic. The data showed a broad trend of prestigious, closed access journals publishing novel, high-value findings closer to the start of an epidemic. Closed access journals also referenced more recently published articles than open access journals. However, open science has expanded its reach over time as seen in both the percentages of open science papers published and open science articles referenced by year. While closed science primary papers had a relatively earlier impact within each given epidemic, open science articles and references held a distinct and growing numerical advantage across the study period. Each epidemic showed a notable sequence in the volume of citations by research methodology over time, starting with the fine-level, wet laboratory sciences such as genetics and virology, and working up to broader, population-level epidemiological studies. These findings show a promising and irreversible niche for open access science within the international outbreak response community as well as the differential importance of scientific disciplines over time. Knowledge of such may be used to more effectively allocate funding and resources across the time-domains of future epidemics as each discipline plays its role in charting effective interventions. These findings provided a bird's-eye view of recent trends in outbreak response, as the early knowledge **collection** efforts of the laboratory sciences laid the groundwork which increasingly complex knowledge-building efforts later **connected** and **calibrated**.

In **Chapter 3** we demonstrated a viable method for improving organizational strategies in message-crafting in response to the opioid epidemic in the United States. We explored tweet virality as a function of message content and profile source, time and day of week, and audience-perceptible aspects of a tweet's profile source such as tone, geography, and suspected bot status. We showed unique advantages to posting tweets ahead of weekly and daily periods of high tweet volume to ensure dissemination. We explored the challenges in sustaining engagement over time and across a targeted geographic region as message engagement plummeted after mere hours and as retweets showed little true affinity for their point of origin. Combining these findings, a public health agency may now select optimal outlets for **conveying** specific messages in response to crises such as the presence of contaminated drugs within a community.

Further, this method is readily generalizable, albeit dependent on manual classification. Provided a suitable corpus of tweets, the tools developed herein could be applied to numerous societal problems where institutional risk messaging plays a role.

Finally, in **Chapter 4,** we explored the interplay between the timing and potency of pharmaceutical and non-pharmaceutical interventions in reducing the threat of novel influenza to a highly connected military base population. To do this, we leveraged mass data **collected** from previous studies to project a **calibrated,** agent-based model of influenza transmission. We assessed the burden of disease in terms of both the cumulative cases as well as the peak number of daily infections, as shorter, higher-intensity epidemics may pose a greater stressor to health systems and readiness than longer-burning epidemic with similar cumulative case counts. Under logistically feasible intervention programs, timely response was shown to play a much stronger role than the strength of pharmaceuticals in reducing the overall burden of disease for a given epidemic. We observed that the sequestration of infected cases served only to delay the onset of an epidemic, buying time for the application of pharmaceutical interventions but not replacing them in function. While strong combined interventions were able to reduce both the peak daily infection counts as well as the cumulative infection rates for simulated epidemics, they could neither prevent the majority of cases nor substantially postpone the day of peak infections once a significant number of cases had been introduced. This demonstrates the unique threat to readiness posed by novel influenza and ILI within a military base and other such highly connected populations.

5.2 Methodological Contributions

In **Chapter 2** we developed a novel system for assessing the impact of open access science within a given field using publicly available data and computational resources. These methods may be applied across disciplines and crisis, black swan events, or otherwise notable occurrences of interest to characterize the structure of the academic response to novel challenges. We present this for public consumption in the research analysis notebook linked in ***Appendix A.1.*** In pursuit of this goal, we also developed an algorithmic means of identifying and visualizing communities of interest within complex networks of interconnecting labels. Such labels at present include study research methods, grants, keywords, and funding bodies, though additional data sources such as semantic networks, the dispersal of illicit funds, or emergency communications bear consideration. This tool is rendered available for the public under the Creative Commons **CC BY-NC** license for non-commercial use with attribution via the public sandbox linked in ***Appendix A.1.*** Sample visualizations are also presented in ***Appendices A.2, A.3, and A4.***

In **Chapter 3** we formalized the application of ChatterGrabber for public health message mapping applications. We developed a system of tools and interfaces to collect clean, organized tweet data and to coordinate the tasks of labeling tweet messages and profile sources. In addition, we developed an analysis pipeline for cleaning the resulting labeled data and conducting rich virality analyses automatically. We humbly offer ChatterGrabber for use by the broader scientific community, and have worked to support this use via publicly available code repositories, workshops, and virtual machine images via ***Appendices B.1, B.2, and B.3.*** ChatterGrabber has long served as both an alarm system and an analytical data **collection** system for the local, state, federal, and international public health community. We hope that by creating such a pipeline it may also see greater use in developing strategies for **conveying** messages for both long-term health campaigns and urgent dispatches to elusive populations in the face of novel threats.

Lastly, in **Chapter 4** we developed a novel front end to the EpiFast agent-based model. This provided a simple yet highly functional collaborative environment specifically targeted at crafting scenarios accommodating for logistical delays. We also developed a visualization pipeline capable of cleanly organizing large scale, multidimensional data sweeps and identifying key inflection points via the application of correlation and regression tree analysis as seen in ***Appendices C.1 and C.2.*** Finally, and perhaps most uniquely, we developed a method for analyzing EpiFast simulation data to extract the transmission rates over time between arbitrary subpopulations as well as the effective reproductive number of infected individuals within each such population over time. While the sheer number of variable sweeps conducted largely enabled the former correlation and regression tree analysis, the latter subpopulation-specific transmission visualization leaned upon unique opportunities provided by the ABM. Such methods may prove invaluable for future work exploring the impact of ILI within highly connected populations.

Bibliography

"2018 Slang Terms and Code Words." n.d. Accessed March 14, 2019. https://www.dea.gov/documents/2018/07/01/2018-slang-terms-and-code-words.

Abuse, Opioid, and Opioid Overdose. 2016. "Community Announcements." https://stacks.cdc.gov/view/cdc/42237/cdc_42237_DS1.pdf.

Allen, Rachel. n.d. "Official Newsletter of the Utah Poison Control Center." *Poisoncontrol.utah.edu*. https://poisoncontrol.utah.edu/newsletters/pdfs/toxicology-today-archive/Vol18_Iss2.pdf.

Azhar, Esam I., Sherif A. El-Kafrawy, Suha A. Farraj, Ahmed M. Hassan, Muneera S. Al-Saeed, Anwar M. Hashem, and Tariq A. Madani. 2014. "Evidence for Camel-to-Human Transmission of MERS Coronavirus." *The New England Journal of Medicine* 370 (26): 2499–2505.

Barrett, Chris, Keith Bisset, Jonathan Leidig, Achla Marathe, and Madhav Marathe. 2011. "Economic and Social Impact of Influenza Mitigation Strategies by Demographic Class." *Epidemics* 3 (1): 19–31.

Barrett, Christopher, Keith Bisset, Shridhar Chandan, Jiangzhuo Chen, Youngyun Chungbaek, Stephen Eubank, Yaman Evrenosoglu, et al. 2013. "Planning and Response in the Aftermath of a Large Crisis: An Agent-Based Informatics Framework." *2013 Winter Simulations Conference (WSC)*. https://doi.org/10.1109/wsc.2013.6721535.

