RESEARCH ARTICLE

SARS-CoV-2 genome datasets analytics for informed infectious disease surveillance [version 1; peer review: 1 approved with reservations]

Moses Effiong Ekpenyong ^{1,2}, Ifio James Udo¹, Mercy Ernest Edoho¹, EnoAbasi Deborah Anwana³, Francis Bukie Osang⁴, Joseph Ikim Geoffery¹, Emmanuel Ambrose Dan¹, Aliu Bolanle Momodu¹, Nnamso Michael Umoh¹, Kingsley Christopher Udonyah⁵

¹Department of Computer Science, University of Uyo, Uyo, Akwa Ibom, 520003, Nigeria

²Centre for Research and Development, University of Uyo, Uyo, Akwa Ibom, 520003, Nigeria

³Department of Botany and Ecological Studies, University of Uyo, Uyo, Akwa Ibom, 520003, Nigeria

⁴Department of Computer Science, National Open University, Abuja, 900271, Nigeria

⁵Department of Computer Science, Akwa Ibom State University, Mkpato Enin, Akwa Ibom, 524106, Nigeria

V1 First published: 14 Sep 2021, 10:919
<https://doi.org/10.12688/f1000research.55007.1>

Latest published: 14 Sep 2021, 10:919
<https://doi.org/10.12688/f1000research.55007.1>

Abstract

Background: The COVID-19 pandemic has ravaged economies, health systems, and lives globally. Concerns surrounding near total economic collapse, loss of livelihood and emotional complications ensuing from lockdowns and commercial inactivity, resulted in governments loosening economic restrictions. These concerns were further exacerbated by the absence of vaccines and drugs to combat the disease, with the fear that the next wave of the pandemic would be more fatal. Consequently, integrating disease surveillance mechanism into public healthcare systems is gaining traction, to reduce the spread of community and cross-border infections and offer informed medical decisions.

Methods: Publicly available datasets of coronavirus cases around the globe deposited between December, 2019 and March 15, 2021 were retrieved from GISAID EpiFluTM and processed. Also retrieved from GISAID were data on the different SARS-CoV-2 variant types since inception of the pandemic.

Results: Epidemiological analysis offered interesting statistics for understanding the demography of SARS-CoV-2 and helped the elucidation of local and foreign transmission through a history of contact travels. Results of genome pattern visualization and cognitive knowledge mining revealed the emergence of high intra-country viral sub-strains with localized transmission routes traceable to immediate countries, for enhanced contact tracing protocol. Variant surveillance analysis indicates increased need for continuous monitoring of SARS-

Open Peer Review

Approval Status ?

1

version 1

14 Sep 2021

?

view

1. **Mario Coccia** , CNR -National Research Council of Italy, Torino, Italy

Any reports and responses or comments on the article can be found at the end of the article.

CoV-2 variants. A collaborative Internet of Health Things (IoHT) framework was finally proposed to impact the public health system, for robust and intelligent support for modelling, characterizing, diagnosing and real-time contact tracing of infectious diseases.

Conclusions: Localizing healthcare disease surveillance is crucial in emerging disease situations and will support real-time/updated disease case definitions for suspected and probable cases. The IoHT framework proposed in this paper will assist early syndromic assessments of emerging infectious diseases and support healthcare/medical countermeasures as well as useful strategies for making informed policy decisions to drive a cost effective, smart healthcare system.

Keywords

Disease surveillance, infectious disease, genome pattern mining, SARS-CoV-2, self-organizing map

This article is included in the **Genomics and Genetics** gateway.

This article is included in the **Pathogens** gateway.

This article is included in the **Coronavirus** collection.

Corresponding author: Moses Effiong Ekpennyong (mosesekpenyong@uniuyo.edu.ng)

Author roles: **Ekpenyong ME:** Conceptualization, Investigation, Methodology, Writing – Original Draft Preparation; **Udo IJ:** Data Curation, Investigation, Visualization, Writing – Review & Editing; **Edoho ME:** Data Curation, Validation, Writing – Original Draft Preparation; **Anwana ED:** Data Curation, Investigation, Writing – Review & Editing; **Osang FB:** Investigation, Writing – Review & Editing; **Geoffery JI:** Data Curation, Software; **Dan EA:** Data Curation, Visualization; **Momodu AB:** Data Curation, Writing – Review & Editing; **Umoh NM:** Data Curation, Investigation, Writing – Review & Editing; **Udonyah KC:** Data Curation, Visualization

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Ekpennyong ME *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ekpennyong ME, Udo IJ, Edoho ME *et al.* **SARS-CoV-2 genome datasets analytics for informed infectious disease surveillance [version 1; peer review: 1 approved with reservations]** F1000Research 2021, 10:919 <https://doi.org/10.12688/f1000research.55007.1>

First published: 14 Sep 2021, 10:919 <https://doi.org/10.12688/f1000research.55007.1>

Introduction

Coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pathogen, was initially detected in Wuhan, China in December 2019 and has progressively become a global pandemic with continuous negative impacts on health policies, education, social relationships, and national economies. The rapid rate of transmission with no medically proven prophylaxis and high numbers of recorded deaths has kept the scientific community apprehensive, working towards understanding the virus and possibly finding avenues to contain the virus. Epidemiological studies have recognized the variations in demographic, clinical and genomic features of COVID-19 among and within continents, countries, and regions. This variation has led to a wide range of outcomes where a proportion of people with positive reverse transcription polymerase chain reaction (RT-PCR) tests have been found to be asymptomatic and symptomatic patients have also reportedly exhibited different symptoms ranging from a mild cold to intense illness and in some cases, death. Hence, understanding the underlying mechanism of variation appears crucial.

A myriad of studies has characterized epidemiologic and genomic features in hospitalized patients in specific regions while some inter-regional/continental studies have utilized COVID-19/SARS-COV-2, curated, and published primary and secondary data (in open access databases). In the literature, the Global Initiative on Sharing All Influenza Data (GISAID) (Elbe and Buckland-Merrett, 2017) has been the most frequently accessed for genomic data mining. The criticism cast on the World Health Organization's (WHO's) Global Influenza Surveillance Network, on the limited global access to avian H5N1 influenza virus sequence data managed by the Los Alamos National Laboratories in the United States of America brought into focus the need for optimal transparency in data sharing with no infringement of intellectual property rights. In addition to GISAID, 34 of such databases and computational resources together with their resource description and data types have been outlined by the Office of Data Science Strategy, National Institute of Health. The data, which predominantly includes genomic, chemical structure, epidemiologic data and digital images, have been made available for sharing, verification, and collaborative research through the efforts of medical personnel and researchers. Besides medical scientists, digital technologists have also strived to play a role in tackling the pandemic, in areas such as risk assessment and patient prioritization, screening and diagnosis, contact tracing, supporting drug discovery and treatment. This technological effort stems from the artificial intelligence (AI) community and thrives on data from open-source repositories as its essential element. However, the inherent constrictions of these repositories possess some systematic challenges, as possible limitations in sample sizes of biomedical data, incompatible data extensions and schemas as well as non-uniformity of data elements of similar features across repositories impede the integration of data from different sources.

Providing capacity to contain emerging and recurrent outbreaks of infectious diseases is insufficient to satisfy effective medical countermeasures such as diagnostics, therapeutics, and vaccines. Often, non-pharmaceutical interventions such as contact tracing, outbreak investigation, isolation, social distancing, and the use of face masks remain the only viable options/tools for slowing down an emerging outbreak in the absence of such medical countermeasures. Once a novel infectious threat is identified, surveillance of the disease transmission becomes paramount for immediate risk assessment. The COVID-19 pandemic has presented the urgent need for governments, healthcare decision makers, providers, and others, to refocus on balancing their response with wider health needs. The increased availability of novel web-based data sources can substantially improve infectious disease surveillance (Choi *et al.*, 2016; Ainsworth *et al.*, 2021). Globally, web-based surveillance tools and epidemic intelligence methods are providing new prospects to facilitate risk assessment and timely outbreak detection. These tools are increasingly adopted for rapid detection of changes in the incidence rate of endemic diseases and the early detection and characterization of syndromes caused by previously unknown pathogens of epidemic potential. Furthermore, customized systems utilizing robotics (Yang *et al.*, 2020) are appearing in infectious disease situations to manage disease situations in the areas of clinical care (telemedicine and decontamination), logistics (delivery and handling of contaminated waste), and reconnaissance (monitoring compliance with voluntary quarantines); and are actively being deployed in Asian countries such as China and Israel, as well as in Western countries such as the United States.

