Rapidly scalable and low-cost public health surveillance reporting system for COVID-19

Vivek Jason Jayaraj, Chiu-Wan Ng, Victor Chee-Wai Hoe, Diane Woei-Quan Chong, Sanjay Rampal

ABSTRACT

Objective Data-driven innovations are essential in strengthening disease control. We developed a low-cost, open-source system for robust epidemiological intelligence in response to the COVID-19 crisis, prioritising scalability, reproducibility and dynamic reporting.

Methods A five-tiered workflow of data acquisition; processing; databasing, sharing, version control; visualisation; and monitoring was used. COVID-19 data were initially collated from press releases and then transitioned to official sources.

Results Key COVID-19 indicators were tabulated and visualised, deployed using open-source hosting in October 2022. The system demonstrated high performance, handling extensive data volumes, with a 92.5% user conversion rate, evidencing its value and adaptability.

Conclusion This cost-effective, scalable solution aids health specialists and authorities in tracking disease burden, particularly in low-resource settings. Such innovations are critical in health crises like COVID-19 and adaptable to diverse health scenarios.

INTRODUCTION

The transmission of COVID-19 has led to more than 535 million cases and 6 million deaths in almost every country in the world between January 2020 and June 2022. Data and digital solutions are essential in strengthening pandemic management and preparedness. Dynamic reporting systems have been used as a data solution for decision-making within various fields. Dynamic reporting systems as part of surveillance systems can be critical for disease control, resource allocation and health planning. However, developing such a system within lower-income and middle-income countries can be challenging due to the scarcity of human and economic resources. Malaysia has experienced high levels of COVID-19 transmission. Despite this, there has been a paucity of data collation and translation into epidemiological intelligence for public health interventions within the public space. We aimed to develop an analytical platform that had: (1) a data collation system, (2) an automated analytical and reporting platform, and (3) a system to monitor and evaluate the solution. The solution had to be low-cost and rapidly deployable without programming expertise.

METHODS

Solution architecture

A five-tiered workflow was developed using open-source tools that allowed work process automation. These tiers include: (1) data acquisition; (2) data processing; (3) databasing, sharing and version control; (4) data visualisation; and (5) dashboard deployment and monitoring (figure 1). Data acquisition was performed with automated scripts using the ‘rvest’ and ‘magick’ packages for web scraping and image handling, respectively. The collected data were then processed using the ‘tidyverse’ and ‘markdown’ packages to facilitate data manipulation and dynamic report generation. These clean, structured data were managed using GitHub for databasing, sharing and version control, ensuring data integrity and accessibility for collaborative efforts. Data visualisation transitioned from the flexdashboard framework to the Shiny package for more interactive web applications. The dashboard was initially hosted on GitHub pages, subsequently migrating to shinyapps.io for improved performance. An in-line frame plugin was used to embed the dynamic reports within the Universiti Malaysia’s Department of Social and Preventive Medicine’s WordPress website (https://spm.um.edu.my/knowledge-centre/covid19-epidive/). Google Analytics plugins monitored the dashboard, tracking usage and user interactions. This streamlined, automated workflow was adapted from various published methods, ultimately creating a comprehensive tool for epidemiological intelligence.
for the solution are available from https://github.com/spm-um/c19-epi4msia.

