Participatory syndromic surveillance tool for the early prediction of COVID-19 outbreak in Bangladesh

By Sangita Paul (UNDP), Anir Chowdhury (UNDP), Kawsar Hossain Sojib (UNDP) and Tara Min Ju Lee (UNDP)

Very few countries affected by COVID-19 have sufficient viral testing capacity to adequately monitor cases occurring in the community. Alternative indicators, such as hospitalization and death rates, have a lag time from two to three weeks, disrupting the mitigation and control process of the COVID-19 pandemic. To address this challenge, UNDP Bangladesh has introduced a participatory syndromic surveillance tool whereby symptoms consistent with COVID-19 are reported by the public. This brief aims to portray the efficacy and usefulness of participatory surveillance as a tool of mapping disease spread at the initial phase of a pandemic. Syndromic data based on self-reported symptoms by the public demonstrate a strong association with lab-confirmed cases on a local scale. Moreover, the syndromic data also suggest an earlier spread of the outbreak across Bangladesh than is evident from the confirmed case counts. The use of alternative data sources for mapping the disease burden, along with cultivating its potential for early strategic preparedness, thus concerns national and international stakeholders alike in combatting ongoing and future pandemics.

Participatory syndromic surveillance augmenting hospital-based surveillance

The ongoing coronavirus disease 2019 (COVID-19) pandemic has overwhelmed healthcare systems around the world, exposing the challenges faced by public health agencies when responding to rapidly emerging outbreaks. The global counts of confirmed and fatal cases of COVID-19 were 22.4
DEVELOPMENT FUTURES SERIES

UNDP Global Policy Network Brief

million and 789,000, respectively, as of Aug 20, 2020. However, the actual count of cases and deaths was much higher. One possible reason for this miscalculation was the worldwide limited testing labs and kits during the early time of the pandemic. This caused reporting delays and data were not updated regularly and in a timely manner.

Syndromic surveillance can supplement the traditional hospital-based surveillance of symptoms reported by people (crowdsourced data). Since symptoms are reported before the outbreak of the disease, it can help in identifying the spread, spikes and prevalent disease areas. This eventually assists in making decisions regarding lockdowns, quarantines, testing allocations and promotional campaigns. The disease can be prevented and controlled with effective interventions beforehand.

Japan introduced a COVID-19 syndromic surveillance system called Covid-19: Operation for Personalized Empowerment to Render smart prevention and care-seeking (COOPERA), which is linked via a mobile messenger application. COOPERA collects peoples’ ages, gender, occupations, postcodes and COVID-19 symptoms. The symptoms included fever, fatigue and shortness of breath. Statistical analysis of the crowdsourced data of the COOPERA system showed that it had a correlation with lagged COVID-19 incident data, providing an early warning of the COVID-19 disease outbreak in Japan at local levels.

In a study conducted by Johns Hopkins University, body temperature and COVID-like illness symptoms were collected from people in the United States using a smartphone application. Using spatio-temporal clustering techniques and cross-correlation analysis, it was found that there was a strong correlation between the symptoms and temporal lagged COVID-19 cases. Participatory crowdsourced syndromic surveillance has been applied in many developed countries including COVID-19 with great success. It was found that the data on syndromic surveillance indicators had a strong correlation with confirmed COVID-19 cases.

On a national scale, control efforts should be guided by accurate data on cases and disease burden ideally captured through widespread surveillance. However, the public health domain of Bangladesh had a dearth of COVID-19 incident data to help respond to the outbreak efficiently and timely.

Confirmed case counts remained low in Bangladesh in the month following the first detected case on 8 March 2020, reflecting the very limited testing capacity at the start of the outbreak. Initial testing was limited to travellers arriving from abroad and known contacts of confirmed cases. Testing was also largely limited to Dhaka until the end of March. Until 27 March, a single laboratory was responsible for administering and analysing COVID-19 tests in Bangladesh, and only about 1,600 total samples were tested nationally by the end of March. The testing capacity of SARS-CoV-2 increased significantly from a daily average of fewer than 100 tests in March to about 15,000 in June 2020. The total confirmed COVID-19 cases in Bangladesh reached nearly 200,000 by mid-July. As in most countries, the testing capacity could only cover a small fraction of symptomatic cases. Reporting delays, especially in the rural and remote areas, added to challenges in monitoring the spread of the epidemic across the country in real-time.