While the 'Big Data' revolution is ongoing, most public health systems such as those in low- and medium-income countries (LMICs) still rely on traditional surveillance. Simonsen *et al.* (2016) advocate increased use of hybrid systems that combine traditional surveillance information and Big Data sources, to ensure a faster system that is more relevant to local needs. Today AI and data science are redefining our response to infectious disease situations, by powering existing devices to deliver cost effective services through predictive modelling (a data mining and probability method for estimating more granular, specific outcomes) and effectively complementing forecasting methods (trend analysis for estimating future events based on past and present data) for proper service integration into decision-making processes of surveillance systems. Consequently, decision-making under emerging infectious disease situations has greatly improved using appropriate data and advanced analytics (George *et al.*, 2019) for reinforced public health actions such as resource requirement determination, situational awareness refinement, and control efforts monitoring (Chretien, Riley, and George, 2015; Rainisch *et al.*, 2015; CDC FluSight; Meltzer *et al.*, 2016).

One major challenge facing the response, recovery, and resilience to emerging pandemics is the weak system for collecting timely, disaggregated data related to the pandemic. Underfunding, mixed data standards, data integration, and the need to share quality data at local, national, regional, and international levels, are challenging epidemiological protocols, hence, undermining health system capacity, and decreasing support for open access tools. Nevertheless, with the availability of other data sources such as social media data, mobile phone data, satellite data and citizen-generated data, there is hope that these gaps can be addressed in real-time. We summarize in [Table 1](#) the effective technologies and resources used at national and international levels for predicting and controlling emerging infectious disease threats. Examples of each resource are highlighted as well as their main applications.

Increased uncertainty of SARS-CoV-2 mutation and the ability of the virus to adapt to changing environments has confirmed the emergence of novel sub-types and strains ([Wang et al., 2020, 2021](#); [Grabowski, Kochanczyk, M., and Lipniacki, 2021](#), [Richmond et al., 2020](#); [Koyama et al., 2020](#)). Furthermore, inadequate capacity of under-resourced countries to contain the sudden spread of emerging infectious diseases has diminished sound medical protocol and is gradually breeding a ‘careless’ society that resorts to self-help activities such as patronizing quacks.

This research study seeks a patient-centered healthcare system with smart components for driving robust decisions on emerging infectious disease surveillance. To initiate this investigation, the following hypotheses are proposed:

H1: Localizing transmission routes of intra-country sub-strains to immediate countries and continents would most likely assist early contact tracing and stem the spread of infectious diseases.

H2: Implementing a smart healthcare system powered by Internet of Things technology would support disease surveillance in overburdened and poor health systems.

H3: Implementing a healthcare system powered by Internet of Things technology offers patient-centered services and enhances healthcare policy decisions.

This paper tackles the first hypothesis by mining retrieved genome datasets from GISAID, to provide a demography and transmission routes of SARS-CoV-2 viral sub-strains among infected patients, by gender. An Internet of Things (IoT) framework with collaborative components for driving robust decisions on emerging infectious disease surveillance including a feasible workflow for implementing same is proposed to initiate the second and third hypotheses. To actualize the later hypotheses, a research project grant is currently being pursued. The research is expected to impact the research community and society in the following areas:

Table 1. Technologies and resources for infectious disease surveillance and typical application area (Adapted from [Christaki, 2015](#)).

Technologies/ resources	Example	Application
Event-based surveillance	GPHIN, ProMED-mail, HealthMap, EpiSPIDER, BioCaster	Outbreak and emerging public health threat detection of SARS
Web-based real-time surveillance	Google Trends, Google Flu Trends, Johns Hopkins University Interactive Web-based Dashboard (Dong et al., 2020)	Real-time monitoring of disease activity (seasonal influenza activity, COVID-19)
Early warning systems and alert response networks	GOARN	Detection of public health threats, Communication between institutions, Implementation of preventive and control measures (WHO Global Alert and Response).
Infectious disease modelling	Agent-based models, metapopulation models (GLEAM, FRED, gravity model)	Epidemic simulation, Assessment of disease spread determinants, Design of containment interventions.
Social media	Flu Near You, Outbreaks Near Me, COVID Near You	Participatory epidemiology (seasonal influenza activity, COVID-19 activity).
New technologies in pathogen discovery	Genome wide sequencing, microarrays, bioinformatics	Pathogen/virus discovery, predictive modelling, determinants of host susceptibility.

- *Localized transmission route discovery* – Whereas Ekpenyong *et al.* (2021a) adopted a global transmission route classification, with a greater proportion of the genome sequences belonging to the intra-country sub-strains cluster, this paper localizes the transmission routes of intra-country sub-strains to immediate countries and continents for enhanced contact tracing and supports the deployment of early countermeasures to prevent further transmissions. Furthermore, Governments' understanding of the source of the disease or infection will guide enhanced border regulations and ascertain which sub-strain(s) is (are) spreading within their countries.
- *Disease information sharing and AI-knowledge extraction* – Retrieval of datasets has always been met with complexity imposed by unstructured data, hence, increasing the difficulty of restructuring the database for AI-knowledge extraction. A novel taxonomy re-defining the unique annotation of entries into clinical information databases is demonstrated in this paper, to aid knowledge simplification and minimization of inconsistencies.
- *Processed global disease datasets* – The datasets provided in this paper are useful for accurate characterization and prediction of SARS-CoV-2 genome pattern(s)/sub-strain(s) by gender, a contribution that is currently missing in the literature. This data would benefit computational scientists in the development of classification models as well as expert/recommendation systems for global disease surveillance. Clinicians/physicians and pharmacists can also exploit the proposed expert system framework, to support efficient decisions on contact tracing, disease diagnosis, and recommendations. Furthermore, by providing access to processed clinical (control) SARS-CoV-2 data, research could be advanced towards individualized/precision medicine. Finally, the developed models and algorithms would provide open-source tools with domain adaption and research replication possibilities.
- *Optimized response and recovery system for patient diagnosis, care, and management* – The proposed Internet of Health Things (IoHT) framework will support overburdened and poor health systems with insufficient numbers of health workers, when responsibly deployed. It will provide better logistics for resource allocation, support self-diagnosis, psychosocial care, and enhance the timely communication of protocols such as real-time medical response and quality policy decisions, contact tracing, and early information on treatment tips/recommendations.

Methods

Data description and patient status - specimen sources taxonomy

A total of 29225 complete, high coverage genome sequences (sequences with lengths of above 29000 bp and <1% undefined or ambiguous bases, 'N's, and <0.05% unique amino acid mutations), were retrieved as a single FASTA file from GISAID EpiFluTM between December 2019 and March 15, 2021. Statistics of retrieved genomes are distributed by continent as follows; Africa: 2288, Europe: 8592, Asia: 13210, North America: 1829, South America: 3289, and Oceania: 17. Only sequences with patient status (age and gender) and complete collection date were filtered and retained for the study, resulting in 9164 genome sequences consisting of 5269 male and 3895 female samples from 61 different countries of the world, across six continents, Antarctica exempt (as no SARS-CoV-2 data deposits were found from Antarctica at the time of retrieval).

Python 3.9.5 was used to extract individual genome sequences from the retrieved FASTA file. The following functions/libraries/packages were used to aid extraction process: Pandas (a library for data manipulation and analysis, used to convert the extracted sequences to CSV file); PyCountry (a library containing ISO information about countries of the world, used for converting country names and their long-forms into their ISO-alpha-3 form); re (the Python regular expression library, used to match characters and sequence of characters during the extraction and cleaning process); os module (a module with several functions for interacting with host's operating system, used for creating directories, moving and deleting files during the cleaning process). The Python code used for the extraction is provided in Algorithm 1 (*Extended data*).

Metadata was compiled to document specific isolate details including: Isolate Code (three-letter country code_isolate number), Country, Accession Number, Gender, Age (Child: <18 y, Adult: 18-59 y, Senior: >59 y), Status, Specimen Source and Additional Information. The Additional Information column holds both location and host information such as transmission history, treatment history, date sample was taken, etc. Specimen sources include swabs (nasal, oral, throat, nasal and oral), fluids (bronchoalveolar lavage, saliva, sputum, stool) and unknown. FASTA files of the genome isolates can be located at GISAID using the accession numbers (see *Data availability*). The statistical methods used to analyze the metadata were summation, simple measures of central tendencies (mean and mode), as well as additional logical statements for conditional count, sum, and approximate and exact matching. These were achieved using the following Microsoft Excel functions: SUM, AVERAGE, MODE, COUNTIF, SUMIF, and VLOOKUP.