Data source and epidemiological indicators

Reporting of COVID-19 data in Malaysia has been carried out by the Ministry of Health (MOH), Malaysia, since January 2020 via daily press releases (https://kpkesihatan.com/), infographics (https://COVID-19.moh.gov.my/) and an official instant messaging channel (https://t.me/cprckkm). Data extraction workflow was developed to scrape case, death, testing, variant and vaccination using open-source tools beginning in September 2020. Cluster data were extracted manually daily as the MOH published these in a non-machine readable format. Data on healthcare capacity for certain states were sporadically provided via press releases and extracted when released. Data on state healthcare capacity, with no reported press release data, were extracted from previous national surveys and reports.12–14 Data on midyear populations were extracted from the population projections provided by the Department of Statistics, Malaysia.15 Additionally, data on mobility (https://www.google.com/COVID-19/mobility/) provided by Google are extracted directly from the Google data repository. Reporting on variants by the MOH was stopped in February 2022. Data on variants of concern were subsequently extracted from the GISAID Initiative (https://www.gisaid.org). Scripts were executed automatically on a daily basis to extract COVID-19 data from the above sources. Since 21 July 2021, the MOH, Malaysia has consolidated these data sources into a single GitHub data repository (https://github.com/MoH-Malaysia/covid19-public), and as such, the crawler was retired on 31 August 2021. Data input for the dashboard was subsequently shifted to the new MOH data repository. Datasets, including retired datasets and code used, are available on GitHub (https://github.com/spm-um/c19-epi4msia-data). A random 10% of data points were validated each week to ensure accuracy. Data sources and indicators considered important for policy response were determined from a review of reports, dashboards, advisories and expert epidemiologist opinions with frequent updates based on the changing science16–18 (online supplemental appendix 1).

Monitoring and evaluation

Google Analytics plugins were used to monitor the dashboard’s utilisation in real time. Indicators monitored include cumulative views, new views, active views, sessions per user, user location, device and platform used in engagement, engagement event type and traffic comparison with other website elements. Additionally, user experience forms were made available to all users to identify potential issues and gaps for improvement as rapidly as possible.

RESULTS

The architecture, developed in October 2020, initially provided semiautomated updates on COVID-19 data and analytics. With Malaysia’s transition into endemity in April 2022, the dashboard underwent enhancements, deploying extended functionalities in June 2022. This fully automated version collated and published analytics on transmission, disease burden, testing, healthcare capacity, variants and human mobility, using a variety of visual aids (online supplemental appendix 2). Using a web-based plugin, usage statistics were monitored, yielding weekly and biannual performance reports. Remarkably, 92.5% of new users converted into active users, with changes in user engagement paralleling shifts in measures for uptake and transmission dynamics (online supplemental appendix 3). The dashboard drew a median of 0.37% of all traffic to the department website, peaking at 1.9% in November 2020 (online supplemental appendix 4).

DISCUSSION

The escalating COVID-19 transmission within Malaysia necessitated the development of a local, low-cost, rapidly scalable and reproducible solution for data collation,
analysis and dashboarding. This initiative was a collaboration of public health experts which yielded a straightforward, agile architecture that can be deployed with limited programming expertise. This open-source model demonstrated robust performance, handling large volumes of data, converting 92.5% of new users into active ones and reflecting stakeholder feedback in its ongoing enhancements. Compared with other open-source solutions in different countries, development by non-expert programmers was the key differentiator, offering potential for use in other low-resource settings. Smaller disease control programmes can use this workflow in developing dynamic reporting systems in said programmes.19 20

Several limitations were observed, including data acquisition challenges, relatively low engagement, limited monitoring of data volumes and collection efficiency, constraints in user experience. Lessons learnt from this project emphasised the importance of stakeholder involvement in the development process, an effective engagement strategy prior to deployment and the necessity of flexibility in the face of changing data formats.

Future iterations should incorporate these insights, and we propose four potential solutions to mitigate current limitations. First, in response to the volatile data sources, improving the system’s adaptability to various data formats is essential. Second, to boost engagement, an inclusive and robust communication strategy must be crafted. Third, detailed data volumes and exact efficiency metrics should also be monitored for a more comprehensive understanding of the system’s impact. Lastly, the user interface can be refined to enhance user experience and stability, without compromising speed and simplicity. Several of these issues were partially addressed in the recent dashboard update of the MOH’s recent use of the GitHub platform for data dissemination and extend functionalities to include more depth in the exploration of indicators.