For the purpose of augmenting disease surveillance, UNDP Bangladesh deployed a participatory surveillance system designed to collect syndromic information based on self-reported symptoms. This amassed reports of multiple sources ranging from hotlines to web applications through a public-private partnership involving telecommunication partners and government agencies. The system is also assisted by a telemedicine team comprised of clinicians and aims to identify disease hot spots to inform policymakers of corresponding response measures. Additionally, the system was rapidly scaled over the course of the first few months of the outbreak.

This paper is an assessment of the utility of participatory syndromic surveillance in evidence-based decision-making within the COVID-19 response context, which includes identifying high-risk zones, allocating resources accordingly and enacting lockdowns with social distancing measures.

Digging into the syndromic data

The self-reported syndromic surveillance data analysed here has been compiled from three main sources: 1) an interactive voice response (IVR) system, 2) several internet and mobile applications and 3) an unstructured supplementary service data (USSD) based messaging system (See Figure 1). All systems are available free of charge to the user. Individuals are asked to report their symptoms.
(cough, fever and shortness of breath), any contact with symptomatic people and/or with someone who tested positive for COVID-19 and in some cases their age and gender. Coughing and sore throat symptoms were mostly reported by people via various channels per month during the period of April 2020 to June 2020 (Table 1). 21 Each response in the system is geolocated based on the nearest cell phone tower of the respondent and mapped to upazilas (subdistricts), the operationally relevant administrative unit in Bangladesh.

Table 1: Average frequency distribution of reported symptoms per month from April 2020 to June 2020

<table>
<thead>
<tr>
<th>Types of symptoms</th>
<th>Average number per month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fever</td>
<td>324</td>
</tr>
<tr>
<td>Coughing</td>
<td>754</td>
</tr>
<tr>
<td>Shortness of breath</td>
<td>234</td>
</tr>
<tr>
<td>Sore throat</td>
<td>732</td>
</tr>
</tbody>
</table>

Source: Director General of Health Services (DGHS), hotline equipped with an interactive voice response (IVR) system (333), internet and mobile applications and unstructured supplementary service data (USSD) based messaging system (*3332#) verified by clinicians

A subset of individuals reporting symptoms is connected with a human verifier, i.e., telemedicine doctors, healthcare professionals or trained field workers, for a preliminary diagnosis based on responses to questions and algorithms specific to each mobile phone operator. For example, if an individual reported multiple symptoms, such as coughing, fever, sore throat, etc., and had contact with a COVID-19 patient in the past 14 days, he/she would be at high risk. The classification process of COVID-19 patients through responses are agreed and verified by the Director General of Health Services (DGHS) of Bangladesh. Following an evaluation over the phone with a human verifier, individuals are then classified as having high or low risk for COVID-19. This classification is based on the reported symptoms, contact with people with symptoms, and, to some extent, the judgement of the human verifier. Since self-reported data is inherently “noisy”, we focus on the subset of people reporting multiple symptoms consistent with COVID-19 and those who were connected with a human verifier (Figure 1).

Figure 1: Self-reported data sources and the human verification process

Source: Authors’ own elaboration
**Association between lab confirmed cases and syndromic data**

The number of cases per subdistrict are adjusted by the population of the subdistrict with data from WorldPop. Data from the syndromic surveillance suggest that the outbreak was much more widespread at the start of April 2020, about a month after the first confirmed case, than was evident from the data on confirmed cases (Figure 2). Confirmed cases remained low in the month following the first detected case on 8 March, reflecting the very limited testing capacity at the start of the outbreak. Initial testing was limited to travellers arriving from abroad and known contacts with confirmed cases. In fact, until 27 March, a single laboratory was responsible for administering and analysing COVID-19 tests in Bangladesh, and only about 1,600 total samples were tested nationally through the end of March.

The number of people reporting multiple symptoms and classified as high risk showed an increasing trend over the specified epidemic time period (15 April to 15 June) (Figure 2). Testing was also largely limited to Dhaka until the end of March. While the number of laboratory-confirmed cases remained low in March and April, even as early as 1 April (when the syndromic surveillance system was not yet fully set up), nearly all upazilas had individuals reporting symptoms consistent with COVID-19.

