PAWS: Towards a globally integrated outbreak surveillance system for public health

Chris von Csefalvay[1]*, Tamas Foldi[1]

Abstract
The novel emergent coronavirus now designated SARS-CoV-2 emerged in Wuhan, Hubei Province, China, in late 2019. Since then, this rapidly spreading pathogen has created one of the most significant pandemics in recent history. Globalisation and international trade have created a unique potential for fast-moving pathogens to spread globally overnight. In this paper, we describe a globally integrated system of data-driven viral surveillance by analogy to the CTBTO's technical infrastructure. Established as a result of the Comprehensive Nuclear Test Ban Treaty signed in 1996, it is backed by a robust, globally distributed monitoring infrastructure that may serve as an inspiration for developing a similar monitoring network for public health. We propose to build on the experiences of the CTBTO's technical infrastructure to conceptualise a Pandemic Advance Warning System, an early warning infrastructure for emerging pathogenic threats.

Keywords
COVID-19 — Biosecurity — Threat reduction — Serosurveillance

[1] *Starschema Inc., Arlington, VA.*
***Corresponding author**: csefalvayk@starschema.net

1. Introduction

In the waning days of 2019, Zhu et al. documented a case cluster of viral pneumonia in three adults in Wuhan, China, with sequencing of samples from bronchoalveolar lavage revealing a novel coronavirus within subgenus *Sarbecovirus*, genus *Betacoronavirus*, subfamily *Orthocoronavirinae*.[?] Tentatively named 2019-nCoV and later named SARS-CoV-2 due to its phylogenetic similarity with the sarbecovirus responsible for the outbreak of Severe Acute Respiratory Syndrome (SARS) in 2002-2004,[?] this novel pathogen has shown surprising virulence, leading to the WHO convening an Emergency Committee on 20 January 2020.[?] By 30 January 2020, following two inconclusive meetings, the World Health Organization declared the outbreak of SARS-CoV-2 and its related syndrome of viral pathogenesis, COVID-19, a public health emergency of international concern (PHEIC), followed by the declaration of a pandemic on 11 March 2020.

For perhaps the first time in the history of public health, near real-time information from a pandemic has been shared by, and with, public health authorities and the wider public. Despite early issues with a lack of suitable testing capability,[?] the volume and velocity of information about the novel coronavirus outbreak became an impressive testament to the ability of modern public health systems to assess, collate and disseminate critical information rapidly.

The challenge to global health security

The rapid global spread of SARS-CoV-2 has highlighted the threat that fast-moving pathogens pose in the age of globalisation and international trade. Writing before the pandemic declaration by the WHO, MacIntyre has identified the high pandemic potential of SARS-CoV-2 unless the People's Republic of China and the first-affected nations (South Korea, the Islamic Republic of Iran and Italy) managed to stop further transmission.[?] Previous outbreaks, such as the H1N1/09 influenza pandemic, have attested to the rapid potential in particular of viral pathogens to rapidly spread.[?,?,?] With growing networks of transportation and the increased volume and affordability in particular of air travel, airborne and droplet-transmitted pathogens now have an unprecedented global reach and velocity.[?,?,?]

Tentative data about the COVID-19 pandemic strongly suggests that the same global transportation networks have had a significant effect on the spread of SARS-CoV-2. Porcheddu et al. (2020) indicate that most of the early cases in Italy were travel-related,[?] as were most early cases of COVID-19 in the United States.[?] While the work by Anzai et al. (2020) and Chinazzi et al. (2020) cast some doubt on whether travel restrictions alone would have had a significant impact on pathogenic dynamics,[?,?] it is undeniable that the freedom of movement afforded by cheap air travel and more liberal immigration regimes throughout most of the planet are posing a novel challenge to global threat reduction.

As a general approach, global health security (GHS) and threat reduction has largely adopted what Davies (2010) referred to as the "statist perspective" of global health[?] – a kind of foreign internal defence against pathogenic threats, funded by affluent Western nations and delivered in developing nations that are often the sites of emergence of new viral pathogens. The COVID-19 pandemic is incontrovertible proof that this approach is no longer workable. As Morens et al. (2020) write,

> [w]e must realize that in our crowded world of 7.8 bil-

> lion people, a combination of altered human behaviors, environmental changes, and inadequate global public health mechanisms now easily turn obscure animal viruses into existential human threats.[?]