Beckman, Richard J., Keith A. Baggerly, and Michael D. McKay. 1996. "Creating Synthetic Baseline Populations." *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/0965-8564(96)00004-3.

Bisset, Keith, and Madhav Marathe. 2009. "A Cyber-Environment to Support Pandemic Planning and Response." *DOE SciDAC Magazine* 13: 36–47.

Bisset, Keith R., Jiangzhuo Chen, Xizhou Feng, V. S. Anil Kumar, and Madhav V. Marathe. 2009. "EpiFast." In *Proceedings of the 23rd International Conference on Conference on Supercomputing - ICS '09*. https://doi.org/10.1145/1542275.1542336.

Blancou, Jean, Bruno B. Chomel, Albino Belotto, and Francois Xavier Meslin. 2005. "Emerging or Re-Emerging Bacterial Zoonoses: Factors of Emergence, Surveillance and Control." *Veterinary Research* 36 (3): 507–22.

CDC. 2015. "Investigation of Undetermined Risk Factors for Suicide Among Youth, Ages 10–24 —Fairfax County, VA, 2014." presented at the Exit Briefing, June 23. https://www.fairfaxcounty.gov/health/sites/health/files/Assets/images/suicide-epi-aid-exit-briefing.pdf.

"CDC SARS Response Timeline | About | CDC." 2018. July 18, 2018. https://www.cdc.gov/about/history/sars/timeline.htm.

Centers for Disease Control and Prevention (CDC). 2003. "Multistate Outbreak of Monkeypox--Illinois, Indiana, and Wisconsin, 2003." *MMWR. Morbidity and Mortality Weekly Report* 52 (23): 537–40.

———. 2011. "Vital Signs: Overdoses of Prescription Opioid Pain Relievers---United States, 1999--2008." *MMWR. Morbidity and Mortality Weekly Report* 60 (43): 1487–92.

Cervellin G Comelli I Lippi. 2017. "Is Google Trends a Reliable Tool for Digital Epidemiology? Insights from Different Clinical Settings." *Journal of Epidemiology and Global Health* 7 (3): 185–89.

Chadha, Mandeep S., James A. Comer, Luis Lowe, Paul A. Rota, Pierre E. Rollin, William J. Bellini, Thomas G. Ksiazek, and Akhilesh Mishra. 2006. "Nipah Virus-Associated Encephalitis Outbreak, Siliguri, India." *Emerging Infectious Diseases* 12 (2): 235–40.

Chen, Jiangzhuo, Achla Marathe, and Madhav Marathe. 2010. “Coevolution of Epidemics, Social Networks, and Individual Behavior: A Case Study.” *Advances in Social Computing*. https://doi.org/10.1007/978-3-642-12079-4_28.

Chen, Yu, Weifeng Liang, Shigui Yang, Nanping Wu, Hainv Gao, Jifang Sheng, Hangping Yao, et al. 2013. “Human Infections with the Emerging Avian Influenza A H7N9 Virus from Wet Market Poultry: Clinical Analysis and Characterisation of Viral Genome.” *The Lancet* 381 (9881): 1916–25.

Chretien, Jean-Paul, Caitlin M. Rivers, and Michael A. Johansson. 2016. “Make Data Sharing Routine to Prepare for Public Health Emergencies.” *PLoS Medicine* 13 (8): e1002109.

Clauset, Aaron, M. E. J. Newman, and Cristopher Moore. 2004. “Finding Community Structure in Very Large Networks.” *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 70 (6 Pt 2): 066111.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics* 25 (11): 1422–23.

Coleman, Christopher M., and Matthew B. Frieman. 2013. “Emergence of the Middle East Respiratory Syndrome Coronavirus.” *PLoS Pathogens* 9 (9): e1003595.

Curran, Kathryn G., James J. Gibson, Dennis Marke, Victor Caulker, John Bomeh, John T. Redd, Sudhir Bunga, Joan Brunkard, and Peter H. Kilmarx. 2016. “Cluster of Ebola Virus Disease Linked to a Single Funeral - Moyamba District, Sierra Leone, 2014.” *MMWR. Morbidity and Mortality Weekly Report* 65 (8): 202–5.

Dalmaris, Peter, William P. Hall, and W. R. Philp. 2006. “The Time-Value of Knowledge: A Temporal Qualification of Knowledge, Its Issues, and Role in the Improvement of Knowledge Intense Business Processes.” In *Proceedings of the 3rd Asia-Pacific International Conference on Knowledge Management (KMAP06)*. https://www.researchgate.net/profile/Wayne_Philp2/publication/235703221_The_time-value_of_knowledge_a_temporal_qualification_of_knowledge_its_issues_and_role_in_the_improvement_of_knowledge_intense_business_processes/links/02e7e531c03a242576000000.pdf.

“DanceSafe.” n.d. Accessed March 14, 2019. https://twitter.com/DanceSafe.

Davis, C. A., O. Varol, E. Ferrara, and A. Flammini. 2016. “Botornot: A System to Evaluate Social Bots.” *Proceedings of the 25th*. https://dl.acm.org/citation.cfm?id=2889302.

Dicker, Richard, Fatima Coronado, Denise Koo, and R. Gibson Parrish. 2006. “Principles of Epidemiology in Public Health Practice.” *Atlanta GA: US Department of Health and Human Services*. https://rtyhvqtnb01.storage.googleapis.com/QjAwNVY5OUg0Sw==01.pdf.

Ebrahim, Shah, and George Davey Smith. 2018. “Who Needs Editors? The Epidemiology of Publications in the IJE.” *International Journal of Epidemiology*, July. https://doi.org/10.1093/ije/dyy143.

“Epidemics | The RAPIDD Ebola Forecasting Challenge | ScienceDirect.com.” n.d. Accessed April 2, 2019. https://www.sciencedirect.com/journal/epidemics/vol/22/suppl/C.

Eubank, Stephen, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. 2004. “Modelling Disease Outbreaks in Realistic Urban Social Networks.” *Nature*. https://doi.org/10.1038/nature02541.

“Face the Fentanyl.” n.d. Accessed March 14, 2019. https://twitter.com/facethefentanyl.

Faye, Ousmane, Pierre-Yves Boëlle, Emmanuel Heleze, Oumar Faye, Cheikh Loucoubar, N ’faly Magassouba, Barré Soropogui, et al. 2015. “Chains of Transmission and Control of Ebola Virus Disease in Conakry, Guinea, in 2014: An Observational Study.” *The Lancet Infectious Diseases* 15 (3): 320–26.

Fiore, B. J., L. P. Hanrahan, and H. A. Anderson. 1990. “State Health Department Response to Disease Cluster Reports: A Protocol for Investigation.” *American Journal of Epidemiology* 132 (1 Suppl): S14–22.

Glik, Deborah C. 2007. "Risk Communication for Public Health Emergencies." *Annual Review of Public Health* 28 (1): 33–54.

Gramlich, John. n.d. "How We Identified Bots on Twitter." Pew Research Center. Accessed March 14, 2019. http://www.pewresearch.org/fact-tank/2018/04/19/qa-how-pew-research-center-identified-bots-on-twitter/.

