SMJ Singapore Medical Journal

Leveraging electronic medical records for passive disease surveillance in a COVID-19 care facility

Hao Sen Andrew Fang[1], MBBS, MMed, Jonathan Kia-Sheng Phua[2], MBBS, Terrence Chiew[3], MBBS(Hons), Daniel De-Liang Loh[4], MBBS, MRCS, Ming Han Lincoln Liow[5], MBBS, FRCSEd, Weien Chow[6], MBBS, MMed, Xian-Yang Charles Goh[7], MBBS, FRCR, Hian Liang Huang[7], MBChB(Hons), MMed

[1]SingHealth Polyclinics, [2]Department of Diagnostic Radiology, Singapore General Hospital, [3]Singapore National Eye Centre, [4]Department of Neurosurgery, [5]Department of Orthopaedic Surgery, Singapore General Hospital, [6]Department of Cardiology, Changi General Hospital, [7]Department of Nuclear Medicine and Molecular Imaging, Singapore General Hospital, Singapore

Correspondence: Dr Fang Hao Sen Andrew, Hybrid Doctor, Doctor Anywhere, 460 Alexandra Road, mTower #40-01, Singapore 119963. andrew.fang@doctoranywhere.com

Singapore Med J 2022, 1–11
https://doi.org/10.11622/smedj.2022010
Published ahead of print: 10 February 2022

INTRODUCTION

On March 11, 2020, the World Health Organization (WHO) declared the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) outbreak as a pandemic. Singapore, a city-state and international travel hub reported its first imported case on January 23, 2020 and was one of the first countries to be affected by Coronavirus Disease 2019 (COVID-19). Since then, there have been 59,975 confirmed cases and 28 deaths in Singapore, as of this writing on Oct 27, 2020.

In order to contain the surge of COVID-19 infection and preserve hospital capacity for critically ill patients, Singapore's Ministry of Health (MOH) set up several out-of-hospital isolation facilities.[1] Community care facilities (CCFs) were designed to accommodate COVID-19 patients with mild symptoms and low risk factors. These CCFs provided capacity to hold up to 20,000 patients for the duration of the isolation period.[1]

Closely shared living spaces within CCFs posed an increased risk of communicable disease spread among residents. This drove the need to develop a disease outbreak surveillance capabilities to provide early detection of potential outbreaks within the CCF. With a lean headcount, our team was inspired to leverage available electronic medical records (EMR) data to develop a passive disease surveillance system (DSS).

METHODS

Singapore EXPO, an exhibition hall venue was repurposed into a 8,000 bed CCF (CCF@EXPO). SingHealth was tasked by MOH to manage the healthcare needs of 3,200 residents housed in four halls. The halls were organized into rows and rooms. Each room measured 2.4- x 3.6-m and was shared by two residents, separated by partition boards. Residents were young to middle-aged adults who had tested positive for COVID-19 infection and were generally well or had mild symptoms.

At the time, Singapore's policy was to isolate COVID-19 positive individuals for twenty-one days from the day of symptom onset or diagnosis, whichever earlier, before they could be discharged. Primary healthcare teams provided basic medical screening at point of admission and primary-care services for CCF@EXPO residents throughout their stay. Physicians utilized GP Connect (GPC), which was a primary-care clinic based EMR system developed by Integrated Health Information Systems (IHiS), and sponsored by MOH, to perform clinical documentation, medication management and results review for CCF residents. Residents were required to measure their own vital signs, including temperature, and input their results into an electronic data capture system (Health Discovery) at least once a day. IHiS, which managed GPC and Health Discovery in CCF@EXPO, aided in the extraction of EMR and vital signs data used for this project.

The healthcare management team convened daily to discuss outstanding issues in the CCF. Daily updates included admission, discharge, occupancy and primary care statistics. Bearing in mind the increased risk of communicable disease spread within the CCF, the management team requested clarity on the spectrum of primary care cases seen at the sickbay, highlighting the need to detect potential communicable disease outbreaks. This led to the development of our passive disease surveillance system (DSS).