The value of agile, open-source solutions like ours becomes apparent in combating diseases like COVID-19. The tool’s adaptability makes it beneficial for tracking various health concerns, underscoring the importance of continued development and enhancement.

CONCLUSION

The development of a low-cost, open-source, rapidly scalable and reproducible workflow as described here can be useful in many situations, especially in low-resource settings. Innovative and low-cost solutions can be critical in crises such as COVID-19 and other health- and healthcare-related scenarios, especially in automating collation, processing and reporting of data. There remains a need for more collaboration in strengthening workflows such as this to allow more rapid adoption of dynamic reporting systems in the health sector.

Acknowledgements The authors would like to thank the Director-General of Health Malaysia for his permission to publish this article.

Contributors Conceptualisation—SR and VJ-CH. Data curation—VJJ. Formal analysis—SR and VJ-CH. Funding acquisition—SR. Investigation—VJ-CH. Methodology—SR, VJ-CH and VJJ. Project administration—C-WN. Resources—C-WN. Software—VJJ. Supervision—SR, VJ-CH and C-WN. Validation—SR and DW-OC. Visualisation—VJJ. Writing (original draft preparation)—VJJ. Writing (review and editing)—C-WN, SR, VJ-CH and DW-OC. All authors read and agreed to the published version of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Datasets, including deprecated datasets and code used, are available on GitHub (https://github.com/spm-um/c19-epi4msia-data).

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD Vivek Jason Jayaraj http://orcid.org/0000-0002-8131-9878

REFERENCES


17 World Health Organisation, 2021. Available: https://covid19.who.int/?gclid=Cj0KCQiwub-HBhCysApctr7wt4KGje6JUmyWQjSbMQWh8jvXTZSdsSR5ay9OAp2p1uOWMFJb0aAmAXEALw_wcB


**Supplementary Appendix 1**

The 7-day incidence and mortality rate, test positivity ratio and testing rate, hospital occupancy rate, ICU occupancy rate, ventilator occupancy rate, percentage of population vaccinated, and the ratio of variants to total cases are given in Supplementary table 1.

### Supplementary table 1: Epidemiological indicators of transmission

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Function</th>
<th>Base unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-day incidence rate</td>
<td>( \frac{\text{No. of reported cases (in time period)}}{\text{Mid \ – \ year population}} \times 100,000 )</td>
<td>Per 100,000 population</td>
</tr>
<tr>
<td>7-day mortality rate</td>
<td>( \frac{\text{No. of reported mortalities (in time period)}}{\text{Mid \ – \ year population}} \times 100,000 )</td>
<td>Per 100,000 population</td>
</tr>
<tr>
<td>Test positivity ratio</td>
<td>( \frac{14 \text{ – day lagged reported cases (per day)}}{\text{Mid \ – \ year population}} )</td>
<td>Percentage</td>
</tr>
<tr>
<td>Testing rate</td>
<td>( \frac{\text{No. of individuals tested (per day)}}{\text{Mid \ – \ year population}} \times 1,000 )</td>
<td>Per 1,000 population</td>
</tr>
<tr>
<td>Hospitalisation bed utilisation rate</td>
<td>( \frac{\text{No. of admitted patients}}{\text{No. of hospital beds available}} \times 100 )</td>
<td>Percentage</td>
</tr>
<tr>
<td>ICU bed utilisation rate</td>
<td>( \frac{\text{No. of patients admitted in ICU}}{\text{No. of ICU beds available}} \times 100 )</td>
<td>Percentage</td>
</tr>
<tr>
<td>Ventilator utilisation rate</td>
<td>( \frac{\text{No. of patients utilising ventilators}}{\text{No. of ventilators available}} \times 100 )</td>
<td>Percentage</td>
</tr>
<tr>
<td>Percentage vaccinated</td>
<td>( \frac{\text{(1st or 2nd dose cumulative up to date)}}{\text{Mid \ – \ year population (Adult or Total)}} \times 100 )</td>
<td>Percentage</td>
</tr>
<tr>
<td>Ratio of variants to cases</td>
<td>( \frac{\text{Cumulative no. of variants detected}}{\text{by strain in time period}} )</td>
<td>Ratio</td>
</tr>
</tbody>
</table>