Gurin, Joel. 2014. "Open Governments, Open Data: A New Lever for Transparency, Citizen Engagement, and Economic Growth." *SAIS Review of International Affairs* 34 (1): 71–82.

Hagberg, Aric, Pieter Swart, and Daniel S Chult. 2008. "Exploring Network Structure, Dynamics, and Function Using Networkx." LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab. (LANL), Los Alamos, NM (United States). https://www.osti.gov/biblio/960616.

Halloran, M. E., N. M. Ferguson, S. Eubank, I. M. Longini, D. A. T. Cummings, B. Lewis, S. Xu, et al. 2008. "Modeling Targeted Layered Containment of an Influenza Pandemic in the United States." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0706849105.

Henry, Maria, Lorraine Francis, Virginia Asin, Karen Polson-Edwards, and Babatunde Olowokure. 2017. "Chikungunya Virus Outbreak in Sint Maarten, 2013-2014." *Revista Panamericana de Salud Publica = Pan American Journal of Public Health* 41 (August): e61.

Hoots, Brooke E., Likang Xu, Mbabazi Kariisa, Nana Otoo Wilson, Rose A. Rudd, Lawrence Scholl, Lyna Schieber, and Puja Seth. 2018. "2018 Annual Surveillance Report of Drug-Related Risks and Outcomes--United States." https://stacks.cdc.gov/view/cdc/58547/cdc_58547_DS1.pdf.

Hutto, Clayton J., and Eric Gilbert. 2014. "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In *Eighth International AAAI Conference on Weblogs and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109.

Jeffs, Benjamin, Paul Roddy, David Weatherill, Olimpia de la Rosa, Claire Dorion, Marta Iscla, Isabel Grovas, et al. 2007. "The Medecins Sans Frontieres Intervention in the Marburg Hemorrhagic Fever Epidemic, Uige, Angola, 2005. I. Lessons Learned in the Hospital." *The Journal of Infectious Diseases* 196 Suppl 2 (November): S154–61.

Jones, Christopher M. 2013. "Heroin Use and Heroin Use Risk Behaviors among Nonmedical Users of Prescription Opioid Pain Relievers - United States, 2002-2004 and 2008-2010." *Drug and Alcohol Dependence* 132 (1-2): 95–100.

Kang, Gloria J., Sinclair R. Ewing-Nelson, Lauren Mackey, James T. Schlitt, Achla Marathe, Kaja M. Abbas, and Samarth Swarup. 2017. "Semantic Network Analysis of Vaccine Sentiment in Online Social Media." *Vaccine* 35 (29): 3621–38.

Kuhlman, C. J., Y. Ren, B. Lewis, and J. Schlitt. 2017. "Hybrid Agent-Based Modeling of Zika in the United States." In *2017 Winter Simulation Conference (WSC)*, 1085–96.

Lee, Vernon J., Jonathan Yap, Alex R. Cook, Mark I. Chen, Joshua K. Tay, Boon Huan Tan, Jin Phang Loh, et al. 2010. "Oseltamivir Ring Prophylaxis for Containment of 2009 H1N1 Influenza Outbreaks." *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa0908482.

Majumder, Maimuna S., Caitlin Rivers, Eric Lofgren, and David Fisman. 2014. "Estimation of MERS-Coronavirus Reproductive Number and Case Fatality Rate for the Spring 2014 Saudi Arabia Outbreak: Insights from Publicly Available Data." *PLoS Currents* 6 (December). https://doi.org/10.1371/currents.outbreaks.98d2f8f3382d84f390736cd5f5fe133c.

Marathe, Achla, Bryan Lewis, Christopher Barrett, Jiangzhuo Chen, Madhav Marathe, Stephen Eubank, and Yifei Ma. 2011. "Comparing Effectiveness of Top-Down and Bottom-Up Strategies in Containing Influenza." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0025149.

Marathe, Achla, Bryan Lewis, Jiangzhuo Chen, and Stephen Eubank. 2011. "Sensitivity of

Household Transmission to Household Contact Structure and Size." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0022461.

Markel, Howard, Alexandra Stern, J. Navarro, Joseph Michalsen, Arnold Monto, and Cleto DiGiovanni. 2006. "Nonpharmaceutical Influenza Mitigation Strategies, US Communities, 1918–1920 Pandemic." *Emerging Infectious Diseases*. https://doi.org/10.3201/eid1212.060506.

Medlock, J., and A. P. Galvani. 2009. "Optimizing Influenza Vaccine Distribution." *Science*. https://doi.org/10.1126/science.1175570.

Meltzer, Martin I., Nancy J. Cox, and Keiji Fukuda. 1999. "The Economic Impact of Pandemic Influenza in the United States: Priorities for Intervention." *Emerging Infectious Diseases*. https://doi.org/10.3201/eid0505.990507.

Mohr, Amanda L. A., Melissa Friscia, Donna Papsun, Sherri L. Kacinko, David Buzby, and Barry K. Logan. 2016. "Analysis of Novel Synthetic Opioids U-47700, U-50488 and Furanyl Fentanyl by LC-MS/MS in Postmortem Casework." *Journal of Analytical Toxicology* 40 (9): 709–17.

Murhekar, Manoj, Ron Moolenaar, Yvan Hutin, and Claire Broome. 2009. "Investigating Outbreaks: Practical Guidance in the Indian Scenario." *The National Medical Journal of India* 22 (5): 252–56.

Parikh, Nidhi, Samarth Swarup, Paula E. Stretz, Caitlin M. Rivers, Bryan L. Lewis, Madhav V. Marathe, Stephen G. Eubank, Christopher L. Barrett, Kristian Lum, and Youngyun Chungbaek. 2013. "Modeling Human Behavior in the Aftermath of a Hypothetical Improvised Nuclear Detonation." In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, 949–56. AAMAS '13. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Pentz, Ed. 2001. "CrossRef: A Collaborative Linking Network." *Issues in Science and Technology Librarianship* 10: F4CR5RBK.

Pergolizzi, Joseph V., Jo Ann LeQuang, Robert Taylor, and Robert B. Raffa. 2018. "Going beyond Prescription Pain Relievers to Understand the Opioid Epidemic: The Role of Illicit Fentanyl, New Psychoactive Substances, and Street Heroin." *Postgraduate Medicine* 130 (1): 1–8.

Peroni, Silvio, Alexander Dutton, Tanya Gray, and David Shotton. 2015. "Setting Our Bibliographic References Free: Towards Open Citation Data." *Journal of Documentation* 71 (2): 253–77.

Porta, Miquel, Jan P. Vandenbroucke, John P. A. Ioannidis, Sergio Sanz, Esteve Fernandez, Raj Bhopal, Alfredo Morabia, Cesar Victora, and Tomàs Lopez. 2013. "Trends in Citations to Books on Epidemiological and Statistical Methods in the Biomedical Literature." *PloS One* 8 (5): e61837.

pubmeddev. n.d. "Home - PubMed - NCBI." PubMed. Accessed January 23, 2019. https://www.ncbi.nlm.nih.gov/pubmed/.