We designed the DSS with the following considerations:

1. Comprehensive and early detection of communicable diseases prevalent in CCF@EXPO
2. Efficient passive reporting by leveraging EMR data
3. Able to provide intuitive insights through spatiotemporal information
4. Low-cost of development

To ensure that the system was targeted in its coverage of prevailing communicable diseases, we consulted an Infectious Disease specialist whom advised us to monitor for

gastroenteritis, chickenpox, measles, mumps, rubella, dengue and scabies outbreaks. Influenza was excluded as all residents were COVID-19 positive and were expected to have acute respiratory symptoms. We decided to implement syndromic surveillance, in addition to passive surveillance. Syndromic surveillance enabled early detection of outbreaks before confirmed diagnoses were made.[2,3] Two disease syndromes – (1) acute diarrhoeal illness and (2) potentially infectious rash were identified for syndromic surveillance.

For syndromic surveillance, we leveraged EMR data to analyse (1) visit dates, (2) patient location and (3) finalised diagnoses as input. From a restricted list of 1,306 Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) primary care diagnoses available in GPC, we studied recent records and adopted a consensus-driven process of elimination to select a list of diagnoses as indicators for the two disease syndromes of interest. Table I lists the diagnoses that were used as indicators for syndromic surveillance.

In addition to syndromic surveillance, we also monitored for cases of fever, which we defined as temperature ≥ 37.5 degree Celsius. We used vital signs data from Health Discovery for the fever monitoring. Fever was selected as it is known to be a common and early objective indicator of infection, manifesting before other vital signs would turn abnormal.

With the knowledge that communicable diseases were likely to cluster in space, adding spatial information to traditional time-based methods would provide additional power and efficacy in detecting outbreaks.[4] A recent review found about a third of public health surveillance algorithms made use of spatial information.[5] We thus designed the system to generate two outputs: (1) a control chart to identify time-based aberrations and (2) a geospatial map of cases to highlight any physical clustering of cases.

For the control charts, we analysed historical data and computed the mean and standard deviation of the daily number of cases for each of the two disease syndromes. An initial threshold of two standard deviation from the daily mean was chosen as a reasonable balance between sensitivity and specificity. As the daily numbers followed a normal distribution curve, we expected a trigger rate of about 2.5%. The absolute numbers were also tracked prospectively to evaluate the need to adjust the threshold.

For the geospatial map, we used the CCF@EXPO hall bed layouts and highlighted positive cases within windows of the past five and three days for syndromic surveillance and fever surveillance respectively. (Fig. 1) Unique maps were generated for each disease. Diseases were tagged to bed location and flagged as coloured triangles to allow cluster detection by visual inspection. The reason for the look-back window was based on the rationale that cases of transmissible diseases tend to disseminate and develop over a period of time, instead of all within the same day.

To address the consideration of cost, we developed the system with Python 3.7, which is an open-source programming language. In addition, we used the following Python packages: “xlwings” package for reading the EMR data, the “numpy”, “pandas” and “re” packages for data cleaning, transformation and analysis, and the “matplotlib” and “seaborn” packages for data visualisation.

RESULTS

The DSS was a lightweight Jupyter notebook file. It only required users to open the file, select the path to the raw EMR data file on the local computer and then click to run. The application would

load the abovementioned Python packages and raw data, transform the data and finally generate the control charts and geospatial maps.

The process from parsing the raw EMR data to generating the outputs went through four phases as a seamless data pipeline. In the first phase, the data was imported to the Python session and converted to data objects ('dataframes') that could be manipulated using Python. In the second phase, the data was cleaned to exclude non-valid cases (e.g. discharges), to extract the data values of interest (i.e. visit date, bed location and diagnoses) and to vectorise bed locations (e.g. by x- and y- dimensions). As the EMR data was generated from real-world processes, there were cases which had been incorrectly registered multiple times, so unique identifiers were used to de-duplicate them. In the third phase, the data was aggregated by counting the cases-of-interest in a time-series manner, setting the stage for the control chart to be generated. Finally, the fourth phase created the visualizations by plotting out the control chart and geospatial map. A modified Jupyter notebook file and a mock dataset are available for download on GitHub (see Availability of Data and Materials section below.)

