RESEARCH ARTICLE

REVISED Unified framework for the ingestion of early epidemic data for downstream data analytics

[version 2; peer review: 1 approved, 2 approved with reservations]

Everlyn Kamau[1], Sadie Kelly[2], Dhruv Darji[3,4], Amrish Y. Baidjoe[5,6], John S. Brownstein[7,8], Finlay Campbell[9], Abhishek Dasgupta[2,10], Marie-Amélie Degail[11], Anastasiia Demidova[12], Luca Ferretti[3,4], Aimee Han[7], Stephen Leshan Koyie[11], Patricia Ndumbi Ngamala[11], Olivier Le Polain[11], Amanda Rojek[3,4], Jacquelin Sauer[13], Samuel V. Scarpino[14-16], Kara Sewalk[7], Juliana Sopko[7], Stanislaw Zakrzewski[17], Laura Merson[4,18], Moritz U. G. Kraemer[2,4]

[1]Francis I. Proctor Foundation, University of California San Francisco, San Francisco, California, USA
[2]Department of Biology, University of Oxford, Oxford, England, UK
[3]Nuffield Department of Medicine, University of Oxford, Oxford, England, UK
[4]Pandemic Sciences Institute, University of Oxford, Oxford, England, UK
[5]Medecins Sans Frontieres, Operational Centre Brussels (OCB), Brussels, Belgium
[6]Médecins Sans Frontières, Luxembourg Operational Research Unit (LuxOR), Luxembourg, Luxembourg
[7]Computational Epidemiology Lab, Boston Children's Hospital, Boston, USA
[8]Harvard Medical School, Boston, Massachusetts, USA
[9]WHO Hub for Pandemic and Epidemic Intelligence, Berlin, Germany
[10]Research Software Engineering Group, University of Oxford, Oxford, England, UK
[11]World Health Organization, Geneva, Switzerland
[12]Independent Researcher, Limassol, Cyprus
[13]Department of Community Health Sciences, Boston University School of Public Health, Boston, Massachusetts, USA
[14]Institute for Experiential AI, Northeastern University, Boston, Massachusetts, USA
[15]Department of Public Health and Health Sciences, Northeastern University, Boston, USA
[16]Santa Fe Institute, Santa Fe, New Mexico, USA
[17]Technical University of Lodz, Lodz, Poland
[18]Institut Pasteur de Dakar, Dakar, Senegal

v2 **First published:** 22 Sep 2025, **10**:524
https://doi.org/10.12688/wellcomeopenres.24776.1
Latest published: 08 May 2026, **10**:524
https://doi.org/10.12688/wellcomeopenres.24776.2

Abstract

Background

Early-phase data during an epidemic are often heterogeneous and difficult to integrate across systems, therefore a need for standard tools and reporting guidelines to facilitate timely and reliable data collection. The Global.health team have developed a data schema for

Open Peer Review

Approval Status ? ? ✓

	1	2	3
version 2 (revision) 08 May 2026			✓ view
version 1 22 Sep 2025	? view	? view	

the ingestion of epidemic data, allowing interoperability where data curated to this schema are readily ingested into existing systems for analysis. This paper describes the definition of 'core data' within the Global.health schema to focus data collection on the most relevant and available data to inform epidemic response during the first 100 days of an outbreak.

Methods

We used expert consultation and a structured literature review to identify key epidemiological questions and parameters that must be addressed during the first 100 days of an outbreak. Relevant digital toolkits and reporting frameworks were reviewed, and minimum data variables required for parameter estimation were identified. These variables were mapped to the existing Global.health schema and assessed for availability in early outbreak data from four recent epidemics. Variables were categorized by availability and those with sufficient early availability were retained in a proposed core schema. Data formats were harmonized with WHO Epi Core, T0 and T1 toolkits to enhance interoperability. A complementary modular schema was defined to capture pathogen-specific variables.

Results

The literature review yielded 78 key epidemiological parameters relevant to early outbreak assessment, organized into eleven categories. Analysis of variable availability in early outbreak datasets showed that 42 of 140 variables in the existing Global.health schema were consistently available and suitable for inclusion in a core early-epidemic schema. Variables related to demographics, case status, symptom reporting, confirmation dates, outcomes, and exposure history were frequently available, while vaccination history, detailed treatment data, and certain clinical variables were less consistently reported. The resulting core schema comprises 42 interoperable variables across seven domains and aligns with WHO data standards and controlled terminologies.

Conclusions

Standardized, interoperable data capture during the early phase of epidemics is essential to enable timely estimation of key epidemiological parameters and to inform response strategies. The Global.health core schema provides a minimum, evidence-informed dataset for early outbreak investigation while maintaining compatibility with WHO reporting standards. By prioritizing variables that are both epidemiologically critical and realistically available in early data streams, this framework supports improved data

1. **Juan Torres Munguía**, Georg-August-Universität Göttingen, Göttingen, Germany
2. **Geneva Tamunobarafiri Igwama**, University of Akron, Akron, USA
3. **Lilia Perfeito**, LIP–Laboratory for Instrumentation and Experimental Particle Physics, Lisboa, Portugal

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

harmonization, analysis, and decision-making during the first 100 days of an epidemic.