In lieu of the statist perspective, a global perspective (to use Davies's term, the "globalist approach") is required. A global health security agenda premised on this approach

> starts with individual health needs and then takes into account how global actors and structures impact on the individual, considering factors ranging from poverty and poor education to the actions of states and the health effects caused by international organizations, multinational corporations and others. The state remains a core actor in this perspective, but globalists see it as just one among a wide range of actors and situates the individual as the core referent.[?]

This paper outlines a technological approach to foster and support a globalist approach to GHS while also affording the benefits of "statist" approaches in terms of national threat reduction priorities.

Surveillance as a pillar of global health security

We draw analogies to an immensely successful international system of ensuring security, in the sphere of nuclear disarmament and arms control.[?] Signed on 10 September 1996, the Comprehensive Nuclear Test-Ban Treaty (CTBT) prohibits "any nuclear weapon test explosion or any other nuclear explosion". At the heart of the CTBT is an extensive network of over 170 seismic monitoring stations, hydroacoustic monitors, infrasound monitoring and radionucide detection equipment all across the signatory nations, bolstered by national data centres and the International Data Centre at the CTBTO Preparatory Commission's headquarters in Vienna, Austria.[?]

From the technical perspective, the CTBTO's capacities are uniquely powerful, albeit quite specifically premised on the ability to detect nuclear explosions due to the seismic impact and radionuclide emission. Pathogens, whether naturally occurring or released intentionally or negligently, do not have the physical impact that makes their detection using the CTBT surveillance regime feasible. However, by analogy to the global surveillance network set up by the CTBTO, a Pandemic Advance Warning System (PAWS) can be constructed by reliance on integrating local healthcare data. Leveraging the advances in healthcare technology and the growing ubiquity of electronic health and medical records worldwide, along with new advances in the field of anomaly detection and machine learning, a PAWS solution can be designed that may provide early warning about emerging transmissible diseases and facilitate local action before pandemic spread.

2. Design of a Pandemic Advance Warning System

2.1 Design principles

In general, a PAWS solution needs to comply with three principal requirements:

1. **Integrability**: by leveraging national and transnationally used standards, such as HL7, a PAWS solution needs to be able to integrate into a wide range of electronic health record (EMR/EHR) software and a large number of other information capture systems (such as field epidemiology data capture terminals). In addition, a PAWS solution must be able to integrate with ancillary systems, such as labs, state epidemiological authorities and other sources of data.

2. **Reliability**: just as the fundamental strength of the technical infrastructure behind CTBTO is its tamper-evident infrastructure (e.g. via seals on radionuclide detectors and regular state-of-health reporting through public key encrypted data streams), the fundamental value of a PAWS solution would be its ability to afford a reliable view of early prodromic signs to all participant states. This is best accomplished by a protocol that is self-monitoring, tamper-proof and integrated at the lowest level. Public key encryption suites can secure such information, while blockchain based approaches have been considered in the health sector to preserve the integrity of data.[?] In particular, blockchains offer immutability and auditability at scale, which can be used to ensure that information is not tampered with.

3. **Privacy**: in order to respect individual privacy vis-a-vis healthcare information as one of the most sensitive domains of personal information, and to ensure compliance with domestic laws, any PAWS solution needs to be privacy-preserving by design. This may be by the use of pseudonymous identifiers or even restrictions to aggregate data. Since the primary aim is not contact tracing or identification of individual patients but rather a risk early warning system, anonymised clinical data, possibly aggregated at high geographic and temporal granularity, is sufficient.

2.2 Data sources

PAWS solutions integrate with EHR/EMR, laboratory, sequencing and clinical analytics/population health software at the lowest possible level in order to facilitate 'fail-deadly' tamper-proof security, wherein tampering with the solution's connection to the international data centre, or a combination of randomly selected data centres or sentinel sites, would immediately disable the solution itself. This principle of Permissive Actioning (PA) acts as a safeguard against 'jamming' or other simple circumvention of the clinical data sources.