Using Python, raw genome sequences of male and female patients/isolates were retrieved and processed into vertical columns of individual genomes and stored in a CSV file for the various continents under study, in the following order: Africa, Europe, Asia, North America, South America and Oceania. During the retrieval, we observed that the GISAID database was inconsistent at rendering the patient status and specimen sources, and numerous incoherent annotations introduced inherent redundancy. To assist efficient documentation and processing of data for intelligent analytics, taxonomies re-classifying these two fields are given in Figure 1 and Figure 2, respectively. A semi-automated method was used to develop the taxonomies. To obtain patient status, the mode function of Microsoft Excel was used to obtain the most frequent occurring status (unique status) and such similar comparisons manually verified before the search and replace function was used to replace redundant statuses with unique ones. Hence, our taxonomies subsume incoherent or redundant annotations (annotations in square text boxes) into unique specifications (annotations in oval shapes), ready for efficient data mining (Edoho *et al.*, 2020). Hence, for patient status (Figure 1), the unique annotation sequence used to relabel the datasets includes:

Figure 1. GISAID COVID-19 patient status taxonomy. GISAID, Global Initiative on Sharing All Influenza Data; COVID-19, coronavirus disease 2019; ICD, International Classification of Diseases; EHPAD, Établissement d'hébergement pour personnes âgées dépendantes.

Figure 2. GISAID COVID-19 specimen source taxonomy. GISAID, Global Initiative on Sharing All Influenza Data; COVID-19, coronavirus disease 2019; VTM, viral transport medium.

- Symptomatic -> Hospitalized -> [Live, Mild, Moderate, Severe, Critical, Recovering, Recovered, Released, Deceased].
- Symptomatic -> Not Hospitalized (Outpatient) -> [Home]
- Symptomatic -> [Isolation]
- Symptomatic -> [Quarantine]
- Asymptomatic
- Unknown

For specimen sources (Figure 2), the unique annotations used to relabel the datasets include:

- Pharyngeal Swab
- Pharyngeal Swab, Serum
- Pharyngeal Swab, Serum, Urine
- Saliva
- Saliva, Pharyngeal Swab
- Serum
- Serum, Urine
- Urine
- Sputum
- Autopsy Material
- Brain Tissue
- Cerebrospinal Fluid
- Anal Fluid
- Lung Tissue
- Unknown

SARS-COV-2 variants processing

In June 2021, WHO designated seven variants of interest (VOIs) and four variants of concern (VOCs) (Konings *et al.*, 2021). Submissions and statistics documenting SARS-CoV-2 VOIs/VOCs were retrieved as a Microsoft Excel workbook using the GISAID EpiCov™ download option. The retrieved Workbook documents VOI/VOC cases between December 16, 2019 and August 8, 2021. Using Microsoft Excel SUM and VLOOKUP functions the following number of SARS-CoV-2 VOIs/VOCs cases were retrieved and processed; VOI Lambda: 647826, VOI Kappa: 981843, VOI Iota: 974080, VOI Eta: 1472927, VOI Zeta: 19012, VOI Gamma: 1036999, VOC Beta: 659580, VOC Alpha: 2200327, and VOC Delta: 89961.

Pattern visualization model

Our pattern visualization model is defined by a self-organizing map (SOM) – a single neural network with neurons defined along the grids, that projects data into a low-dimensional space (Kangas *et al.*, 1990). Using an unsupervised, competitive learning process, a low-dimensional, discretized representation of the input space or training samples, known as the feature map, is produced. During training, weights of the winning neuron and neurons in a predefined neighborhood are adjusted towards the input vector using equation (1),

$$w_{id}^{t+1} = w_{id}^t + rf(i, q)(x_d - w_{id}^t); 1 \leq d \leq D. \quad (1)$$

where r is the learning rate and $f(i, q)$ is the neighborhood function, with value 1 at the winning neuron q and decreasing as the distance between i and q increases. At the end, the principal features of the input data are retained. The batch unsupervised weight/bias algorithm of MATLAB 2017b (*trainbu*) with mean squared error (MSE) performance evaluation, was adopted to drive the proposed SOM. This algorithm trains a network with weight and bias learning rules using batch updates. The training was carried out in two phases: a rough training with large (initial) neighborhood radius and large (initial) learning rate, followed by a finetuned training phase with smaller radius and learning rate. A freely available alternative software that can be used to replicate this study is Python with the following data science

libraries: NumPy—a fundamental package for scientific computing in Python. Pandas. Matplotlib.pyplot—a collection of functions that make matplotlib work like MATLAB.

Each genome sequence was mapped or transformed into an equivalent genomic signal (a discrete numeric sequence) using the following encoding of the individual nucleotide (i.e., A = 1; C = 2; G = 3; T = 4). As base input, we maintained nucleotide pairs above 29000 bp (the input vector), indicating approximate (maximum) length of DNA sequences of the SARS-CoV-2 genome. Next, all ambiguous sequences were removed. A vector representation for pairwise Euclidean distance computation among the vectors in the form of a distance matrix was achieved using the SOM algorithm implemented in MATLAB 2017b. As the distance matrix is highly dimensional, a suitable representative sequence of each isolate was adopted and the individual component planes transformed into a cognitive map using their similarity scores, useful for labelling classification targets for predictive systems.

Cognitive knowledge mining

Knowledge mining is an emerging field in AI and has had huge benefits for quick learning from Big Data. We applied natural language processing to the genome datasets and extracted knowledge of similar viral sub-strain(s). An iterative technique (using Python Similarity, and Microsoft Excel's COUNTIF and CORRELATION commands/functions) was then imposed on the SOM isolates ($i = 1, 2, 3, \dots, n$), where n is the maximum number of isolates. For each isolate pattern, similar patterns with the rest of the isolates (i.e., $i + 1, i + 2, \dots, n$) were compiled. Compiled isolate(s) were concatenated into a list ($j_1, j_2, \dots, j_m$) where j is an element of the list. The compiled list was dumped into *CogMap*($k_i \in j_1, j_2, \dots, j_m$).

Results

Epidemiological analysis

Epidemiological analysis was carried out on the Metadata file. Table 2 documents the continent, isolate distribution by country, isolate distribution by gender, and total number of samples/isolates retrieved. Using Python, the following statistics were compiled:

Symptomatic and asymptomatic cases

Table S1 (*Extended data*) shows statistics for symptomatic and asymptomatic cases. We observed more hospitalized cases (7625/9164, 83.21%) than not-hospitalized cases (391/9164, 4.27%), with more male patients hospitalized (M = 4338/7625, 56.89%; F = 3287/7625, 43.11%). Furthermore, more males died of COVID-19 than females (M = 541/9164, 5.90%; F = 248/9164, 2.71%). Asymptomatic cases represent 0.76% (40/5269) and 1.05% (41/3895) of total male and female patients, respectively.

Table 2. Distribution of retrieved data.

Continent	Country	Male	Female	Total
Africa	Algeria (3), Cameroon (1), DRC (8), Egypt (35), Gambia (13), Ghana (15), Madagascar (3), Morocco (6), Mozambique (7), Nigeria (18), Rwanda (27), Senegal (135), South Africa (1507), Tunisia (26).	701	1103	1804
Europe	Andorra (1), Austria (18), Belgium (11), Bosnia and Herzegovina (4), Bulgaria (1), Croatia (15), Cyprus (8), Czech Republic (173), Denmark (3), Faroe Islands (14), Finland (2), France (131), Georgia (4), Germany (12), Greece (30), Hungary (80), Italy (561), Moldova (3), Norway (1), Poland (7), Portugal (2), Romania (52), Russia (125), Slovakia (4), Spain (256), Sweden (3), Switzerland (2), Ukraine (13).	802	743	1545
Asia	Bahrain (1), Bangladesh (29), Cambodia (1), China (319), India (1598), Indonesia (91), Iran (11), Iraq (2), Israel (38), Japan (300), Kazakhstan (24), Kuwait (3), Lebanon (18), Malaysia (89), Mongolia (6), Myanmar (1), Nepal (1), Oman (58), Pakistan (4), Philippines (12), Saudi Arabia (500), Singapore (540), South Korea (18), Sri Lanka (29), Taiwan (64), Thailand (2), Turkey (134), United Arab Emirates (111), Vietnam (74).	2757	1321	4078
South America	Argentina (2), Brazil (519), Chile (1), Colombia (186), Ecuador (28), Peru (2), Venezuela (3).	394	347	741
North America	Canada (27), Costa Rica (58), Dominican Republic (6), Guadeloupe (17), Mexico (110), Panama (253), Saint Martin (8), USA (499).	603	375	978
Oceania	Guam (2), New Zealand (2), Australia (14).	12	6	18
	Total: Number of countries excavated per continent: Africa (14), Asia (29), South America (7), North America (8), Oceania (3).	5269	3895	9164

Patient age and status across continents

Distribution of patients by age and status across African countries is shown in Table S2 (*Extended data*). It was observed that South Africa had the highest number of entries (M = 503/1804, 27.88%; F = 1004/1804, 55.65%). Regarding age, the highest number of patients came from the Adult class (M = 506/1804, 28.05%; F = 869/1804, 48.17%). Regarding patient status, the highest proportion of patients belonged to the Live category (M = 599/1084, 33.20%; F = 1039/1840, 56.47%), followed by the Released category (M = 97/1804, 5.38%; F = 63/1804, 3.50%).

