

# Routes of communication: understanding the information flows that shape public and official reaction to COVID-19

## 1 Project overview

During disease outbreaks of international concern, like 2003 SARS, 2009 H1N1, 2014 Ebola, 2016 Zika ((jd: fill in better names with dates)) ((mli: filled, delete comment if approved)) or the current COVID-19 outbreak, this process ((de: what process? You haven't specified one.)) is both compressed and amplified. The stakes also become higher, because public behaviour directly affects the spread of infectious disease: people who avoid large gatherings may slow disease spread, while people who flee infected areas may accelerate it, for example. Excessive fear of disease spread can have severe economic effects, and may also lead to bias and discrimination against groups seen as linked to the disease, or merely seen as others. ((de: What does "seen as others" mean?)) ((cs: "others" in quotation.))

The proposed research will study how information flows between forums, including scientific publications; governmental policies and agency recommendations; and mass and social media – and investigate how it affects public perceptions and behaviours. Our interdisciplinary group will analyze these flows by combining textual and contextual analysis; AI-assisted human-supervised data mining; time-series analysis; and dynamical modeling. We will study how good information competes with misinformation, and look for factors correlated with successful spread of good information. We will gather information on communication and surrogates for behaviour from a wide range of sources. Sources about communication will include agency websites; preprint servers and publicly available scientific journals; major newspaper websites; social-media platforms; and twitter data. ((cs: why single out twitter from social-media?)) Surrogates for public perceptions and behaviour will include twitter data (again) ((cs: and Sina Weibo?)); google trends; publicly available box-office information for movies and major sporting leagues; and information ((cs: public transportation rides?)) about cancellations and shortages (for example of face masks or pharmaceutical or pseudo-pharmaceutical products) from our textual analysis.

The project will be organized around three Research Questions (RQs):

**RQ1** How does information (and misinformation) travel between scientists, public-health workers, mass media and social media? (*Research sub-area*: “cultural dimensions of the epidemic”) ((cs: Make how mass and social media convey information a sub questions? Its content is crucial in this proposal; and I believe this is where we will get most of details on cultural perspective.)) ((mli: What is Research sub-area anyway? Do we even need it?))

**RQ2** How does communication affect public behaviour and the course of the outbreak? (*Research sub-area*: “public health response”) ((jd: improve this quote)) ((de: Behavioural responses and effects on outbreaks?))

**RQ3** How can scientists and policy-makers evaluate and improve the effectiveness of their communication? (*Research sub-area*: “strategies to combat misinformation”)

## 2 Background

Public-health communication is a balancing act. Officials are often caught between the need to be heard, and the danger of causing panic. This problem is particularly acute in the case of an infectious disease outbreak, since the presence of a novel pathogen increases both the importance of being heard and the danger that the public will over-react.

In the case of COVID-19, scientists are still scrambling to understand the pathogen's biology; public-health workers are scrambling to decide on the best recommendations and policy decisions given current knowledge at any given time; and the mass media is scrambling to understand the situation and decide how best to communicate with the public.

There are other complicating factors. An outbreak of global concern represents an opportunity for mainstream and peripheral media, and for social-media actors to increase their "clicks" and "likes" and therefore prestige and/or profitability. These motivations work against the balancing act, and instead favor over-simplification and sensationalization.

It is known that traditional media can strongly influence public perceptions, creating fear by overestimating risk during the SARS outbreak [?] and the influenza pandemic [?], or feeding into bias – e.g., anti-Chinese bias during the SARS crisis [?]. **((naveed: There are reports of bias against Chinese people in social media and in news reports. Could be cited here. <https://www.theguardian.com/uk-news/2020/feb/09/chinese-in-uk-report-shocking-levels-of-racism-after-coronavirus-outbreak><https://www.vox.com/2020/2/7/21126758/coronavirus-xenophobia-racism-china-asians><https://www.theguardian.com/world/2020/jan/28/canada-chinese-community-battles-racist-backlash>** Media is also the tool public health authorities rely on to promote their concerns and recommendations during outbreaks. Understanding media effects on disease spread (e.g., media attention increases self-protection) can help enhance epidemic forecasting and preventive measures to slow the disease spread [?].

While traditional news media (including online presence) remains influential, social media plays an increasingly important role in shaping how we communicate and understand information [?]. Social media can play a positive role spreading good information, [?, ?, ?], but may also spread misinformation and feed bias [?, ?]. **((naveed: Misinformation is mainly coming from non-medical community and various groups promoting their political or strategic interest or even without any interest.))** Since the initial reports, a cluster of acute severe respiratory disease (COVID-19) and the potential for global spread, there has been widespread discussion and dissemination of information through social media [?, ?, ?]. **((hyuju: mentioning "social media as an important source for identifying how people feel and react to the situation" could be relevant here, connecting to the methods (especially topic modeling and sentiment analysis) later.)) ((mli: I feel like background is getting heavy. I suggest moving topic modeling and sentiment analysis to the Analysis section and describe clearly what these things are and how we are going to use it)) ((naveed: However, there have been exchanges and disagreements between scientists on the likely risk and spread of disease, and validity of claims and creditability of these scientists have been questioned.))**

**((jd: What should we put back here?))**

**((jd: More about how information spreads from scientists and policy makers to social media (both directly and indirectly). Not sure we need google trends**

here; we can talk about it nicely in metrics, I hope.))