Qudrat-Ullah, Hassan, and Peter Tsasis. 2017. *Innovative Healthcare Systems for the 21st Century*. Springer.

Rhonda Cook, Jeremy Redmon. 2017. "Middle Georgia's Opioid Overdose Outbreak Remains Unsolved." Myajc. The Atlanta Journal-Constitution. December 12, 2017. https://www.ajc.com/news/crime--law/georgia-drug-overdose-outbreaks-stunning-harming-many-families/HR9WvDJzk4yDUdovkUNVWP/.

Rivers, Caitlin M., Eric T. Lofgren, Madhav Marathe, Stephen Eubank, and Bryan L. Lewis. 2014. "Modeling the Impact of Interventions on an Epidemic of Ebola in Sierra Leone and Liberia." *PLoS Currents* 6 (October). https://doi.org/10.1371/currents.outbreaks.fd38dd85078565450b0be3fcd78f5ccf.

Rodriguez, Dania M., Michael A. Johansson, Luis Mier-y-Teran-Romero, moiradillon, eyq, YoJimboDurant, Bianca N. Doone, et al. 2017. *Cdcepi/zika: May 29, 2017*.

https://doi.org/10.5281/zenodo.584136.

Rosenstock, Irwin M. 1974. "Historical Origins of the Health Belief Model." *Health Education Monographs* 2 (4): 328–35.

Schlitt, James S., and Bryan L. Lewis. 2017. "Digital Disease Surveillance in Rural Appalachia." In *ASA Annual Conference*. mds.marshall.edu. http://mds.marshall.edu/asa_conference/2017/accepted_proposals/307/.

Schlitt, James T., Bryan Lewis, and Stephen Eubank. 2015. "ChatterGrabber: A Lightweight Easy to Use Social Media Surveillance Toolkit." *Online Journal of Public Health Informatics* 7 (1). https://doi.org/10.5210/ojphi.v7i1.5717.

Schlitt, J., P. Bordwine, V. Truong, B. Lewis, and S. Eubank. 2019. "Optimizing Public Health Opioid Messaging with ChatterGrabber." presented at the Virginia Department of Health Preparedness Conference, Roanoke, VA. https://docs.google.com/presentation/d/11ty3BdAav8Sq9qD_mudwldFcHnSX2iLgIxrvDFJo0ek/edit.

Scholl, Lawrence. 2019. "Drug and Opioid-Involved Overdose Deaths — United States, 2013–2017." *MMWR. Morbidity and Mortality Weekly Report* 67. https://doi.org/10.15585/mmwr.mm6751521e1.

Skiles, Keith. 2017. "Harmful Algal Blooms (101) –Bloom Basics and Public Health Response." presented at the TRAIN Course ID: 1070588, July 25. http://www.vdh.virginia.gov/content/uploads/sites/12/2017/08/HAB-101-training_2017_onepdf.pdf.

US Census Bureau Geography. 2012. "Cartographic Boundary Shapefiles - Urban Areas," September. https://www.census.gov/geo/maps-data/data/cbf/cbf_ua.html.

"WHO | Disease Outbreaks." 2019, January. http://www.who.int/emergencies/diseases/en/.

"WHO | Novel Coronavirus Infection in the United Kingdom." 2015, June. http://www.who.int/csr/don/2012_09_23/en/.

"WHO | Origins of the 2014 Ebola Epidemic." 2015, September. http://www.who.int/csr/disease/ebola/one-year-report/virus-origin/en/.

"WHO | Plague – Madagascar." 2015, June. http://www.who.int/csr/don/21-november-2014-plague/en/.

"WHO | The History of Zika Virus." 2017, February. http://www.who.int/emergencies/zika-virus/history/en/.

Yi, Ming, and Achla Marathe. 2015. "Fairness versus Efficiency of Vaccine Allocation Strategies." *Value in Health*. https://doi.org/10.1016/j.jval.2014.11.009.