Distribution of patients by age and status across European countries is shown in Table S3 (*Extended data*). It was observed that Italy had the highest number of entries (M = 308/1545, 19.94%; F = 253/1545, 16.38%). Regarding age, adults (M = 395/1545, 25.57%; F = 392/1545, 25.37%) and seniors (M = 381/1545, 24.66%; F = 327/1545, 21.17%) shared the highest proportion (about 24%) of the total patients. Regarding patient status, the highest proportion of patients belonged to the Live category (M = 441/1545, 28.54%; F = 436/1545, 28.22%) followed by the Mild category (M = 122/1545, 7.90%; F = 109/1545, 7.06%).

Distribution of patients by age and status across Asian countries is shown in Table S4 (*Extended data*). It was observed that India had the highest number of entries (M = 1041/4078, 25.53%; F = 557/4078, 13.66%). Regarding age, adults (M = 2169/4078, 53.19%; F = 914/4078, 22.41%) constituted the highest proportion of patients. Regarding patient status, the highest proportion of patients belonged to the Live category (M = 1744/4078, 42.77%; F = 740/4078, 18.15%), followed by the Released category (M = 636/4078, 15.60%; F = 340/4078, 8.34%).

Distribution of patients by age and status across North American countries is shown in Table S5 (*Extended data*). It was observed that USA had the highest number of entries (M = 318/978, 32.52%; F = 181/978, 18.51%). Regarding age, adults (M = 392/978, 40.08%; F = 222/978, 22.70%) constituted the highest proportion of patients. Regarding patient status, the highest proportion of patients belonged to the Live category (M = 165/978, 16.87%; F = 123/978, 12.58%), followed by the Home category (M = 159/978, 16.26%; F = 120/978, 12.27%).

Distribution of patients by age and status across South American countries is shown in Table S6 (*Extended data*). It was observed that Brazil contributed the highest number of entries (M = 258/741, 34.82%; F = 261/741, 35.22%). Regarding age, adults (M = 232/741, 31.31%; F = 219/741, 29.55%) had the highest proportion of patients. Regarding patient status, the highest proportion of patients belonged to the Live category (M = 100/741, 13.50%; F = 109/741, 14.71%), followed by the Deceased category (M = 147/741, 19.84%; F = 120/741, 16.19%).

Distribution of patients by age and status across Oceanian countries is shown in Table S7 (*Extended data*). Although there is paucity of data in this continent, we observed that Australia contributed the highest number of entries (M = 9/18, 50%; F = 6/18, 33.33%). Regarding age, adults (M = 9/18, 50%; F = 3/18, 16.67%) had the highest proportion of patients. Regarding patient status, the highest proportion of patients belonged to the Recovering category (M = 6/18, 33.33%; F = 3/18, 16.67%), followed by the Recovered category (M = 3/18, 16.67%; F = 1/18, 5.56%).

Specimen sources across continents

Table S8 (*Extended data*) reveals that in the highest proportion of cases, pharyngeal swabs (M = 2925/9164, 31.92%; F = 2850/9164, 31.10%) were used as sequence samples. However, annotation evidence indicates the use of saliva (3/9164, 0.03%) in Europe (Spain), Asia (Turkey), and North America (Canada). The use of sputum (98/9164, 1.07%) as a sequence sample was found in Africa (Ghana and Nigeria), Asia (China, Indonesia, Japan, Lebanon, Mongolia, South Korea, Sri Lanka, and Thailand), and Oceania (New Zealand and Australia). Lung tissue (69/9164, 0.75%) was used as sequence sample in Asia (China and Japan), Europe (Austria, Belgium, Czech Republic, France, Italy, Russia, and Spain), and South America (Colombia and Ecuador). Anal fluid (14/9164, 0.15%) was used as sequence samples in Asia (China) and North America (USA). Serum or blood, and urine (4/9164, 0.15%) were used as sequence samples in Asia (India). A hybrid of samples (32/9164, 0.35%) was collected for sequencing SARS-CoV-2 virus in Africa (Nigeria), and Asia (India). Other forms of sequence samples (9/9164, 0.10%) came from Europe (Russia). Statistics also show that unknown sequence samples formed 31.21% (2860/9164) of the total sequence samples collected.

Intra- and inter-country transmissions across countries

Analysis of intra- and inter-country transmissions was performed on the Additional Information column of the metadata and are documented in Table 3. In Africa, we found that more intra-country transmissions were reported, especially in South Africa. Senegal had few family cluster transmissions, while few imported cases or inter-country transmissions were observed in Madagascar, Nigeria, and Senegal.

Table 3. Country transmissions statistics across continents.

Continent	Country	Intra-country transmission	Inter-country transmission
Africa	Madagascar	Not recorded	France (2)
	Nigeria	Not recorded	Italy (1), Europe (1)
	Senegal	Family cluster (4)	Imported cases (3), Spain (1), Italy (1)
	South Africa	King Cetshawayo (1), Umkhanyakude (66), Ilembe (122), Umgungundlovu (41), eThekweni (115), Amajuba (29), Zululand (71), uThukela (62), Ugu (55), Umzinyathi (16), Harry Gwala (12), Sisonke (8)	Not recorded
Europe	Austria	Vienna (2),	Not recorded
	Czech Republic	Cinovecka (1), Liberec (3), Dobra Voda (3), Benatky nad Jizerou (1), Bassova (1), Ukraine (3), Madona di Campiglio (1), Austria (1), Darkov (3), Slovakia (1), Switzerland (1), Doksy (1), Vojkov (2), Brno (1)	Italy (2)
	Germany	Not recorded	Italy (1)
	Italy	Not recorded	China (4)
	Slovakia	Not recorded	France (1)
	Spain	Barcelona (1)	Not recorded
Asia	Cambodia	Wuhan (1)	
	China	Wuhan (18), Pakistan (1), Iran (1),	United Kingdom (5), USA (6), Mexico (1), Spain (3), Germany (1), Italy (5), France (4), Budapest (3), Greece (1), Switzerland (1), Norway (1),
	India	Faridabad (2), Dubai (6)	Not recorded
	Kuwait	Not recorded	USA (1), United Kingdom (1), Italy (1)
	Lebanon	Iran (4)	United Kingdom (2), Italy (1), Egypt (1)
	Mongolia	Not recorded	Imported cases (3)
	South Korea	Domestic infection (4)	Oversea inflow (11)
	Taiwan	Turkey (1), China (2), Japan (2), Indonesia (1), Czech-Republic (1)	France (1), USA (4), England-Belgium-Germany (1), Portugal (1)
	Thailand	Wuhan (2)	Not recorded
North America	Costa Rica	Florida-New York-USA (5), Mexico (1)	Spain (2), Argentina-Peru (1)
	Mexico	USA (1), Guadalajara (1), Gomez Palacios (1)	Italy (1), Argentina (1),
	Panama	Local (231), USA (5), Puerto Rico (1)	Germany (1)
South America	Brazil	Local contact (3)	Germany-Italy-Spain (7), Milan-Italy (4), United Kingdom (1), USA (1)
	Chile	Asia-Europe (1)	Not recorded
	Ecuador	Local transmission (4)	Netherlands (1)
Oceania	New Zealand	Not recorded	Iran (2)
	Australia	Not recorded	China (1)

In Europe, intra-country transmissions were mostly observed in the Czech Republic and Spain, while inter-country transmissions were observed in Italy, China, and France.

Contact with patient zero (0), family clusters, congregations, seafood wholesale markets, local communities, passengers on a Nile river cruise ship, and medical college hospitals were means of intra- and inter- country transmissions in Asia. Inter-country transmissions specifically came from Africa, Europe, and North America.

Evidence of reinfection was reported in North America, specifically the US. Private events, contact with infected patients, local communities and local airports, were avenues of intra- and inter- country transmissions. Europe and South America were the sources of transmission.

Evidence of reinfection was also reported in South America, specifically Brazil. Local hospitals and international travel were the means of inter- and intra-country transmissions in this continent.

Oceania witnessed intra- and inter-country transmissions from Asia, specifically Iran and China.