((jd: Should we try to bring in factual themes? The key ones are risk of worldwide spread; case-fatality proportion; something about control strategy effectiveness. The fuzziness of the last one is why I haven't tried to do this yet.))

## 3 Methods and feasibility

### 3.1 Data

**Science** We will develop systematic search and screening strategies to extract relevant peer-reviewed publications from Google Scholar and PubMed. To account for the strong influence of preprints early in the epidemic [?], we will also include preliminary scientific findings posted on medRxiv through 31 March 2020. We will index these papers and track their appearances in mass media and social media; we will also track which of the preprints are published after peer review.

**Public health recommendations** We will collect and analyze reports, guidelines and recommendations available from the World Health Organization and from the central disease control agency of each of our focal countries. It is worth noting that all of these agencies have launched special COVID-19 web pages.

**Mass media** We will use the Lexis-Nexis search engine (via McMaster University) and OriProbe Information Services to collect articles relevant to the outbreak, going back to the outbreak start in December 2019, and continuing throughout the grant period. We will focus on the top English- and Mandarin-language newspapers (taking both circulation and online access into account) from Canada, China, England, Singapore, Taiwan, and the USA. We will include the top Mandarin-language newspapers in both Canada and the USA.

**Social media** We will efficiently collect data from Twitter and Weibo by purchasing API access, using data going back to November 2019 – before the epidemic started – to give a baseline for comparison.

**Public response** Twitter (and Weibo) ((mli: I suggest we take out Weibo unless someone on the team signs it off and know how it works and how hard it is to get data)) data will give us information not only on information flows, but also on public interest, attitudes and topics being discussed on the social media. We will also probe public interest and concern using publicly available data from GoogleTrends, which tabulates frequency of searches (by search times and topics) in various regions across the world [?, ?]. Economic data relating to the outcome of the COVID-19 outbreaks will be gathered as reference for contextual analysis (see below): for example, travel, movies box offices, cancellation of public events.

## 3.2 Analysis

**Textual analysis** We will use both human coders and AI approaches for textual analysis of scientific papers, government recommendations, mass media and social media. Coders will develop codebooks based on scanning information, then refine these codebooks while systematically coding a randomly selected subset of articles from each stream. Codebooks will contain both themes and frames for the analysis. We will train AIs built using established approaches for textual analysis to these human-scored articles. Programmers will work back and forth with coders and subject-matter experts to reach a final codebook that is consistent with study aims and can be scored reliably by computer, allowing us to review a very large number of articles. This approach will allow us to scan for tweets (and Sina Weibo) spreading misinformation in real time. Such information would be shared with public-health practitioners to assist with counter-messaging strategies. ((naveed: This covers most of it. However, we had ad discussion that we could start with aspect based sentiment analysis and topic modeling, which will require relatively less human involvement. Traditional classification algorithms that we use require a coded data as you proposed but the amount of coded data to make it work may take some time to get developed while sentiment analysis and topic modeling could start the text analysis process.))

**Time-series analysis** We will use cross-correlation analyses to look for indicators that information is moving from one communication forum to another; that events (like disease spread or public behavior) are affecting communication; or that communication is affecting events. Cross-correlation analysis is complicated and prone to false-positive results. Importantly, therefore, we will be able to use the cross-correlation analysis to generate hypotheses that can be checked by more detailed textual analysis. For example, if we hypothesize that tweets about fatalities are being driven at a certain time and place by mass media, or by government policies, we can sample from those tweets and examine them for detailed information or citations; if we hypothesize that a trend in self-isolation is driven by social media, we can search for mass media stories that interview people about their motivation. The ability to compare large-scale trends with detailed texts should amplify our pattern to detect and confirm patterns.

**Dynamical modeling** Dynamical modeling provides the link between individual events and emergent phenomena. We will make a range of simple dynamical models to further probe our time-series results by asking what mechanisms may underlie our observed connections, and what these connections might imply for the future. Dynamical models will allow us to explore hypotheses about what factors affect behaviour, and also to explore new hypotheses about how changes in behaviour are likely to loop back to disease transmission or to panic responses that might lead to shortages or to impacts on regional economies or the global economy. ((naveed: This is not clear to me what we are trying to model here and how this will link back with the other themes.))