Appendices

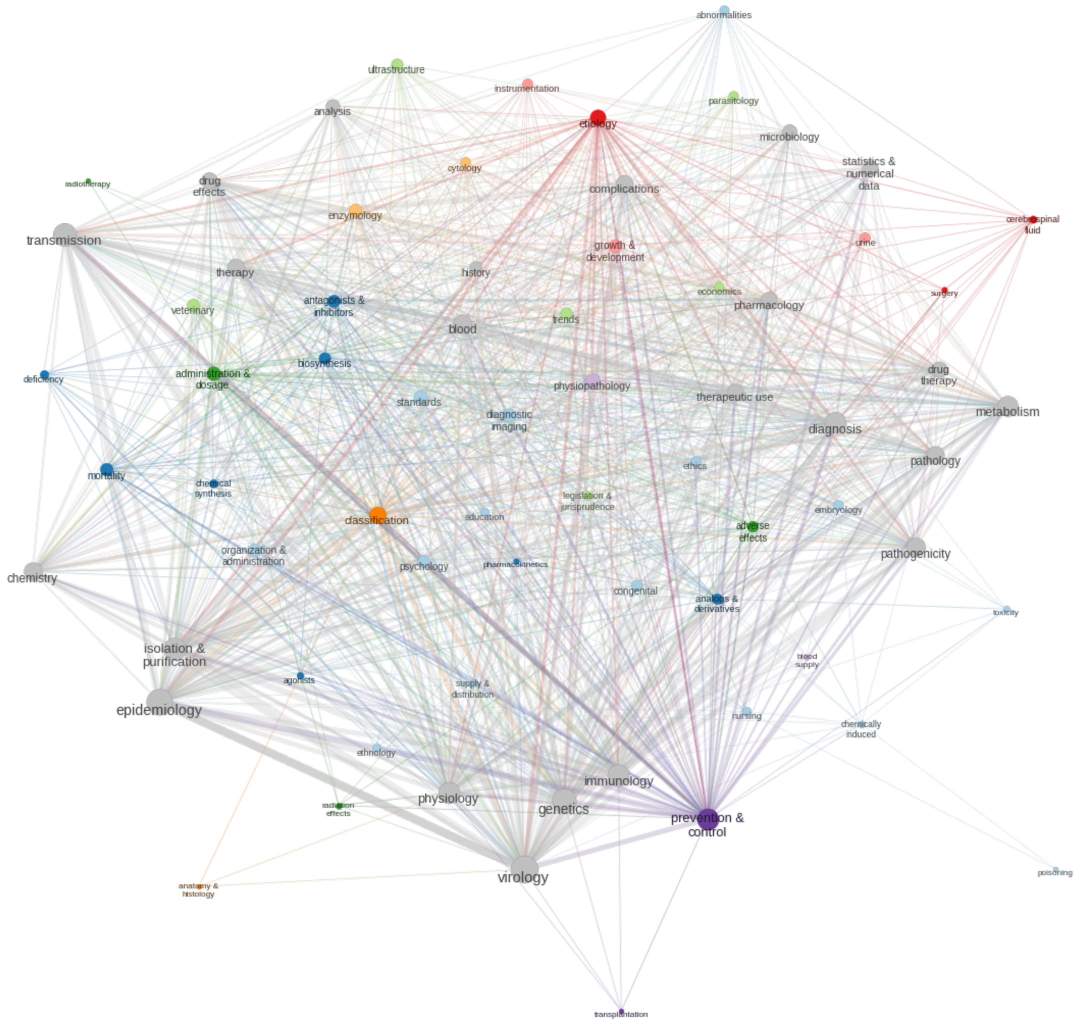
Appendix A

Supplemental Information: The State of Open Access in Outbreak Science

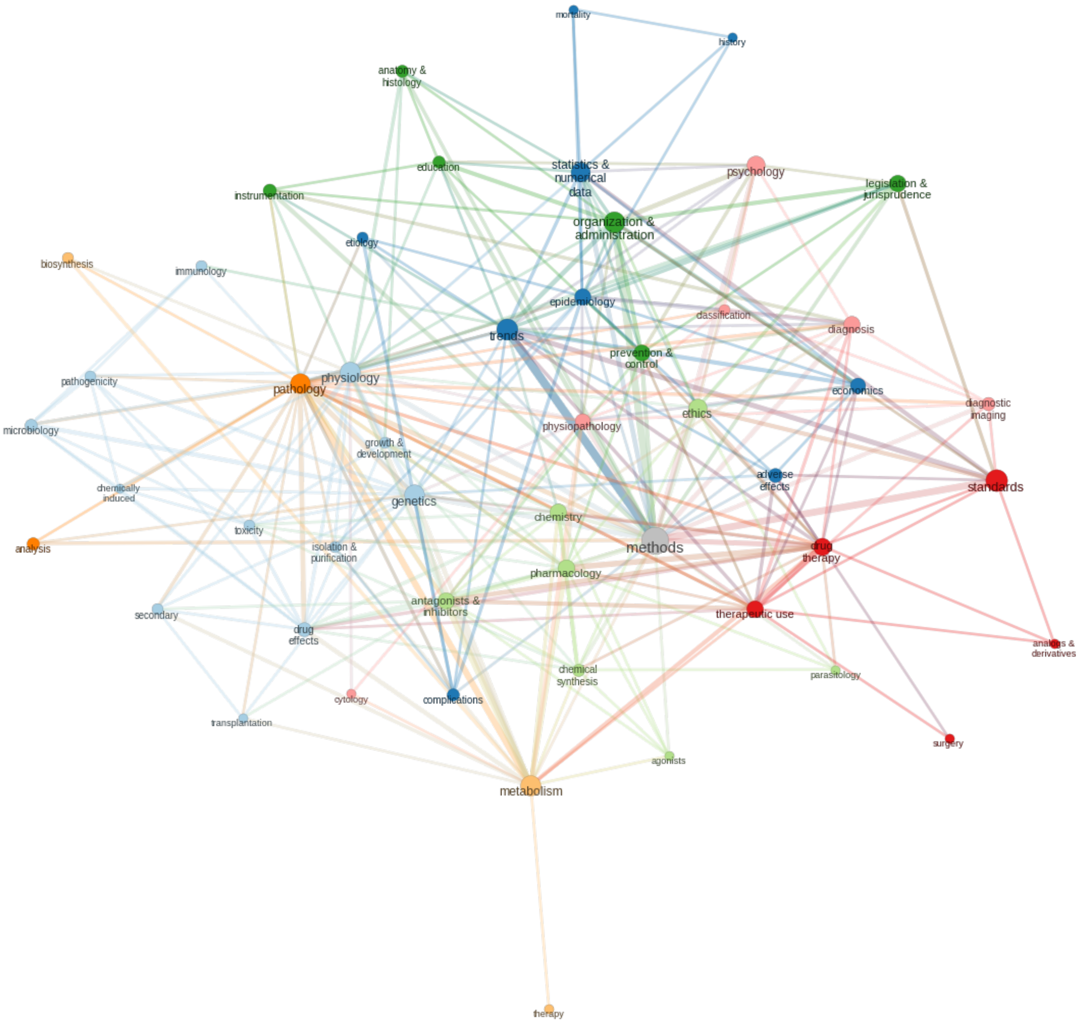
A.1 ScienceWebVis interactive Python notebooks

- **Research notebook:**
  - https://colab.research.google.com/drive/1Xp4_YhgVKlJMGuahEcC_6lXr9C-rsNFt
- **Public sandbox**
  - https://colab.research.google.com/drive/1GobsutmWjddSm4iovgcFGiIWgI_-29WM

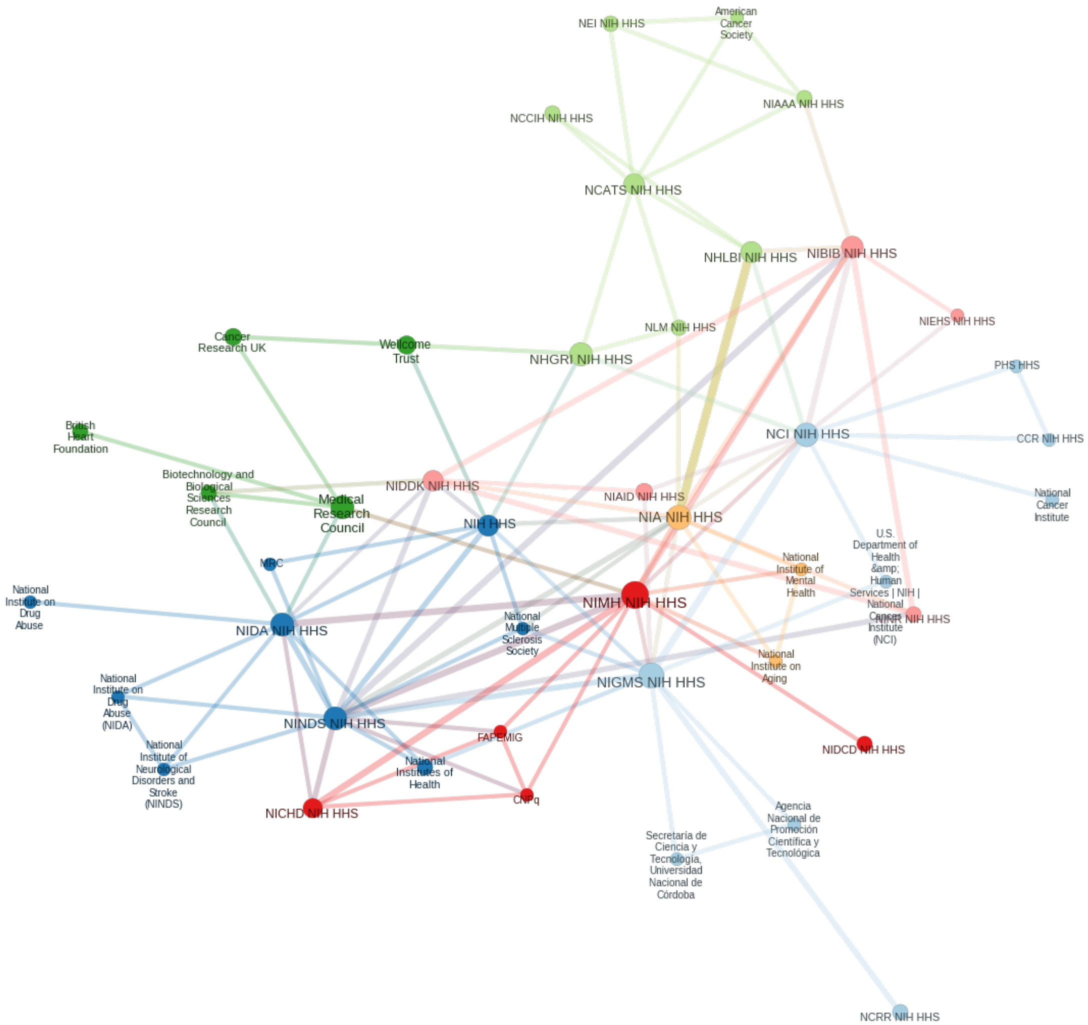
A.2 NetVis Chapter 2 publications qualifier network