Genome pattern analysis

The SOM component planes allowed an investigation of countries that share similar genome pattern expressions of SARS-CoV-2 and which patterns permeate the different regions. To account for the variability of SOM neighborhood structure at every SOM run, the reference genome was included in the experiment datasets during each training phase. Our topologies possess random (but controllable) discontinuities that permit more flexible self-organization with high-dimensional data, thus, preserving the ensuing map structure as much as possible. The training was performed by gender, per continent, for clean whole genome sequences (>29000 bp), i.e., without ambiguous nucleotides. During the training, the male and female samples were trained separately for each continent. However, due to the paucity of data in the Oceanian region, its dataset was merged with South America, resulting in 10 different maps (Figures 3-7). We observed that globally, there are inter- and intra- country transmissions evident in the pattern (dis) similarities exhibited by the various SOM maps. Most component planes exhibiting intra-country sub-strains show highly disparate and variable cluster patterns with well separated boundaries, indicating emergence of new sub-strain(s) with rapid nucleotide mutations. However, component planes exhibiting inter-country sub-strain patterns possess clear patterns without sharp boundaries, indicating fewer nucleotide changes. Interestingly, the gradual evolution of the cluster patterns into well separated boundaries can be traced, hence providing opportunities for predicting the emergence of new sub-strains. Furthermore, some of the patients retained the reference genome pattern (i.e., had similar pattern as component plane 1, encircled in red), indicating no significant mutation in the nucleotide composition.

Cognitive knowledge extraction

Next, we decoupled the SOM correlation hunting matrix space (Vesanto and Ahola, 1999), and attributed these associations to disparate clusters of discovered viral sub-strains, resulting in a cognitive map that links similar transmission routes. Table S9 (*Extended data*) and Table S10 (*Extended data*) distinguish transmission routes for male

Figure 3. Self-organizing map component planes for African countries. Component plane 1 is the reference genome pattern. (a) Male patients – Component planes: [2-129] are patterns from South Africa; [130] is the pattern from Gambia; [131] is the pattern from Algeria; [132-136] are patterns from Egypt; [137-143] are patterns from Nigeria; [144-145] are patterns from Morocco; [146-147] are patterns from Mozambique; [148] is the pattern from Nigeria; [149-160] are patterns from Senegal. (b) Female patients: Component planes [2-219] are patterns from South Africa; [220] is the pattern from Algeria; [221-222] are patterns from Egypt; [223-228] are patterns from Tunisia; [229] is the pattern from Mogadishu; [230-233] is pattern from Nigeria; [234-237] are patterns from Senegal.

Figure 4. Self-organizing map component planes for European countries. Component plane 1 is the reference genome pattern. (a) Male patients – Component planes [2] is the pattern from Switzerland; [3-5] are patterns from Faroe Islands; [6] is the pattern from Poland; [7-8] are patterns from Belgium; [9-21] are patterns from Romania; [22-56] are patterns from Spain; [57-58] are patterns from Georgia; [59-110] are patterns from Italy; [111-123] are patterns from Russia; [124-154] are patterns from France; [155] is pattern from Slovakia; [156-159] are patterns from Hungary; [160-163] is pattern from Ukraine; [164] is the pattern from Sweden; [165] is the pattern from Bosnia and Herzegovina; [166-179] are patterns from Czech Republic. (b) Female patients: Component planes [2] is the pattern from Switzerland; [3-4] are patterns from Faroe Islands; [5-6] is pattern from Belgium; [7-12] are patterns from Germany; [13-33] are patterns from Romania; [34-62] are patterns from Spain; [63] are patterns from Georgia; [64-95] are patterns from Italy; [96-118] are patterns from Russia; [119-146] are patterns from France; [147-148] is pattern from Slovakia; [149-153] are patterns from Hungary; [154-157] is pattern from Ukraine; [158] is the pattern from Austria; [159-169] are patterns from Czech Republic.

Figure 5. Self-organizing map component planes for Asian countries. Component plane 1 is the reference genome pattern. (a) Male patients – Component planes: [2-11] are patterns from Singapore; [12] is the pattern from Iraq; [13-53] are patterns from China; [54] is the pattern from Kuwait; [55-71] are patterns from Malaysia; [72] is the pattern from Sri Lanka; [73-82] are patterns from Bangladesh; [83-167] are patterns from India; [168-169] are patterns from South Korea; [170] is the pattern from Kazakhstan; [171-176] are patterns from Indonesia; [177-181] are patterns from Turkey; [182-187] are patterns from Taiwan; [188-196] are patterns from Vietnam; [197] is the pattern from Israel; [198-215] are patterns from Saudi Arabia; [216-222] are patterns from Oman; [223-225] are patterns from Lebanon; [226-229] are patterns from United Arab Emirates; [230-299] are patterns from Japan. (b) Female patients: Component planes: [2-5] are patterns from Singapore; [6-40] are patterns from China; [41-53] is pattern from Malaysia; [54-58] are patterns from Bangladesh; [59-154] are patterns from India; [155-160] is pattern from Kazakhstan; [160-169] are patterns from Indonesia; [170-172] are patterns from Turkey; [173-184] are patterns from Taiwan; [185-201] are patterns from Vietnam; [202-215] are patterns from Saudi Arabia; [216-217] are patterns from Pakistan; [218-225] are patterns from Oman; [226-227] are patterns from Lebanon; [228] is the pattern from United Arab Emirates; [229-300] are patterns from Japan.

Figure 6. Self-organizing map component planes for North American countries. Component plane 1 is the reference genome pattern. (a) Male patients – Component planes: [2-17] are patterns from Mexico; [18-101] are patterns from USA; [102-108] are patterns from Saint Martin; [109-110] are patterns from Guadeloupe; [111-127] are patterns from Canada; [128-144] are patterns from Costa Rica. (b) Female patients – Component planes: [2-15] are patterns from Mexico; [16-64] are patterns from USA; [65-67] are patterns from Saint Martin; [68-74] are patterns from Guadeloupe; [75-79] are patterns from Canada; [80-87] are patterns from Costa Rica.

Figure 7. Self-organizing map component planes for South American and Oceanian countries. Component plane 1 is the reference genome pattern. (a) Male patients – Component planes: [2] is the pattern from Chile; [3] is the pattern from Argentina; [4-15] are patterns from Colombia; [16-19] are patterns from Ecuador; [20] is the pattern from Peru; [21-79] are patterns from Brazil. (b) Female patients – Component planes: [2] is the pattern from Venezuela; [3] is the pattern from Argentina; [4-21] are patterns from Colombia; [22-78] are patterns from Brazil.

and female patients (columns 3 and 4), respectively. Also shown in the tables are dominant transmission cases, and patients that retained the reference genome pattern (column 1). For male patients (Table S9, *Extended data*), no dominant intra-country transmission or spread is observed in Africa, while only South African and Tunisian patients retained the reference genome pattern. Dominant intra-country transmission is observed in Europe (Faroe Islands, France, Sweden, and Czech Republic), while the reference genome pattern appears dominant in Poland, Romania, Spain, Italy, Russia, and France. In Asia, dominant intra-country spread is observed in Singapore, China, Indonesia, Israel, and Japan, while the reference genome pattern is retained in China, Sri Lanka, Bangladesh, Taiwan, Saudi Arabia, and The United Arab Emirates. In North America, intra-country transmission occurs in Mexico and USA, while the reference genome is retained only in Mexico and USA. Dominant intra-country patterns are found in Brazil, while the reference genome pattern is retained in Chile, Ecuador, and Brazil. In Oceania, intra-country transmission is observed in Australia, while the reference genome is retained in Guam.

For female patients (Table S10, *Extended data*), no dominant intra-country transmission is observed in Africa, while only South African patients retained the reference genome pattern. Intra-country transmission is observed in Europe (Faroe Islands, Belgium, Russia, France, with dominant transmission in the Czech Republic), while the reference genome pattern appears dominant in Belgium, Romania, Spain, Italy, France, with only one copy in the Czech Republic. In Asia, dominant intra-country spread is observed in Japan with few intra-country transmissions occurring in Singapore, China, Bangladesh, India, Indonesia, Taiwan, Saudi Arabia, Pakistan, and Oman, while the reference genome pattern is retained in China, Sri Lanka, Bangladesh, Taiwan, Saudi Arabia, and United Arab Emirates. In North America, intra-country transmission appears dominant in Mexico and USA with few transmissions in Saint Martin, Guadalupe, and Canada, while the reference genome is retained in Mexico and USA. Intra-country patterns are found in Colombia and Brazil, while the reference genome pattern is also retained in Colombia and Brazil. In Oceania, no intra-country transmission or reference genome patterns are observed.

The benefits of our cognitive map cannot be overemphasized. While manual annotation of transmission routes indicates few recorded cases (see [Table 3](#)), our cognitive map efficiently traced each patient and groups the patient according to two transmission routes, i.e., inter- and intra-country transmissions. The map would benefit contact tracing and early disease surveillance for informed medical decisions if integrated into current epidemiological protocol.