**Synthesis** We will combine results from our analysis techniques above to formulate hypotheses about what factors lead to effective communication. In particular, we will identify

cases where good information did or did not out-compete bad information. When information from public-health agencies spreads effectively we will also evaluate our behavioural proxies to ask when it led to a calibrated reaction from the public (as opposed to over- or under-reaction). ((cs: Tie findings of the above back to the three RQ?))

### 3.3 deliverables

- software ((naveed: Describe what kind of software. I was envisioning a tool that runs on Twitter stream data and classifies tweets based on contents as well as significance in terms of influence ( how much it has potential to spread based on following network or influencer. See paper for classification of influencers. <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6747-8> ))
- Communication: Writing papers is certainly one, but we should be tweeting and etc. What are communication platforms we should be communicating in? ((cs: attending conference, meeting with policy makers with our findings; possible WHO network?)) ((naveed: It depends on objective, would we be doing counter messaging or public health. The way I was envisioning, creating a tool to identify and alert public health communication team and they addressing it through tweets or other means. )) ((de: Perhaps combine the tweets with a blog, which would engage different audiences. The tweets could summarize more elaborate blog posts. I have no experience writing either a blog or tweets myself...))

## 4 Research Setting & Personnel

The principal applicants have a long history of influential research on disease outbreaks:

The research will principally take place at McMaster University. Nominated principal applicant Dr. David J.D. **Earn** (6 hours per week) led the creation of the International Infectious Disease Data Archive and has expertise in gathering and curating infectious disease data, and in dynamical modelling, including modelling the influence of individual decision-making on epidemic dynamics. Principal applicant Dr. Jonathan **Dushoff** (4 hours per week) is an internationally recognized expert in infectious disease modelling, has extensive experience with statistical frameworks for fitting models to data, and has been involved in the Ebola challenge and other forecasting projects. Co-Applicant Dr. Benjamin **Bolker** (5 hours per week) is a highly accomplished ecological statistician with extensive experience in spatial-dynamical modeling, statistical modeling and statistical software. Co-Applicant Dr. Naveed Z. **Janjua** (5 hrs week) leads data and analytic services at the BCCDC and was involved in the 2009 H1N1 pandemic response; during pandemic, lead or contributed to studies on immuno-epidemiology of pandemic H1N1, household transmission, modelling, effect of prior seasonal vaccine receipt on pandemic H1N1 infection risk and pandemic vaccine effectiveness. Co-Applicant Dr. Chyun-Fung **Shi** (40 hours per week), post-doctoral researcher, has ... Co-Applicant Dr. Michael **Li** (5 hours per week), post-doctoral researcher, has focused his research on epidemic forecasting and is experienced working with large databases.



## 5 Research Time line

((mli: not sure if we need this section, but CIHR want to see we are working on things and are ready to do more work.)) ((naveed: I think we need to see what we could start quickly. HyeJu our postdoc has already started working on getting twitter data and starting to analyze it to make sense. We should say that are ready to get started. ))

### 5.1 Year 1

### 5.2 Year 2

### 5.3 Potential Outcomes

Transfer findings to peer-reviewed publications. In addition to a grand one paper, we plan to transfer findings into subgroups, such as by affected country (North America, Asia-Taiwan, H.K. and Singapore), types of media (news media and social media (twitter)).

- Potential to contribute to the global response to COVID-19
- Social and policy countermeasures and Global Coordination Mechanism

We could say; We will work with public health teams to inform their communicate strategies tackling misinformation and design a tool for real-time monitoring and tackling misinformation.

We will also provide data on magnitude of discriminatory information and its geographical origin to tackle discrimination. ))((naveed: Data for public health response. Our provincial health officer is asking for data and today there was a visit from Federal Health Minister and BC Health Minister to Richmond related to affect on Chinese community.

We could say; We will work with public health teams to inform their communicate strategies tackling misinformation and design a tool for real-time monitoring and tackling misinformation.

We will also provide data on magnitude of discriminatory information and its geographical origin to tackle discrimination. ))

## 6 Challenges and Mitigation Strategies

Social media in China is to included due to data availability. Social media such as twitter and Facebook are barred in China and WeChat does not share their database. Yet, there are information and coverage about the outbreak inside China in the media. We will categorize people's reaction and understanding of COVID-19 on the news content or the tweets to supplement our results.

((naveed: What about Weiboo. Good to highlight that data may not be available or limited in scope. Other mechanism of collecting that data e.g., web crawler for data available in public domain.))

**Data curation** We will go back and document clearly all the policy changes and case definitions.

- data collected from National Health Commission
- Figure out how to use the data effectively (e.g. we are not using death, severity categorizations, number of tested, number of positive, and etc)
- Case definitions
- Media content
- Social media platforms
- Google trends

**Analysis/ Pipelineing/ mainstreaming** Info delays, misinformation and miscommunications

**Communication** how people are interpreting

## **7 Conclusion/summary**

References