Primarily, PAWS is intended to connect to EMR/EHR solutions directly, providing symptom detection surveillance, bolstered by strong communication networks, mutual confirmation of computational results and anomaly detection algorithms deployed strategically at sites where computing power

is both available and affordable. Of course, this requires a thorough recording of symptoms and objective results (medical imaging, nuclear mecidine, medical laboratory sciences, microbiology). While EMR/EHRs are ubiquitous in the United States and widely used in most developed nations, they have seen comparatively less use throughout the developing world. However, over recent years, this tide seems to have turned. The availability of open-source EMR/EHR implementations like OpenMRS has played a significant role in enabling developing countries to have HL7 compatible, standards-compliant EMR/EHR solutions.[?] It is assumed that just as the experience from EMR/EHR programmes in Africa initially targeted at HIV/AIDS intervention shows,[?] widespread adoption of EMR/EHR solutions in more resource-constrained areas of the world is just a matter of time.

In addition, along with structured data from EMR/EHR implementations, recent years have witnessed the development of informal early warning and information exchange networks. An example of these is Epi-X, maintained by the United States' Center for Disease Control and Prevention (CDC). Since coming online in late 2000, Epi-X has received over 60,000 reports, and has greatly contributed to the monitoring of the 2006 New Jersey *Fusarium spp.* keratitis outbreak, the 2009 H1N1 pandemic, the West African EBOV outbreak of 2014 and the recent ZIKV epidemic in 2016.[?,?] Limited to public health officials designated and vetted in advance by the CDC, the system also heavily relies on the capabilities of the CDC to edit and manage submissions. While this is an intriguing implementation of an early warning system, it is not scalable on either end. The demands on the maintainer to manually vet all submissions does not scale to a global health agenda and can only support a relatively small recipient community. In particular, the semi-structured data submitted to Epi-X puts a significant load on recipients to sift and understand the data, while also limiting their ability to quantitatively analyse epidemiological processes reported therein.

The PAWS paradigm of pandemic early alerting is intended to integrate with these developments by providing a globally transparent, shared detection capability for anomalous signals. In other words, by surveillance of presenting symptoms, diagnoses and objective findings, a wealth of data is generated that can then be centrally evaluated for anomalies, including anomalous co-occurrences (suggesting a new, emerging syndromic entity) and anomalous spikes in known symptom constellations (suggesting a rise in a known clinical entity).

2.3 Infrastructure

One of the strengths of the CTBTO regime is a comprehensive communications and integration infrastructure, referred to as the Global Communication Infrastructure (GCI).[?] This connects the 337 monitoring stations (at the time of writing) – comprising primary and secondary seismic detection stations, radionuclide measurement stations, hydrophone arrays and infrasound measurement stations – via a combination of terrestrial communications secured by virtual private networking (VPN) and satellite-based communications using a constellation of six communications satellites through Very Small Aperture Terminals (VSAT).

It is estimated that daily traffic over the GCI, including Quality of Service (QoS) and tamper-security data (referred to as State of Health, or SOH for short, in the CTBTO context), approximates 35GB/day. It is our estimation that a globally integrated infectious disease monitoring system would have to contend with at least 4-5 orders of magnitude more data – rising to potentially 8-9 orders of magnitude if unstructured data, genomic sequencing and clinical imaging are integrated. To respond to this need, the underlying infrastructure should be channel-neutral, i.e. operate as an OSI Layer 4 transport protocol that can leverage a wide range of Layer 3 transport options and a wide range of physical communication infrastructures, from radiofrequency and satellite communications through wired and fiberoptic links. Global disparities in the cost and availability of these connections mean that the resulting architecture must be able to accommodate a vast range of different interconnectivity levels, including batch/bulk uploads of data in a limited time window (e.g. due to intermittent satellite coverage) while maintaining strong security and data integrity.