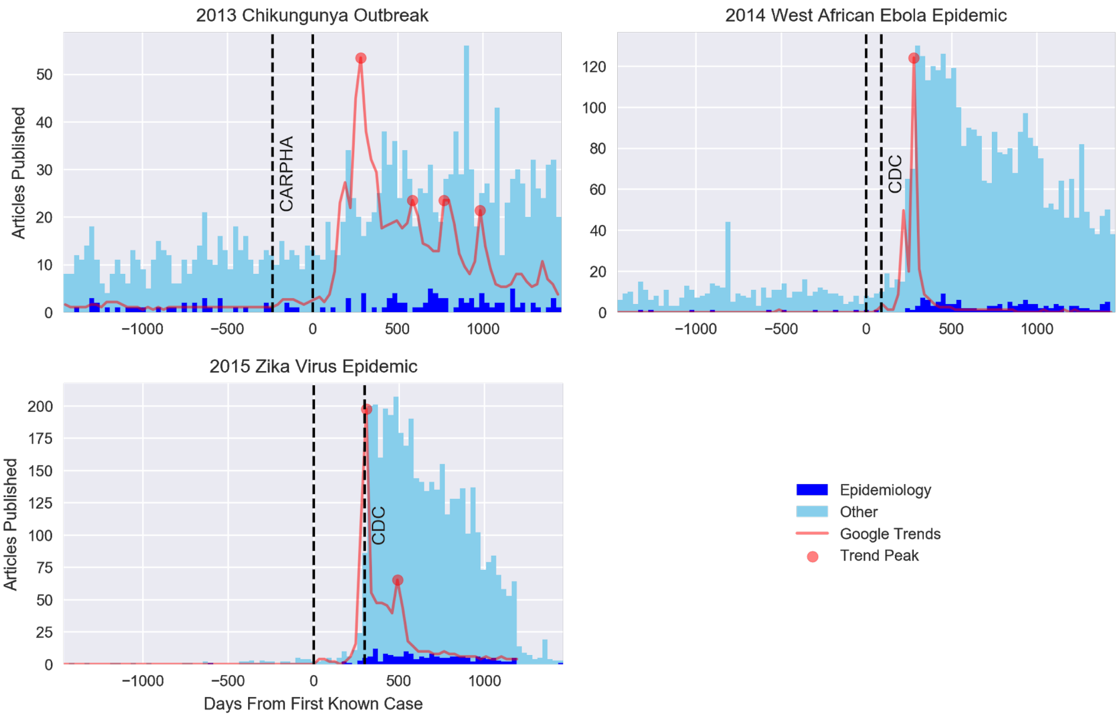
A.3 NetVis 2019 open science qualifier network



A.4 NetVis 2019 open science grant agency network


NEI NIH HHS
American Cancer Society
NIAAA NIH HHS
NCCIH NIH HHS
NCATS NIH HHS
NHLBI NIH HHS
NIBIB NIH HHS
NIEHS NIH HHS
NLM NIH HHS
Cancer Research UK
Wellcome Trust
NHGRI NIH HHS
PHS HHS
British Heart Foundation
NCI NIH HHS
CCR NIH HHS
Biotechnology and Biological Sciences Research Council
NIDDK NIH HHS
Medical Research Council
NIAID NIH HHS
National Cancer Institute
NIH HHS
NIA NIH HHS
U.S. Department of Health &amp; Human Services | NIH | National Cancer Institute (NCI)
National Institute of Mental Health
MRC
National Institute on Drug Abuse
NIMH NIH HHS
NINR NIH HHS
NIDA NIH HHS
National Multiple Sclerosis Society
National Institute on Aging
NIGMS NIH HHS
National Institute on Drug Abuse (NIDA)
NINDS NIH HHS
FAPEMIG
NIDCD NIH HHS
National Institute of Neurological Disorders and Stroke (NINDS)
National Institutes of Health
CNPq
NICHD NIH HHS
Agencia Nacional de Promoción Científica y Tecnológica
Secretaría de Ciencia y Tecnología, Universidad Nacional de Córdoba
NCRR NIH HHS