---

**Figure 2: Time series and spatial distribution of confirmed cases, total number suspected based on reported symptoms and total number classified as high risk**

![Time series and spatial distribution of confirmed cases, total number suspected based on reported symptoms and total number classified as high risk](image)

Note: The blue lines in the trend graphs are the seven-day rolling average

Source: Director General of Health Services (DGHS), hotline equipped with an interactive voice response (IVR) system (333), internet and mobile applications and unstructured supplementary service data (USSD) based messaging system (*3332#*) verified by clinicians
The correlations between the lab-confirmed cases and each of the syndromic indicators for each subdistrict were calculated weekly to account for the delay in lab confirmed testing and reporting. There was a strong positive association between the confirmed cases and the syndromic data (per-capita number of people reporting symptoms consistent with COVID-19 and per-capita number classified as high risk) with a lag time of one or two weeks (Figure 3).

**Figure 3: Correlation between confirmed cases per 100,000 (summed over a one-week window from 8 June to 14 June) and the per-capita number of people classified as high risk for COVID-19 and the per-capita number of people reporting symptoms consistent with COVID-19 infection**

![Correlation Between Confirmed Cases and Syndromic Data](image)

**Source:** Director General of Health Services (DGHS), hotline equipped with an interactive voice response (IVR) system (333), internet and mobile applications and unstructured supplementary service data (USSD) based messaging system (*3332#*) verified by clinicians

**Early spread of the disease**

The vast majority of upazilas had individuals reporting symptoms as well as individuals classified as high risk for COVID-19 before the first confirmed case was reported, with a median lead time of ten days (Figure 4). Confirmed cases in many upazilas were first reported at the beginning of April, and cases across the country began to increase from mid-April onwards, reflecting the increase in testing capacity and the expansion of testing beyond Dhaka.

To understand the mobility of individuals from Dhaka to other parts of the country until 19 March 2020, before a lockdown was imposed, call detail record (CDR) data of the biggest mobile operator in Bangladesh were utilized. However, the estimation of population mobility might be biased, since the number of mobile phone users and the general population were not equal; additionally, the mobile phone subscriber data was obtained from a single mobile operator.
However, a large portion of Dhaka’s population can be represented by the number of subscribers of this mobile operator, which suggests that relative mobility estimates are likely to be representative. This view is supported by previous work in Kenya, which has also shown that mobility estimates from CDR data are not significantly affected by biases in ownership. Employing previously described methods for the spread of COVID-19, the probability of importing one or more infected cases was also estimated in the first 16 days of the outbreak (i.e. prior to any travel restrictions) for each upazila outside the Dhaka district.

The early spread of the outbreak suggested by the syndromic surveillance was also consistent with the predicted importation across the country based on population mobility data. We found a significant positive association between the probability of an upazila having an imported case by 19 March, prior to the lockdown, and before physical distancing measures and the syndromic surveillance indicators were put into place. (Figure 5).

Token together, these results were indicative of an earlier spread of the outbreak, which was not captured in the official case counts. Our findings were also consistent with genomic data from Bangladesh that also suggested an early spread of the outbreak across the country in late March. These results demonstrate the potential usefulness of participatory surveillance as an early-warning system during the initial phases of an emerging outbreak.

**Figure 4: Spatial variation in the time between the first confirmed case and the first high-risk classification shown in (A) and the respective dates by upazila are shown in (B). Upazilas are ordered according to the date of their first high-risk classification in (B). (C) shows the density plot for the lag between first high-risk classification and first confirmed case. The red line indicates zero days of lag time.**

*Source: Director General of Health Services (DGHS), hotline equipped with an interactive voice response (IVR) system (333), internet and mobile applications and unstructured supplementary service data (USSD) based messaging system (*3332#) verified by clinicians*
Figure 5: Correlation between syndromic data (for the first two weeks that data from all surveillance streams was available: 15 April to 29 April) and probability of importation estimated using mobile phone data (x-axis)

Source: Director General of Health Services (DGHS), hotline equipped with an interactive voice response (IVR) system (333), internet and mobile applications and unstructured supplementary service data (USSD) based messaging system (*3332#) verified by clinicians