2.4 Analysis

The rise of algorithmic anomaly detection models in time series, including Long Short-Term Memory (LSTM) based anomaly detection[?,?] and Hidden Markov Models (HMM),[?] attests to a growing ability to isolate anomalies and identify trends even from large and 'noisy' data sets. These techniques have already found use in a number of fields, such as intrusion detection in cybersecurity,[?] identification of price manipulation to support antitrust law[?] and aviation safety.[?] It can be said with confidence that the arsenal of mathematical and quantitative tools at our disposal are sufficient for identifying early prodromic signs of an emerging pandemic. At the same time, massively parallel computation has made many of these techniques affordable and thanks to vendors like Amazon Web Services and Microsoft's Azure offerings, there is an unprecedented availability of highly scalable burst computational capacity at our disposal. In addition, many cloud providers are providing service offerings that come with strong security features and are pre-certified for government clients, such as Amazon Web Services's GovCloud.

The CTBTO's analytical burden is relatively low, as nuclear and thermonuclear explosions present a unique, localisable and well-identifiable seismic, infrasound and radiation signature. In other words, detecting a nuclear or thermonuclear explosion relies on a relatively high signal-to-noise ratio. Even compared to normal seismic activity, nuclear and thermonuclear explosions can be differentiated from normal seismic activity through the presence of radionuclide emissions and anomalous propagation patterns. This is not the case for early outbreaks.