Variant surveillance analysis

Two classes of SARS-CoV-2 variants namely, Variant of Interest (VOI) and Variant of Concern (VOC) have been defined by World Health Organization (WHO) and US Centers for Disease Control and Prevention (CDC). The B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.2 (Delta), and P.1 (Gamma) variants have been classified as variants of concern ([Campbell et al., 2021](#)) and are circulating around the world. Another variant classification type known as Variant of High Consequence (VOHC) has recently been introduced by US CDC, and to date, no VOHC have been identified ([Chadha et al., 2021](#)). As attention on the pandemic shifts to the emergence of new VOC, understanding the variability between new variants and non-VOC lineages is becoming increasingly important for surveillance and maintaining the effectiveness of public health as well as vaccination programs ([Jewell, 2021](#)).

Analysis of the variants type by continent is presented in [Figure 8](#). We observe that Europe records the highest number of VOCs as follows (VOC Gamma: 606955, VOC Beta: 407879, VOC Alpha: 1338089, and VOC Delta: 72290). North America follows with (VOC Gamma: 36441, VOC Beta: 195360, VOC Alpha: 683838, and VOC Delta: 9181). Asia follows with (VOC Gamma: 41946, VOC Beta: 8112, VOC Alpha: 130662, and VOC Delta: 1570). South America follows with (VOC Gamma: 21860, VOC Beta: 538, VOC Alpha: 31848, and VOC Delta: 325). Africa follows with (VOC Gamma: 78, VOC Beta: 6645, VOC Alpha: 12388, and VOC Delta: 249). And Oceania follows with (VOC Gamma: 38, VOC Beta: 56, VOC Alpha: 3467, and VOC Delta: 1343). Oceania appears to be experiencing increased proportion of VOC Delta over its closest form, VOC Alpha ($1343/3667 = 38.74\%$) compared to Africa ($249/12388 = 2.01\%$), South America ($325/31848 = 1.02\%$), Asia ($1570/130662 = 1.20\%$), North America ($9181/683838 = 1.34\%$), and Europe ($72290/1338089 = 5.40\%$).

Figure 8. SARS-CoV-2 variant analysis across continents.

Discussion

Results of epidemiological and surveillance analysis revealed increased intra-country transmissions, demanding localized strategic planning and informed response to the pandemic. Understanding how, when and in what circumstances the virus spreads is crucial to developing effective public health measures for infection prevention and control. Genetic variants of SARS-CoV-2 are circulating around the world with routine surveillance (cases, deaths, health workers, hospitalizations) becoming more critical to monitor viral mutations and variants through sequence-based analysis, laboratory studies, and epidemiological investigations.

Smart healthcare system: a proposal

A novel IoHT framework is proposed in this section to support our drive towards a smart healthcare system. The proposed framework as shown in Figure 9 is a multi-layer expert system architecture that coordinates a set of collaborative components or layers, namely, 1) smartphone/end-user, 2) IoHT, 3) data warehouse and knowledge base (KB), and 4) medical expert opinion.

Figure 9. Proposed Internet of Health Things (IoHT) framework. DS, data source; DW, data warehouse; KB, knowledge base.

- 1) The *smartphone/end-user layer* is intended to offer medical services to end-users (individuals, epidemiologists, physicians, policymakers, government), and serve two major functions: a) assist in primary collection of demographic data—as users register and signup for medical services (e.g., syndrome check/confirm), physiological data—symptoms, medical history, and any other data—assisted by the IoHT; and b) assist in secondary data processing (deposited genomes or any other data of emerging infectious diseases). This layer communicates with the IoHT layer and the data warehouse and KB layers to perform analysis of syndromes, perform global disease surveillance, display disease status and produce surveillance reports.
- 2) The *IoHT layer* holds smart health objects in the form of intelligent apps or plug-in devices that provide personal health records to the KB layer of the framework. It consists of interconnected objects with the capacity to exchange and process data to improve patient health. An interface/plug-in that aids communication between sensors and the mobile layer is proposed. Sensors are required to assist this interface to measure physiological parameters such as temperature, blood pressure, pulse rate, and oxygen level, for use in the diagnosis of infectious diseases. This synergy can transform the mobile layer for regulated display, transfer, storage, or conversion of patient-specific medical data from a connected device.
- 3) The *data warehouse and KB layer* curates and stores clinical datasets (e.g., genome sequences) voluntarily donated by patients or deposited by clinicians or any other data the world over. This layer also contains data modelling tools and data marts (subsets of data). Data can be crowdsourced for disease databases for the purpose of intelligent disease surveillance facilitated by intelligent (machine learning, ML) algorithms, to produce surveillance reports (epidemiological analysis, genome pattern analysis and cognitive knowledge map) useful for tracking disease transmission routes and supporting inter- and intra-country contact tracing. The KB is a collection of data generated by end-users, including rules and intelligent algorithms, for diagnostics and surveillance analytics. The KB layer therefore represents the engine room of the proposed framework and communicates with all other layers for efficient storage and retrieval purposes as well as coordinated health information processing.
- 4) The *knowledge opinion layer* enables contributions from medical experts for knowledge simplification and evidence-based practice that defines thresholds for diagnosis and prevention of emerging infectious diseases. These contributions are processed by the KB, to offer reference labels to input datasets. In this paper, the taxonomies created to normalize patient status and specimen source will improve the precision of feature labelling in intelligent system classification.

To achieve layer 1, the Python programming language and Node.js could be used to develop a Smart Medical Assistant and the app interface for early syndromic assessment of infectious disease situations. To achieve layer 2, a cloud-based system embedding the various devices to guarantee ubiquitous computing, context-awareness and wireless communication could be adopted. The objects would be interconnected for exchanging and processing data for improved patient's healthcare services. Embedded soft sensors could also be programmed for the purpose of collecting physiological measurements such as temperature, blood pressure, and oxygen level. Synergy between the IoHT and the smartphone layers will transform the mobile layer into a regulated retrieval and storage system. To achieve layer 3, AI/data science tools/models (neural network, multi-criteria, fuzzy, and tree-based models) could be applied to mine experiential knowledge from primary/secondary datasets into informed decision charts/reports. To achieve layer 4, contributions from medical experts for knowledge simplification and evidence-based practice that defines the 'universe of discourse' for diagnosis and prevention of emergent infectious diseases could be performed. To achieve context-awareness, a location-based system could be implemented using two major steps: spatial database design – satellite image acquisition, identification, and abstraction of relevant features within the study area, start and end boundary extraction; and geo-modelling, GIS mapping and testbed/prototype design – mapping the location information and details from the health registry data to the testbed/prototype using a geo-analysis software such as ArcGIS. Steps to achieving this activity include geolocation superimposition on the extracted surface after image digitalization, geodatabase modeling and integration and prototype/testbed development.

To test and evaluate the proposed system (for end-users' acceptance), data could be crowdsourced from patients/participants and healthcare workers for the purpose of building an intelligent disease surveillance system facilitated by ML/AI algorithms.

The following subsections discuss the data collection procedures, inference/evaluation criteria and expected research outcomes of the smart medical/healthcare system proposal.

Data collection procedures

To achieve the smart healthcare system, researchers would be drawn from multidisciplinary fields including Computing, Medical, Engineering and Social Sciences. The core system's design, implementation and evaluation will be carried out by researchers in the computing and engineering fields. Experiential knowledge of symptoms, assessments and end-users' evaluation would be obtained from patients and medical experts. Community interaction, development of data instruments, and instruments evaluation would be performed by social scientists.

To ensure that activities are properly monitored during implementation, coordinators from participating/collaborating universities in Nigeria would be appointed. The research population would be drawn from various states and cover a sizable number of participants. Stratified random sampling could be adopted to select required participants for the study.

Inference/evaluation criteria

To test and evaluate the proposed IoHT system for end-users' acceptance, data could be crowdsourced from male and female participants (patients and healthcare workers). A user acceptance test, where actual end-users test the system to determine its ability to carry out the required tasks it was designed to address in real-world situations or validate whether all the specific requirements are satisfied, with appropriate inference/evaluation of the system, would also be carried out.

Expected research outcomes

This research is expected to produce four major outcomes to support investigation of the second and third hypotheses, as follows:

- 1) *Improved healthcare services.* A smart healthcare system powered by IoHT with Smart Medical Assistant, which connects patients and healthcare resources in real-time and aids the syndromic assessment of infectious disease(s), is expected. As healthcare informatics research seeks a nexus between theory and practice, the research promises to engage and build a critical mass of early career researchers from the North-South region of Nigeria to actualize the most-needed knowledge-driven healthcare system that will create national, regional, and global impact.
- 2) *Enhanced healthcare decision support.* A proposed location-based (spatial) system in the form of an app will enable real-time and accurate localization of available healthcare facilities as well as services offered by these facilities. Another deliverable is an expert recommendation system, which will provide useful information such as disease surveillance reports from crowdsourced data for enhanced healthcare decisions. By integrating a decision support system that influences society, the community/society will be involved, and Government policies will be re-modelled towards a smart and healthy society. With this breakthrough, the capacity of the university system will be strengthened to impact the various sectors of the economy and restore confidence in the various stakeholders.
- 3) *Increased number and quality of research outputs.* Quality research publications and patents are expected from this research that will advance science, technology, and innovation, as well as research progress, for the dissemination of innovative and reproducible research results. The method/design, findings, and results will have a wider impact on the global community.
- 4) *New breed of experienced mentees.* This research will produce strong and independent mentees with multi-disciplinary experience to sustain the research into the future.