Practical implications of the system for decision making

Like many developing countries, Bangladesh also faced numerous crises during the COVID-19 pandemic. There were challenges in breaking the silos among the organizations and establishing collaboration among the public, private and technology sectors as well as academicians to develop a data-driven, decision-making system. For better analysis, the government and private data (public health, CDR, population, etc.) were shared with experts of different organizations and universities and put into place. As a result, maintaining data privacy and security was a challenging issue, since the names, addresses, phone numbers, etc. of citizens were confidential information that were at risk of being shared. The data were anonymized and shared with experts of organizations and academicians (parties). Non-disclosure agreements (NDAs) were signed between the parties related to information sharing on different citizen-centric information systems and services. CDR data represented a huge volume of data and maintenance of the quality of such big data was another challenging issue. To mitigate such a challenge, databases were hosted in the National Data Center, and real IP was assigned to access the server. Analysts checked any anomalies in the data on a regular basis and dealt with these.

The right innovations and whole-of-government approaches drove the nation to melt down boundaries, break the silos and generate unprecedented collaboration among public, private and academic sectors. Given the rapid global increase in the use of mobile phones and access to the internet, these surveillance systems provide a useful complement for more traditional surveillance systems. Due to limited laboratory capacity, traditional syndromic surveillance could not provide adequate information about the spread of the pandemic. However, self-reporting using triage allowed rapid data collection through widely accessible devices, such as phones, smart phones and webpages. This data collection process has the added advantage of being less costly to both the respondents and the collecting agency, owing to the digitization of the process.

Data from various sources like test data, test positivity data, death data, hospitalization capacity, equipment’s availability, etc. from the Director General of Health Services are periodically incorporated to develop the system as a proper preventive measure so that decisions can be made immediately to combat any disease pandemic arriving in the future without
needing to create a similar type of new system. The unprecedented collaborations implemented the participatory syndromic surveillance in a matter of weeks, which later developed into the “National COVID-19 Data Intelligence System”. This intelligence system enabled the development and administration of innovative solutions, such as syndromic surveillance, mortality surveillance, contact tracing, epidemiological modelling and health facility management.

The National COVID-19 Data Intelligence System allowed policymakers to identify the hot-spots before the outbreak of the disease in those particular spots at the local levels (upazillas and wards in city corporations). The hot-spots were then identified as red zones and their adjacent areas as yellow zones. The government then imposed lockdowns and movement restrictions in those areas to reduce disease transmission. Preliminary applications of the system were conducted in two cases: the surveillance system in the Rajabazar and Wari areas of Dhaka city were identified as high-risk zones based on the call density and risk profiles, which led to the zoning and lockdown of the areas for 14 days. Since the disease progression can be predicted seven to ten days ahead of testing, hospital facilities could be prepared prior to filling up, based on the real-time data to tackle emergency crisis periods and save lives.

In later days, the COVID-19 Intelligence System integrated the COVID-19 test results data from all the RT PCR labs of the country and mobility data from telecom operators into the system. It integrated the efforts of vast numbers of epidemiologists, data scientists, economists and researchers resulting in the “National COVID-19 Dashboard” for policymakers. This dashboard not only visualizes the analytics but also generates policy recommendations for policymakers at the field levels (field administrators). Currently, 64 civil surgeons, 64 deputy commissioners, 8 divisional commissioners, 8 divisional health officers and secretaries from various ministries are using the dashboard. It has led to the geospatial identification of the most vulnerable people and provided policymakers with the tool to adopt recovery plans for the people who need it the most.

Key policy messages

**Augmentation by innovative solutions for strategic preparedness**

Self-reports indicated high-risk COVID-19 cases ten days before the laboratory tests confirmed the case counts. Combined with analysis of mobility of the population, the “hot-spots” with high caseloads could be identified beforehand. Similar analytical methods can be applied to monitor other disease outbreaks. More advanced analysis can be applied through mortality surveillance, contact tracing and epidemiological modelling to more accurately identify hot-spots, track suspected cases and forecast disease progression even ahead of testing. This can inform policymakers as well as communities to enact appropriate response measures, recovery plans and prepare for expected caseloads. For health facilities, data sources can be expanded to include hospitalization capacity and equipment availability for facilities to plan and allocate resources according to expected caseloads, such as deploying more medical personnel and ICU equipment.