The Cori et al. approach was used to estimate the time-varying (instantaneous) reproduction number (R\(_t\)) on 7-day sliding intervals, with 95% credible intervals. This approach used a time-series of daily case data by reporting date and a serial interval distribution assumed to follow a discretised gamma distribution. It then models the number of secondary infections at each time step relative to the number of primary infections and a period of infectiousness represented by the serial interval.

The time-varying reproduction number is defined as the fraction of the expected number of secondary infections at time \( t \) over the number of infected individuals weighted by their relative infectiousness at time \( t \), given by the generation or serial interval distribution. 44 Serial intervals from a review of published and unpublished literature were assumed to fit the Malaysian profile of cases.

An R\(_t\) of more than 1 suggests that the epidemic will continue to grow, whilst an R\(_t\) of less than 1 suggest transmission is in decay. The R\(_t\) is given by:

\[
R_t = \frac{I_t}{\sum_{s=1}^{t} I_{t-s} w_s}
\]

Where \( I_t \) is the number of infections on day \( t \), and \( w_s \) is the generation interval of \( s \) days separate an infector-infectee pair.
Supplementary Appendix 2

Supplementary table 2 Characteristics of dashboard views

| Period              | Overall, N = 919
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>---------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td><strong>Countries</strong></td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td>871 (96%)</td>
</tr>
<tr>
<td>Others*</td>
<td>41 (4.5%)</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td></td>
</tr>
<tr>
<td>Department website</td>
<td>1 (0.1%)</td>
</tr>
<tr>
<td>Direct*</td>
<td>608 (66%)</td>
</tr>
<tr>
<td>GitHub</td>
<td>310 (34%)</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>911 (99%)</td>
</tr>
<tr>
<td>Non-English*</td>
<td>8 (0.9%)</td>
</tr>
<tr>
<td><strong>Device</strong></td>
<td></td>
</tr>
<tr>
<td>Desktop</td>
<td>763 (83%)</td>
</tr>
<tr>
<td>Mobile</td>
<td>154 (17%)</td>
</tr>
<tr>
<td>Tablet</td>
<td>2 (0.2%)</td>
</tr>
<tr>
<td><strong>Browser</strong></td>
<td></td>
</tr>
<tr>
<td>Chrome</td>
<td>697 (76%)</td>
</tr>
<tr>
<td>Others*</td>
<td>192 (21%)</td>
</tr>
<tr>
<td>Safari</td>
<td>30 (3.3%)</td>
</tr>
</tbody>
</table>

1n (%)  
2Pearson’s Chi-squared test; Fisher’s exact test  
3 Seven NA values in locations  
*Other countries include Australia (3), Brunei (1), Egypt (1), Germany (1), Indonesia (2), Ireland (1), Japan (1), New Zealand (2), Norway (1), Pakistan (1), Singapore (1), South Korea (1), Sri Lanka (3), Ukraine (1), United States (4). Other languages include Chinese and Korean. Other browsers include Android Webview (14), Edge (3), Opera (2), and Samsung Internet (25)  
^Direct source is defined as the use of the exact URL to the website instead of being rerouted via the department website or from GitHub.
Supplementary Appendix 3

Supplementary figure 2 Dashboard and monitoring platform: a) Original dashboard with cards on daily change in cases, deaths and recoveries and national/sub-national epidemic curves, b) weekly progress report on dashboard, and c) Updated dashboard with a weekly situational report of cases in Malaysia and her states.
Supplementary figure 2 Number of 7-day moving average of 7-day cumulative views of dashboard between 1 October 2020-30 September 2021