Limitations

This work is limited to disease surveillance analysis using secondary data, pattern visualization and generic grouping of transmission routes for discovery of intra-country sub-strains. Although our proposed IoHT framework integrates a hybrid solution for collaborative data analytics, practical implementation of this proposal is required to derive the full benefits of the framework.

Conclusions

Community transmission and antiretroviral treatments can engender novel mutations in a virus, potentially resulting in more virulent strains with higher mortality and strains resistant to treatment. Hence, systematic tracking of demographic

and clinical data as well as sub-strain information is indispensable to effectively contain infectious diseases. Mutant variability analysis and precise sub-strain prediction can also serve as useful precursors to emerging viral sub-strain discovery and quality vaccine formulation. This work identified the existence and spread of SARS-CoV-2 viral sub-strains among infected patients. Using cognitive knowledge mining, transmission routes (inter- and intra-country transmissions) were efficiently separated by gender. Clinicians and medical experts could exploit the epidemiology and genome pattern analysis, as well as the proposed system framework, to support efficient medical decisions on real-time contact tracing, disease diagnosis and global disease surveillance.

Data availability

Underlying data

Sequence data are available from [GISAID](#) (Elbe and Buckland-Merrett, 2017). GISAID accession numbers are presented in the Acknowledgment Table. Access to the data requires registration and agreement to the conditions for use at: <https://www.gisaid.org/registration/register/>.

Open Science Framework: SARS-CoV-2 Genome Datasets Analytics. <https://doi.org/10.17605/OSF.IO/U7G4D>. (Ekpenyong *et al.*, 2021b).

This project contains the following underlying data:

- gisaid_hcov-19_acknowledgement_table_2021_08_11_04.pdf (GISAID Acknowledgement table)

Extended data

Open Science Framework: SARS-CoV-2 Genome Datasets Analytics. <https://doi.org/10.17605/OSF.IO/U7G4D> (Ekpenyong *et al.*, 2021b).

This project contains the following extended data:

- Appendix.pdf (Algorithm 1 and Tables S1-10)

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

References

- Ainsworth B, Miller S, Denison-Day J, *et al.*: **Infection control behavior at home during the COVID-19 pandemic: observational study of a web-based behavioral intervention (Germ defence).** *Journal of Medical Internet Research*. 2021; **23**(2): e22197.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Campbell F, Archer B, Laurenson-Schafer H, *et al.*: **Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021.** *Eurosurveillance*. 2021; **26**(24): 2100509.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chadha J, Khullar L, Mittal N: **Facing the wrath of enigmatic mutations: A review on the emergence of SARS-CoV-2 variants amid COVID-19 pandemic.** *Environ Microbiol*. 2021.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Choi J, Cho Y, Shim E, *et al.*: **Web-based infectious disease surveillance systems and public health perspectives: a systematic review.** *BMC Public Health*. 2016; **16**(1): 1–10.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chretien JP, Riley S, George DB: **Mathematical modeling of the West Africa Ebola epidemic.** *Elife*. 2015; **4**: e09186.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Christaki E: **New technologies in predicting, preventing and controlling emerging infectious diseases.** *Virulence*. 2015; **6**(6): 558–565.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dong E, Du H, Gardner L: **An interactive web-based dashboard to track COVID-19 in real time.** *Lancet Infect Dis*. 2020; **20**(5): 533–534.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Edoho ME, Ekpenyong ME, Momodu AB, *et al.*: **Mining the Human Metabolome for Precision Oncology Research.** *Proceedings of the 4th International Conference on Medical and Health Informatics*. 2020: 8–17.
[Reference Source](#)
- Ekpenyong ME, Edoho ME, Inyang UG, *et al.*: **A hybrid computational framework for intelligent inter-continent SARS-CoV-2 sub-strains characterization and prediction.** *Sci Rep*. 2021a; **11**(1): 1–25.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ekpenyong ME, Udo IJ, Edoho ME, *et al.*: **SARS-CoV-2 Genome Datasets Analytics.** 2021b.
[Publisher Full Text](#)
- Elbe S, Buckland-Merrett G: **Data, disease and diplomacy: GISAID's innovative contribution to global health.** *Glob Chall*. 2017; **1**(1): 33–46.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- George DB, Taylor W, Shaman J, *et al.*: **Technology to advance infectious disease forecasting for outbreak management.** *Nat Commun*. 2019; **10**(1): 1–4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grabowski F, Kochanek M, Lipniacki T: **L18F substrain of SARS-CoV-2 VOC-202012/01 is rapidly spreading in England.** *medRxiv*. 2021;
[Publisher Full Text](#)
- Jewell BL: **Monitoring differences between the SARS-CoV-2 B. 1.1.7 variant and other lineages.** *Lancet Public Health*. 2021; **6**(5): e267–e268.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Konings F, Perkins MD, Kuhn JH, *et al.*: **SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse.** *Nat*

Microbiol. 2021; **6**: 821–823.

[PubMed Abstract](#) | [Publisher Full Text](#)

Kangas J, Kohonen T, Laaksonen J: **Variants of self-organizing maps.** *IEEE Transactions on Neural Networks.* 1990; **1**(1): 93–99.

[PubMed Abstract](#) | [Publisher Full Text](#)

Koyama T, Weeraratne D, Snowden JL, *et al.*: **Emergence of drift variants that may affect COVID-19 vaccine development and antibody treatment.** *Pathogens.* 2020; **9**(5): 324.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Meltzer MI, Santibanez S, Fischer LS, *et al.*: **Modeling in Real Time During the Ebola Response.** *MMWR Suppl.* 2016; **65**(Suppl-3): 85–89.

[PubMed Abstract](#) | [Publisher Full Text](#)

Rainisch G, Asher J, George D, *et al.*: **Estimating ebola treatment needs, United States.** *Emerg Infect Dis.* 2015; **21**(7): 1273–1275.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Richmond CS, Sabin AP, Jobe DA, *et al.*: **Interregional SARS-CoV-2 spread from a single introduction outbreak in a meat-packing plant in northeast Iowa.** *MedRxiv.* 2020;

[Publisher Full Text](#)

Simonsen L, Gog JR, Olson D, *et al.*: **Infectious disease surveillance in the big data era: towards faster and locally relevant systems.** *J Infect Dis.* 2016; **214**(suppl_4): S380–S385.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Vesanto J, Ahola J: **Hunting for Correlations in Data Using the Self-Organizing Map.** *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications.* 1999: 279–285.

Wang R, Chen J, Gao K, *et al.*: **Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants.** *Commun Biol.* 2021; **4**(1): 1–14.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Wang R, Chen J, Gao K, *et al.*: **Characterizing SARS-CoV-2 mutations in the United States.** 2020; *arXiv preprint arXiv:* 2007.12692.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Yang GZ, Nelson BJ, Murphy RR, *et al.*: **Combating COVID-19—The role of robotics in managing public health and infectious diseases.** *Sci Robot.* 2020; **5**(40).

[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 09 December 2021

<https://doi.org/10.5256/f1000research.58540.r100582>

© 2021 Coccia M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mario Coccia

Collegio Carlo Alberto, CNR -National Research Council of Italy, Torino, Italy

SARS-CoV-2 genome datasets analytics for informed infectious disease surveillance

The topics of this paper are interesting, but the structure and content must be revised, and results have to be better explained by authors before to be reconsidered for publication.

Title has to be shorter. Abstract has to be shorter focusing on results and explaining reliable health and social implications.

Introduction has to better clarify the research questions of this study and provide more theoretical background. Authors have to better describe the different sources of transmission dynamics of COVID-19 (e.g., mobility, environment, pollution, etc.) and risk factors in society, that can accelerate diffusion of this novel coronavirus. In addition, the application of a collaborative Internet of Health Things (IoHT) is one of the strategies of control of COVID/19 and support of disease surveillance in poor health systems because it has to be associated with a certain level of Health expenditure. IN any case this tool has to be discussed in a general strategy of prevention to cope with COVID/19 transmission (See suggested readings that must be all read and used in the text).

Hypotheses are too general and vague and have to be better defined and specified.

Methods of this study is not clear. The section of Materials and methods must be re-structured with following sections:

- Sample and data
- Measures of variables
- Models and Data analysis procedure.