**Minimization of noise and volatility of self-reported syndromic surveillance data**

While there could be dynamics of the pandemic and surveillance data yet to be observed, self-reported data were positively correlated with confirmed cases. Another comparative strength of the data is in its scalability. However, self-reported data can have initial volatility and “noise” that need to be verified and “cleaned”. The self-reported symptoms are extremely noisy and unstable. When the system was first introduced, people were used to reporting their symptoms incorrectly due to their ignorance about the COVID-19 symptoms, checking whether the system worked properly, making fun of it, etc. The noise and instability of the data reduce the predictive power and acceptability of the data. The awareness among people needs to be raised through various types of awareness-building activities to frame the symptoms correctly and reduce the unnecessary calls.

**Inclusion of women, older and marginalized people**

In a country like Bangladesh, older people, women and marginalized people remain under-reported in the CDR and the mobile data. Older people’s and women’s subscriber identity module (SIM) cards are not registered by their names; instead, they are registered by their husband’s or son’s name. The marginalized people are excluded, because they do not have access to all government services and some of the services are not available to them. This reduces the coverage and representativeness of the CDR and mobile data. While implementing the participatory syndromic surveillance, clinicians verified the information given by the citizens.
and corrected the names, addresses and other important information to ensure coverage and representativeness. Union digital centers situated at the union level of the country are now providing digital government services to the marginalized people effortlessly and helping them give information about their COVID-19 symptoms to the hotline numbers and applications.

Government can cap the registration of at most two SIM cards against each national identity (NID) card. During bio-metric registration of each SIM card, the fingerprints of a SIM card buyer can be verified with the fingerprints given by the same citizen during the smart NID registration to see whether the same person has purchased the SIM card or not. As a long-term mitigation process, the government and UNDP are working for improving the rights of the marginalized people and giving them access to all essential services. The network coverage of the mobile phone operators needs to be improved in the remote and hard-to-reach areas.

Building analytical capacity among the public servants

The huge volume of mobile and CDR data (Big Data) is flowing from the technology sector and other private companies like the mobile network operators (MNOs), but not from the public sector. Consultants are hired and their analytical capacities are outsourced. Therefore, there is no opportunity to build the in-house capacity related to statistical, data or other skills within the civil service. There are risks of losing data authenticity, privacy and security. Several capacity building trainings are needed for public servants to enhance their data management and analysing capacity. During recruitment of public servants, the technical qualifications of the applicants should be considered.

Increase the practice of using updated data

There is a challenge in using the population data. The actual population data comes from census 2011 data, which is quite old. The projected population data has been used here as an alternative source. The technology sector can play a significant role in collecting the actual number of people living in the community, geographic boundaries of the community, etc.

Ensure data privacy and security

To store and process large data-sets, reliable computing system software like Apache Hadoop can be used. This software allows for the processing of varied sources of data with low cost and speedy data processing techniques. In order to maintain the privacy and safeguards with regard to CDR data, FlowKit, an open-source tool developed by Flowminder can be used. This program allows analysts to examine the de-identified mobile data behind the MNOs’ firewalls. FlowKit can be installed in the National Data Center to process and analyse the data at a low cost and with high efficiency.

Conclusion

The traditional surveillance system needs patients to interact with the healthcare system where there are challenges of testing facilities and reporting delays. On the contrary, participatory surveillance depends on self-reporting of symptoms, which is real-time and based on crowdsourcing. Mobile phones and the internet are being used on a massive scale nowadays. Symptoms reported through phone hotlines, mobile phone applications and the internet play a significant role in complementing the traditional surveillance system.

Noise and volatility were found in different reporting mechanisms. The classification of the high and low risks is based on the judgement of the human verifiers. Also, the use of the system may depend on the timing of advertising of the system, awareness of the COVID-19 cases in the locality, etc. Nevertheless, the self-reported system showed a positive correlation with confirmed cases at the subdistrict level seven days later. This surveillance system proposes an early spread of the outbreak that is evident from the confirmed case counts. Combining with other syndromic surveillance data from clinics and hospitals, hospitalization and death rate, etc., early prediction of the outbreak helps in controlling the disease and taking preventive measures as necessary. Hospital beds and necessary medical equipment can be arranged in advance to manage patient overloads. This system also enables the community support team across the entire country to increase awareness and promote non-pharmaceutical interventions.

This surveillance system can be replicated in disease outbreaks where testing capacity is limited and symptoms are relatively non-specific. The system of controlling disease based on crowdsourced syndromic surveillance used here can be simulated in future pandemics in Bangladesh and globally in other countries.