As part of a daily workflow, IHiS would provide our team with a password-protected file containing EMR data of primary care visits from the previous day. As these files contained patient identifiers, they were stored on and accessed via dedicated password-locked laptops by assigned team members. Our team would use the DSS to parse the EMR data which generated the control chart and geospatial maps. Extracting the (1) visit dates, (2) patient location and (3) finalised diagnoses from the EMR data and using these as input to the system, control charts and geospatial maps were generated near instantaneously. The control charts and geospatial maps were then sent to an Infectious Disease specialist for review. If it was deemed that there was potential outbreak, the medical and infection prevention and epidemiology teams would be alerted immediately for

follow-up action. The control chart and geospatial maps were also compiled into daily updates to the healthcare management team.

During the initial two weeks, there were no triggers from the DSS of potential disease outbreaks. In order to test the DSS' performance, we simulated a gastroenteritis outbreak. We conducted the simulation by first defining a time and location of the outbreak. Next, dummy cases of gastroenteritis were inserted into the extracted EMR data within the vicinity of the cluster over a five day period. This was done by entering a variety of the acute diarrhoeal illness diagnoses (Table I) into the dataset. We then serially ran the DSS on this modified dataset over the days prior, during and after the simulated outbreak period. The results from the test showed that the DSS was able to detect the outbreak within a day. It was also able to track the progression of the cluster which would have helped in contact tracing and investigation of the source. (Fig. 2)

DISCUSSION

We described the design of an easy-to-use, low-cost disease outbreak surveillance system that leverages EMR data to provide spatiotemporal information of recent cases. It was successfully implemented as part of our healthcare operations in managing COVID-19 patients within a CCF, and was demonstrated to be capable of detecting potential outbreaks early in a simulated model.

Our low-cost disease surveillance system which leverages EMR data and open-source software can easily be modified for use in similar healthcare operations, such as humanitarian missions and military medical operations, or in resource-limited settings, as long as data with date, location and diagnosis variables were available. Whilst several open-source tools for disease surveillance already exist, the strength of our DSS is that it leverages raw extracted EMR data to reduce inefficiency of manual data entry or cleaning, is EMR agnostic and can be run on computers

without internet connectivity.[6,7] In fact, even settings that do not use EMR, our DSS can still be applied as long as an electronic database of cases with time, location and diagnosis data is available.

From the geospatial maps, it can be seen that our DSS worked well when the population under surveillance was evenly spread out and individuals occupied exclusive spaces. Such a grid-like arrangement would also be typical in healthcare settings like hospitals and domiciliary care institutions. Considering its use in other settings where dense and sparse occupancy areas are inter-mixed, we would still expect our DSS to function well. Moreover, with the ability to adjust marker transparency, cases (coloured markers) that overlapped would be given greater prominence, producing what would be an impressive heatmap visualisation.

This study was not without limitations. The interpretation of each disease surveillance system outputs needs to be done in the context in which it is applied. In our CCF operations, our DSS used residents' bed location as a proxy for their whereabouts, while in reality residents were free to move around within their designated hall. The residents also shared amenities and equipment, such as toilets and vital signs monitoring devices. The fluid, dynamic mixing and interaction among residents could potentially stymie the early predictive capability of the geospatial maps in cluster detection. Nevertheless, from the control charts, our DSS still preserved the capability to detect any intra-hall aberration in daily case numbers to mitigate this limitation. We also acknowledge that our DSS has not been systematically and rigorously back-tested on real-world data. This is an area for further improvement.