Results has a lot of subheadings that create fragmentation of the paper and confusion. All

subheadings can be removed.

Table 2 and 3 have to indicate the period of data.

Figure 3-7 are too small, and they are unreadable so I suggest putting in Appendix and authors can explain better results in the text .

Figure 8 has to indicate the period in the title and all data should be presented per 100 000 people to be comparable...

Discussion can remove subheadings. Figure 9....is messy. The paper has a lot of figures that are difficult to digest, they have to be reduced inserting in the text the most important ones, the other if appendix.

To reiterate, avoiding in the just mentioned sections, sub-headings that create fragmentation of the paper.

The small paragraph of Limitations can be put in conclusions.

Conclusion has not to be a summary, but authors have to focus on manifold limitations of this study and provide suggestions of health and social policy.

Overall, then, the paper is interesting, but structure is confused. Theoretical framework is weak, and some results create confusion... structure of the paper has to be improved; study design, discussion and presentation of results have to be clarified using suggested comments.

If the paper is improved, by using all comments, maybe it can be considered.

Suggested readings of relevant papers that have to be read and all inserted in the text and references.

Ahmed, A., Boopathy, P., Sudhagara Rajan, S. 2022. Artificial intelligence for the novel corona virus (COVID-19) pandemic: Opportunities, challenges, and future directions, International Journal of E-Health and Medical Communications 13(2)

Coccia M. 2021. Preparedness of countries to face covid-19 pandemic crisis: Strategic positioning and underlying structural factors to support strategies of prevention of pandemic threats, Environmental Research, n. 111678, <https://doi.org/10.1016/j.envres.2021.111678>

Saini, G., Swahn, M.H., Aneja, R. 2021. Disentangling the Coronavirus Disease 2019 Health Disparities in African Americans: Biological, Environmental, and Social Factors)(2021) Open Forum Infectious Diseases, 8 (3).

Fountain-Jones, N.M., Appaw, R.C., Carver, S., Didelot, X., Volz, E., Charleston, M. 2020. Emerging phylogenetic structure of the SARS-CoV-2 pandemic ()(2020) Virus Evolution, 6 (2), art. no. veaa082

Coccia M. 2021. The relation between length of lockdown, numbers of infected people and deaths of Covid-19, and economic growth of countries: Lessons learned to cope with future pandemics similar to Covid-19. Science of The Total Environment, n. 145801.

<https://doi.org/10.1016/j.scitotenv.2021.145801>

Pan, D., Sze, S., Minhas, J.S., Bangash, M.N., Pareek, N., Divall, P., Williams, C.M., (...), Pareek, M. 2020. The impact of ethnicity on clinical outcomes in COVID-19: A systematic review (2020) *EClinicalMedicine*, 23, art. no. 100404

Gomes, L., Jeewandara, C., Jayadas, T.P., (...), Waggoner, J., Malavige, G.N. 2022. Surveillance of SARS-CoV-2 variants of concern by identification of single nucleotide polymorphisms in the spike protein by a multiplex real-time PCR, *Journal of Virological Methods* 300, 114374

Candido, D.S., Claro, I.M., de Jesus, J.G., Souza, W.M., Moreira, F.R.R., Dellicour, S., Mellan, T.A., (...), Faria, N.R. 2020. Evolution and epidemic spread of SARS-CoV-2 in Brazil (), (2020) *Science*, 369 (6508), pp. 1255-1260.

Coccia M. 2021. Effects of the spread of COVID-19 on public health of polluted cities: results of the first wave for explaining the déjà vu in the second wave of COVID-19 pandemic and epidemics of future vital agents. *Environmental Science and Pollution Research*. 28(15), 19147-19154. <https://doi.org/10.1007/s11356-020-11662-7>

Nana-Sinkam, P., Kraschnewski, J., Sacco, R., Chavez, J., Fouad, M., Gal, T., Auyoung, M., (...), Behar-Zusman, V. 2021. Health disparities and equity in the era of COVID-19 (2021) *Journal of Clinical and Translational Science*, art. no. e99.

Coccia M. 2021. Pandemic Prevention: Lessons from COVID-19. *Encyclopedia of COVID-19*, 1, pp. 433-444. <https://doi.org/10.3390/encyclopedia1020036>

Forde, A.T., Crookes, D.M., Suglia, S.F., Demmer, R.T. 2019. The weathering hypothesis as an explanation for racial disparities in health: a systematic review (2019) *Annals of Epidemiology*, 33, pp. 1-18.e3.

Davies, N.G., Jarvis, C.I., van Zandvoort, K., Clifford, S., Sun, F.Y., Funk, S., Medley, G., (...), Keogh, R.H. 2021. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7 () (2021) *Nature*, 593 (7858), pp. 270-274

Coccia M. 2021. High health expenditures and low exposure of population to air pollution as critical factors that can reduce fatality rate in COVID-19 pandemic crisis: a global analysis. *Environmental Research*, vol. 199, Article number 111339, <https://doi.org/10.1016/j.envres.2021.111339>

Jeewandara, C., Jayathilaka, D., Ranasinghe, D., Hsu, N.S., Ariyaratne, D., Jayadas, T.T., Panambara Arachchige, D.M., (...), Malavige, G.N. 2021. Genomic and Epidemiological Analysis of SARS-CoV-2 Viruses in Sri Lanka (2021) *Frontiers in Microbiology*, 12, art. no. 722838

References

1. Correction to *Lancet Infect Dis* 2020; published online March 11, [https://doi.org/10.1016/S1473-3099\(20\)30144-4](https://doi.org/10.1016/S1473-3099(20)30144-4). *Lancet Infect Dis*. 20 (5): e79 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Coccia M: Effects of the spread of COVID-19 on public health of polluted cities: results of the first wave for explaining the déjà vu in the second wave of COVID-19 pandemic and epidemics of future

- vital agents. *Environ Sci Pollut Res Int*. 2021; **28** (15): 19147-19154 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Coccia M: Pandemic Prevention: Lessons from COVID-19. *Encyclopedia*. 2021; **1** (2): 433-444 [Publisher Full Text](#)
4. Coccia M: High health expenditures and low exposure of population to air pollution as critical factors that can reduce fatality rate in COVID-19 pandemic crisis: a global analysis. *Environ Res*. **199**: 111339 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Jeewandara C, Jayathilaka D, Ranasinghe D, Hsu NS, et al.: Genomic and Epidemiological Analysis of SARS-CoV-2 Viruses in Sri Lanka. *Front Microbiol*. 2021; **12**: 722838 [PubMed Abstract](#) | [Publisher Full Text](#)
6. Carcereny A, Martínez-Velázquez A, Bosch A, Allende A, et al.: Monitoring Emergence of the SARS-CoV-2 B.1.1.7 Variant through the Spanish National SARS-CoV-2 Wastewater Surveillance System (VATar COVID-19). *Environmental Science & Technology*. 2021; **55** (17): 11756-11766 [Publisher Full Text](#)
7. Forde AT, Crookes DM, Suglia SF, Demmer RT: The weathering hypothesis as an explanation for racial disparities in health: a systematic review. *Ann Epidemiol*. **33**: 1-18.e3 [PubMed Abstract](#) | [Publisher Full Text](#)
8. Nana-Sinkam P, Kraschnewski J, Sacco R, Chavez J, et al.: Health disparities and equity in the era of COVID-19. *J Clin Transl Sci*. 2021; **5** (1): e99 [PubMed Abstract](#) | [Publisher Full Text](#)
9. Candido D, Claro I, de Jesus J, Souza W, et al.: Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*. 2020; **369** (6508): 1255-1260 [Publisher Full Text](#)
10. Gomes L, Jeewandara C, Jayadas TP, Dissanayake O, et al.: Surveillance of SARS-CoV-2 variants of concern by identification of single nucleotide polymorphisms in the spike protein by a multiplex real-time PCR. *J Virol Methods*. 2021; **300**: 114374 [PubMed Abstract](#) | [Publisher Full Text](#)
11. Pan D, Sze S, Minhas JS, Bangash MN, et al.: The impact of ethnicity on clinical outcomes in COVID-19: A systematic review. *EClinicalMedicine*. 2020; **23**: 100404 [PubMed Abstract](#) | [Publisher Full Text](#)
12. Fountain-Jones NM, Appaw RC, Carver S, Didelot X, et al.: Emerging phylogenetic structure of the SARS-CoV-2 pandemic. *Virus Evol*. 2020; **6** (2): veaa082 [PubMed Abstract](#) | [Publisher Full Text](#)
13. Saini G, Swahn MH, Aneja R: Disentangling the Coronavirus Disease 2019 Health Disparities in African Americans: Biological, Environmental, and Social Factors. *Open Forum Infect Dis*. 2021; **8** (3): ofab064 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: COVID-19 and public health

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research