A.5 Limited prior history outbreak publication histories v Google Trends



Appendix B

Supplemental Information: Opioid Social Media Message Mapping

B.1 ChatterGrabber Github repository

- https://github.com/jschlitt84/ChatterGrabber

B.2 ChatterGrabber workshops

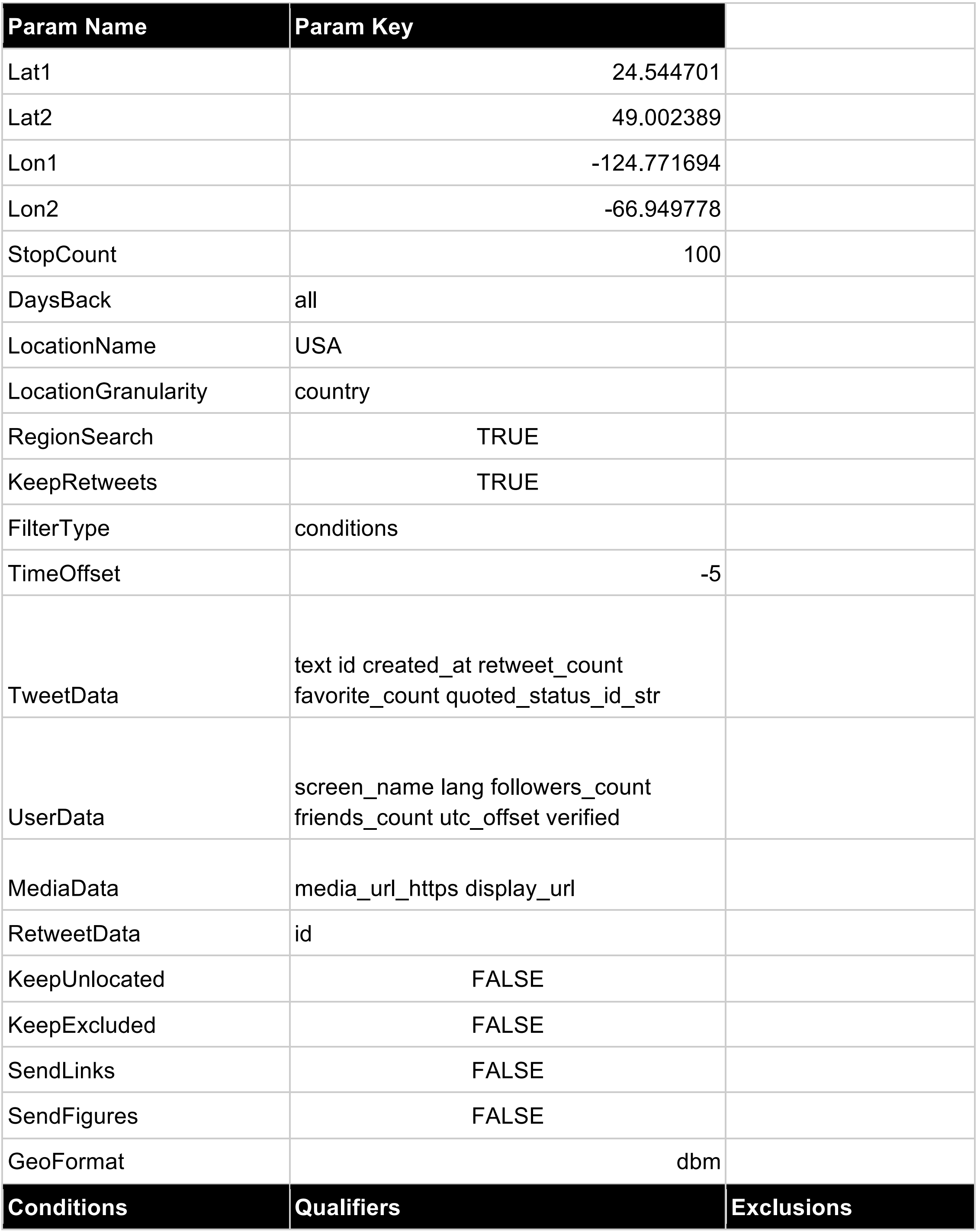
- https://drive.google.com/drive/folders/0B8VhE4yQ6s1FfnpHeEZaTHk0bU9IZjl6QmNlZmxVMUJmd3RIZGthbUl3NFRnNGZfekVhQk0

B.3 ChatterGrabber virtual machines

- https://drive.google.com/drive/folders/0B8VhE4yQ6s1FfnF4aGoxWXNibDhHeDZSQzNBTGluRTBwSEZRVGFYOG1wZHB2OVpGeXpQc2c

B.4 ChatterGrabber collection parameters

- https://docs.google.com/spreadsheets/d/1FX1iM5N_At6kevaKtjJMtS7KkB1kGtDKFKyBTUch_MQ/edit#gid=0



Param Name	Param Key	
Lat1	24.544701	
Lat2	49.002389	
Lon1	-124.771694	
Lon2	-66.949778	
StopCount	100	
DaysBack	all	
LocationName	USA	
LocationGranularity	country	
RegionSearch	TRUE	
KeepRetweets	TRUE	
FilterType	conditions	
TimeOffset	-5	
TweetData	text id created_at retweet_count favorite_count quoted_status_id_str	
UserData	screen_name lang followers_count friends_count utc_offset verified	
MediaData	media_url_https display_url	
RetweetData	id	
KeepUnlocated	FALSE	
KeepExcluded	FALSE	
SendLinks	FALSE	
SendFigures	FALSE	
GeoFormat	dbm	
Conditions	**Qualifiers**	**Exclusions**

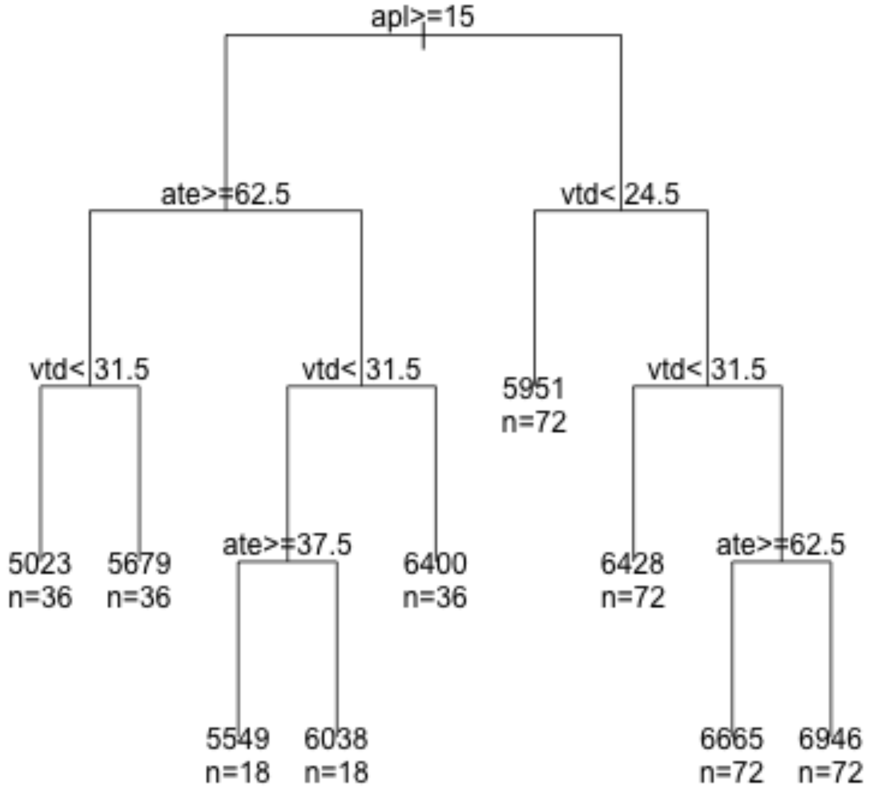
drug	bad trip	song
opioid	self medicated	music
norco	rolled balls	Future
perc	milligram	The Weeknd
oxy blues	gram	Lil Dicky
oxy	milleys	Wiz Khalifa
scag	toast up	Mally Mall
hydro	seizure	Cedric Gervais
oxycodone	can't breathe	Tyga
hydrocodone	trouble breathing	Stripper Joint
codeine	twitchin	Mask Off
vicodin	itchin	Lil Uzi Vert
percocet	traphouse	bad and boujee
morphine	relapse	cookin up dope with the uzi
hydromorphine	tweaked	cookin up dope in the hotpot
fentanyl	dumb tweaked	jazz
methadone	fake	Post Malone
lortab	vomit	Quavo
dilaudid	overdose	Migos
oxymorphine	counterfeit	Schoolboy Q
fiorcet	sick	Lil Durk
heroin	slump	Blac Youngsta
dance fever	OD	Lil pump
suboxone	pain	trump
meperidine	purple	death penalty
demerol	purple skin	duterte
astramorph	limp body	liberals
roxicet	antidote	conservatives

naloxone	dying	
fentora	rolling	
actiq	fatal	
zohydro	geek	
lorcet	tweak	
carfentanil		
prescription		
yellow pill		
buprenorphine		
xan		
fake drugs		
muffin man		
norc		
speedball		
zan		
benzo		
poison pill		
fentanil		
fentenil		
fentenel		
black tar		
narcan		
purple drank		
grey death		
gray death		
lean		