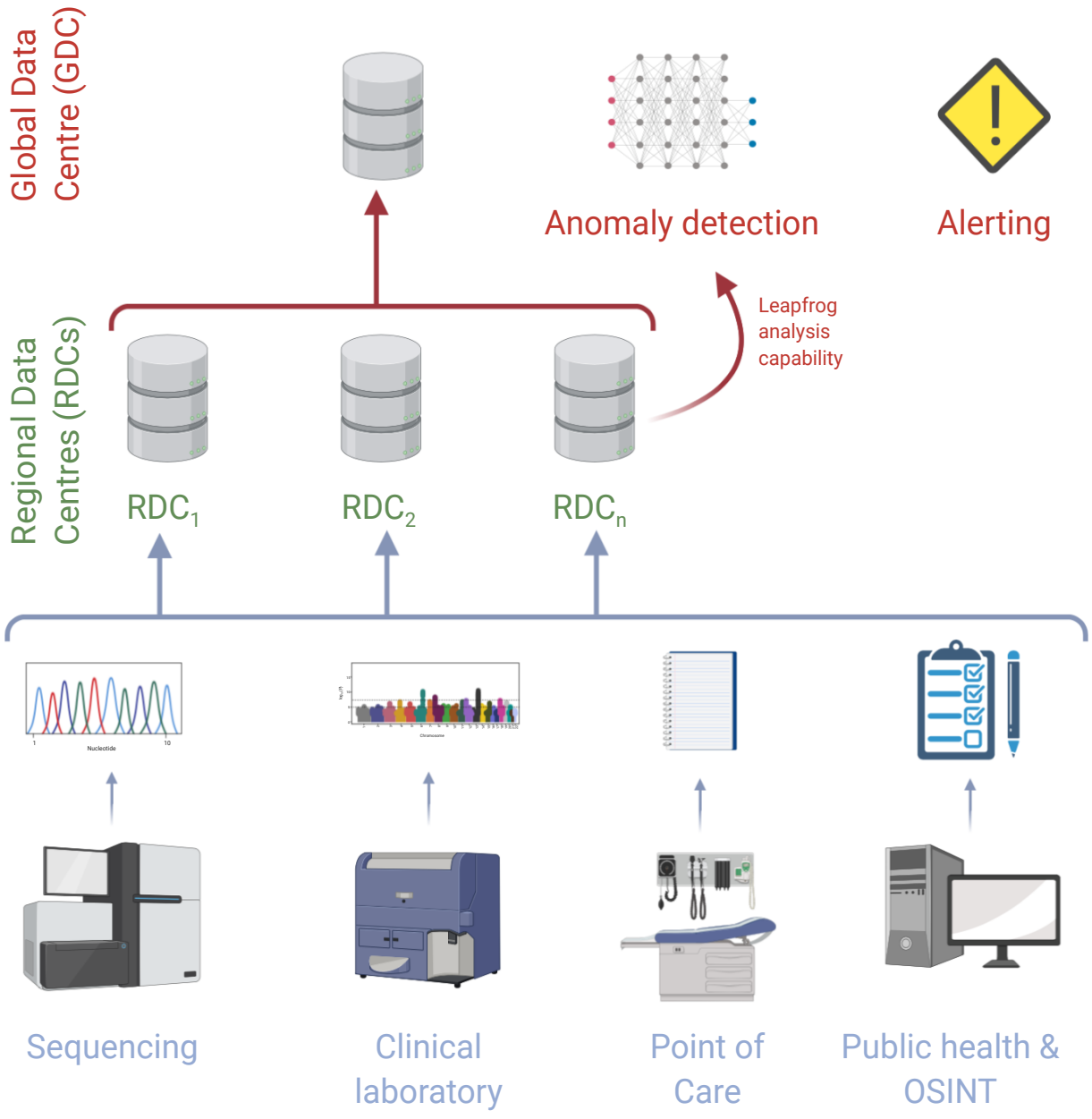

Figure 1. Information flow in a possible PAWS implementation.

With the exception of a few pathological entities, notably viral haemorrhagic fevers (VHFs) such as ebolaviruses, marburgviruses and Crimean-Congo Haemorrhagic Fever, most pandemic threats present with generalised symptoms, such as malaise, pyrexia (fever), asthenia (weakness) and elevated non-specific markers of early immune activity (elevated C-reactive protein, white blood cell count and erythrocyte sedimentation rate). In particular, if these cases occur during a time when presentations characterised by a generalised viral syndrome are not immediately anomalous, such as during influenza season, they may be misidentified as a prevailing known pathogen or 'hide in the noise' of a similar but distinct underlying disease process. Indeed, this has to an extent been the case during the early days of the COVID-19 pandemic.

A PAWS solution must therefore be supported by robust analytical capabilities, both in terms of design and implementation. In terms of design, it is crucial that a PAWS solution be supported by stringent data quality filters. The evolution of anomaly detection methodologies has created an impressive array of technologies that leverage machine learning for the identification of data quality issues in fields as divergent as magnetic resonance spectroscopy,[?] building management systems[?] and citizen science in ornithology.[?] Such measures can go a long way to identify and remedy anomalies that can be accounted for by data quality issues. In terms of implementation, a PAWS solution must be backed by both the expertise and the resources to deploy cutting-edge data science and machine learning capabilities for anomaly detection over the space of tens to hundreds of terabytes of data per day, rapidly detecting anomalies and quantifying emergent trends. The CTBTO's approach relies on a centralised International Data Centre, based in Vienna, Austria. Perhaps a more appropriate approach for a global PAWS solution would be to leverage the already existing interconnections between cloud/PaaS providers throughout the globe and rely on a distributed, fault-tolerant, self-healing infrastructure using redundancy and Byzantine fault tolerance to prevent tampering with results at any level.

3. Conclusion

The rapid emergence and worldwide impact of SARS-CoV-2, the virus responsible for COVID-19, has highlighted not only the global impact that pandemics can have, but also the unprecedented speed at which such effects may develop. In particular given the grave second order social and economic effects of this outbreak, it is hard to deny that emerging novel infectious diseases now form as much a key concern of homeland security and the national defence as terrorism or hostile cyberthreats.

Yet while response to both of those threats has been extensive, well-funded and data-driven, the opportunities in creating a globally integrated pandemic advance warning system have not been exploited successfully to this date. As the final cost – human, economic and sociopolitical alike – of the coronavirus is tallied up, the price of devising and maintaining such a system will hardly be more than a rounding error, but the cost of its absence will be seen to have left lasting damage.

Any such system must reconcile the two perspectives on global health security that Davies (op. cit.) proposed. It must afford states the protection of their national security interests that the 'statist' approach proposes, but it must support global integration that is more in line with 'globalist' thinking on GHS. Neither approach alone has provided the means to successfully detect and curb the initial outbreak before it has gained pandemic spread. A synthesis or reconciliation of these approaches is needed to prevent the next global pandemic – possibly much more costly in all terms than the present one. The collaboration on developing a PAWS solution that equitably caters to the interests of the entire global population might well help usher in that synthesis.

Acknowledgments

The authors wish to thank Prof. Benedicte Callan and Ms. Janet Kathryn Hedrick for the discussions that inspired this paper. All errors and ommissions are the authors' own.

Competing interests

The authors have declared no competing interest.

Funding statement

No external funding received.