The current iteration of our DSS relies on heuristics like visual inspection to detect outbreaks from spatiotemporal information. In this area, research has been done to develop and validate alternative statistical methods for spatiotemporal disease surveillance. Such methods

include cumulative sum, scan statistics and model-based algorithms.[8-10] These methods could be incorporated to further systematize and improve the DSS performance.

Moving forward, we are working with a vendor to enhance our DSS with 3-dimensional mapping and artificial intelligence capabilities so that it can be used in multi-level settings such as hospitals and larger care facilities.

In conclusion, in managing the medical operations of a COVID-19 isolation facility, we were motivated to develop a passive disease surveillance system that leverages EMR data. We have described the design and successful implementation of a DSS in a real-life medical operation. In the spirit of the open science movement, we have made our work open-source so that others will be able to modify it for their own disease surveillance purposes.

AVAILABILITY OF DATA AND MATERIALS

A modified Jupyter notebook file and a mock dataset are available for download on GitHub (https://github.com/andrewfanghs/pydoss).

REFERENCES

1. Chia ML, Chau DH, Lim KS, et al. Managing COVID-19 in a novel, rapidly deployable community isolation quarantine facility. Ann Intern Med 2020; 174:247-51.
2. Pavlin JA. Investigation of disease outbreaks detected by “syndromic” surveillance systems. J Urban Health 2003; 80(2 Suppl 1):i107-14.
3. May L, Chretien JP, Pavlin JA. Beyond traditional surveillance: applying syndromic surveillance to developing settings--opportunities and challenges. BMC Public Health 2009; 9:242.

4. Sparks R, Bolt S, Okugami C. Spatio-temporal disease surveillance. In: Morse S, ed. Bioterrorism. IntechOpen, 2012: 159-78.
5. Yuan M, Boston-Fisher N, Luo Y, Verma A, Buckeridge DL. A systematic review of aberration detection algorithms used in public health surveillance. J Biomed Inform 2019; 94:103181.
6. Nieves E, Jones J. Epi Info™: Now an Open-source application that continues a long and productive "life" through CDC support and funding. Pan Afr Med J 2009; 2:6.
7. Hodanics C. OpenESSENCE: disease surveillance through medical record system integration with OpenMRS. Online J Public Health Inform 2014; 6:e156.
8. Robertson C, Nelson TA, MacNab YC, Lawson AB. Review of methods for space–time disease surveillance. Spat Spatiotemporal Epidemiol 2010; 1:105-16.
9. Zhang HL, Lai SJ, Li ZJ, Lan YJ, Yang WZ. [Application of cumulative sum control chart algorithm in the detection of infectious disease outbreaks]. Zhonghua Liu Xing Bing Xue Za Zhi 2010; 31:1406-9. Chinese.
10. Kulldorff M, Heffernan R, Hartman J, Assunção R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS Med 2005; 2:e59.

Table I. List of diagnoses from electronic medical records for syndromic surveillance

Syndrome	Acute diarrhoeal illness	Potentially infectious rash
Diagnosis (SNOMED-CT Concept ID)	• Acute diarrhoea (409966000) • Diarrhoea (62315008) • Enteritis (64613007) • Gastroenteritis (25374005) • Vomiting (422400008)	• Dengue (38362002) • Disorder of skin (95320005) • Eruption (271807003) • Herpes zoster (4740000) • Herpes zoster without complication (111859007) • Infestation by Sarcoptes scabiei var hominis (128869009) • Itching (418290006) • Itching of skin (418363000) • Measles (14189004) • Measles without complication (111873003) • Rubella (36653000) • Varicella (38907003)

SNOMED-CT = Systematized Nomenclature of Medicine – Clinical Terms

Fig. 1 This figure shows the conversion of the actual CCF@EXPO layout (left image) to a 2-dimensional blueprint (right image) which was then used to guide development of the geospatial maps.