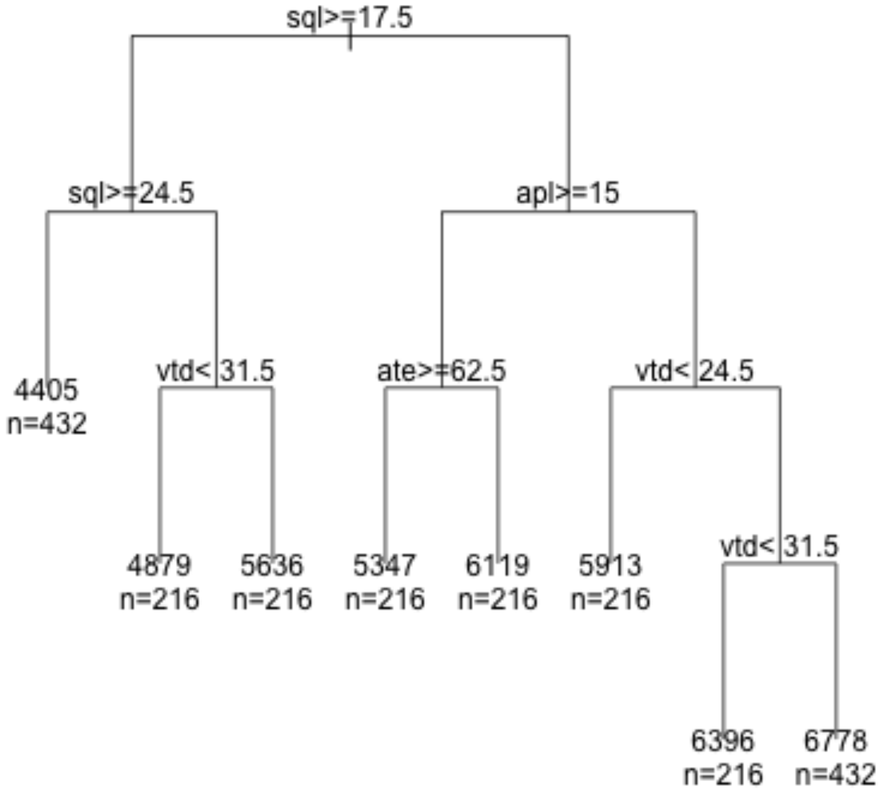
Appendix C

Supplemental Information: Planning Protection of Personnel

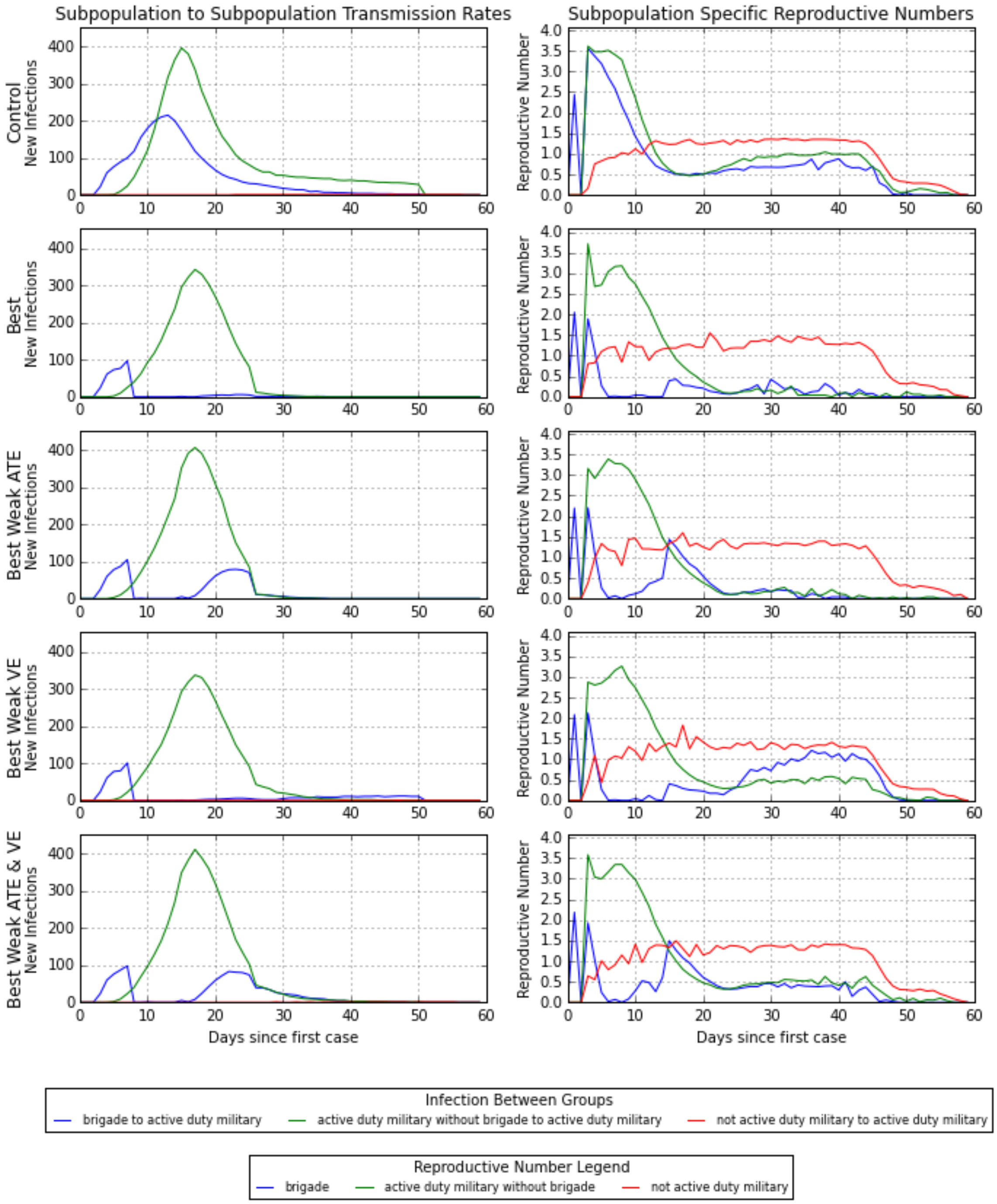
C.1 Correlation and regression tree analysis of short sequestration scenario cumulative case counts



C.2 Correlation and regression tree analysis of short sequestration scenario cumulative case counts



C.3 Subpopulation to subpopulation transmission analysis


Subpopulation to Subpopulation Transmission Rates
Subpopulation Specific Reproductive Numbers
Control
Best
Best Weak ATE
Best Weak VE
Best Weak ATE & VE
New Infections
Reproductive Number
Days since first case
Infection Between Groups
brigade to active duty military
active duty military without brigade to active duty military
not active duty military to active duty military
Reproductive Number Legend
brigade
active duty military without brigade
not active duty military