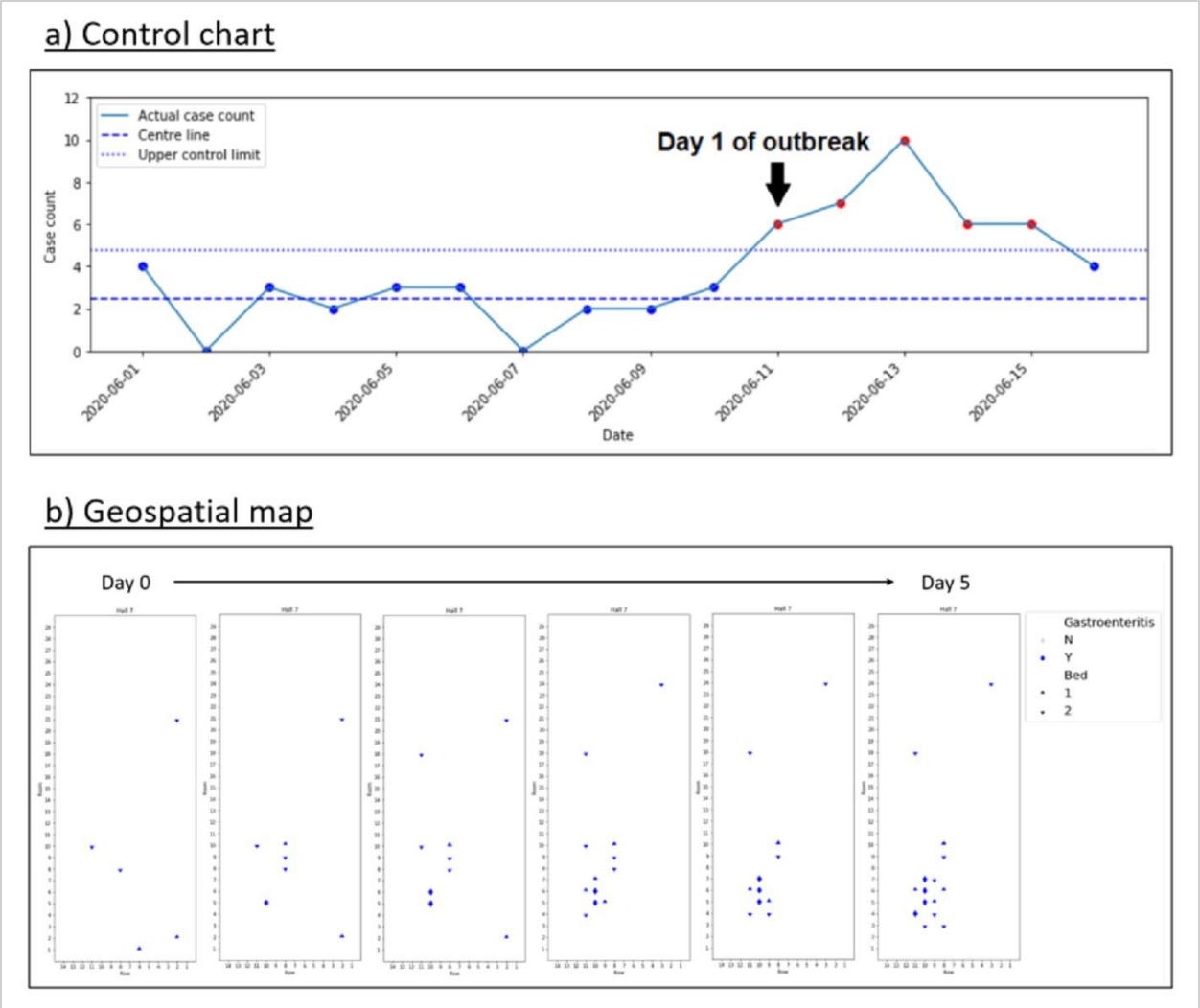

Fig. 2 This figure shows the results from testing the disease surveillance system using a simulated gastroenteritis outbreak. In the control chart (top), the system was trigged within the first day as the case count exceeded the upper control limit. From the geospatial map (bottom), the outbreak cluster progression can be tracked.