Keywords
Epidemic, Outbreak, Data Schema, Toolkits, Reporting Guidelines

Corresponding author: Everlyn Kamau (everlyn.kamau@ucsf.edu)

Author roles: Kamau E: Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kelly S**: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Resources, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Darji D**: Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Baidjoe AY**: Investigation, Writing – Review & Editing; **Brownstein JS**: Investigation, Writing – Review & Editing; **Campbell F**: Investigation, Writing – Review & Editing; **Dasgupta A**: Investigation, Resources, Software, Writing – Review & Editing; **Degail MA**: Investigation, Validation, Writing – Review & Editing; **Demidova A**: Investigation, Resources, Writing – Review & Editing; **Ferretti L**: Investigation, Writing – Review & Editing; **Han A**: Data Curation, Formal Analysis, Investigation, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Koyie SL**: Investigation, Writing – Review & Editing; **Ngamala PN**: Investigation, Writing – Review & Editing; **Polain OL**: Investigation, Writing – Review & Editing; **Rojek A**: Formal Analysis, Investigation, Validation, Writing – Review & Editing; **Sauer J**: Data Curation, Formal Analysis, Investigation, Writing – Review & Editing; **Scarpino SV**: Investigation, Writing – Review & Editing; **Sewalk K**: Data Curation, Investigation, Writing – Review & Editing; **Sopko J**: Data Curation, Formal Analysis, Investigation, Software, Validation, Writing – Review & Editing; **Zakrzewski S**: Data Curation, Investigation, Writing – Review & Editing; **Merson L**: Formal Analysis, Funding Acquisition, Investigation, Resources, Supervision, Writing – Review & Editing; **Kraemer MUG**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [225288; 226052; 228186 to M.U.G.K.]; M.U.G.K. acknowledges funding from The Rockefeller Foundation, Google.org, the Oxford Martin School Programmes in Pandemic Genomics & Digital Pandemic Preparedness, European Union's Horizon Europe programme projects MOOD [#874850] and E4Warning [#101086640], the John Fell Fund, a Branco Weiss Fellowship, United Kingdom Research and Innovation [#APP8583] and the Medical Research Foundation [MRF-RG-ICCH-2022-100069]. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the funders.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Kamau E, Kelly S, Darji D *et al.* **Unified framework for the ingestion of early epidemic data for downstream data analytics [version 2; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2026, **10**:524 https://doi.org/10.12688/wellcomeopenres.24776.2

First published: 22 Sep 2025, **10**:524 https://doi.org/10.12688/wellcomeopenres.24776.1

REVISED **Amendments from Version 1**

We have revised the introduction to clarify the goals and objectives of the paper. We also clarified the concept of data interoperability and have reorganized the section to make it clearer.

Methods

We have clarified about the workflow and methodology used in this paper including the how the literature review was conducted and how the expert consensus process was evaluated.

Results and Discussion

We revised the results to highlight the main components of the Global.health schema and to clarify on the relevance of the schema in outbreak settings. We have clarified on the limitations of data sources without English as the primary language. We have also highlighted the potential of other factors e.g. geographical disparities to influence the quality and harmonization of epidemiological data.

Any further responses from the reviewers can be found at the end of the article

Introduction

The first weeks of an epidemic are crucial for mounting an effective response given the potential health and socioeconomic impacts of epidemics.[1] During this period, it is essential to obtain reliable estimates of key epidemiological parameters such as transmission rates and disease severity, both for known and newly emerging pathogens. This is also a critical time for official reporting systems to assess availability of response resources, evaluate clinical impact, plan rapid testing and interventions (including assessing their effectiveness), monitor disease progression and estimate potential burden on health care systems and society. However, in the early stages of an epidemic, data on infectious disease cases and associated metadata – such as case and death counts, locations, sources of infection, laboratory tests performed, and vaccination status – are often sparse and heterogeneously captured, thus hindering data integration, analysis and interpretation.

These limitations hinder the utility of available data in public health decision making, despite the importance of the information they hold.[2] Heterogeneity can result from variation in content, quality, volume, format, veracity, completeness, definition and management processes.[3] Additionally, a lack of standardization in data capture systems delays data integration for research and statistical analysis, which impedes understanding and forecasting of the trajectory of an epidemic.[4] This limits real-time use of data to effectively assess and optimize public health operations during epidemic response.[5]

Another key issue in understanding disease dynamics is data access, which is limited by privacy concerns, legal and regulatory restrictions, conflicts of interest, complex processes to arrange data sharing agreements between entities, and the absence of trust, particularly in the digital era.[2,6] Socio-technical factors such as language barriers, differences between regional and national structures and rules that govern the dissemination of health information, imbalances in technical capacities, and power dynamics also impact the establishment of effective and meaningful data sharing.[7]

The Global.health platform assembles and curates open-access emerging infectious disease data to support situational awareness and risk assessments for decision-makers, researchers and the public.[8] Global.health aims to make data sharing more efficient and data more openly accessible to communities and groups contributing to epidemic response. A central step towards this aim is data interoperability, which enables seamless access, exchange, integration, and coordinated use of data across different systems, applications, and organizations, while preserving meaning, integrity, and usefulness. Global.health advances interoperability through the development and implementation of a standardized data schema, designed to ensure that epidemiological data are collected or curated in a uniform format that supports rapid analysis and estimation of key epidemiological parameters. This schema, known as the Global.health Day 0 schema (available at https://github.com/globaldothealth/outbreak-schema) has been applied to standardize epidemiological data from six previous outbreaks (https://global.health). Its implementation has exposed a critical gap in the continuum of data available during epidemics, which is most pronounced in the early phase of an outbreak – defined as the first 100 days – when timely information is essential and the opportunity for containment is greatest. We hypothesize that extending and refining the Day 0 schema to better capture data relevant to early outbreak assessment and response will strengthen its utility as a tool for supporting decision-making during the initial stages of epidemics.

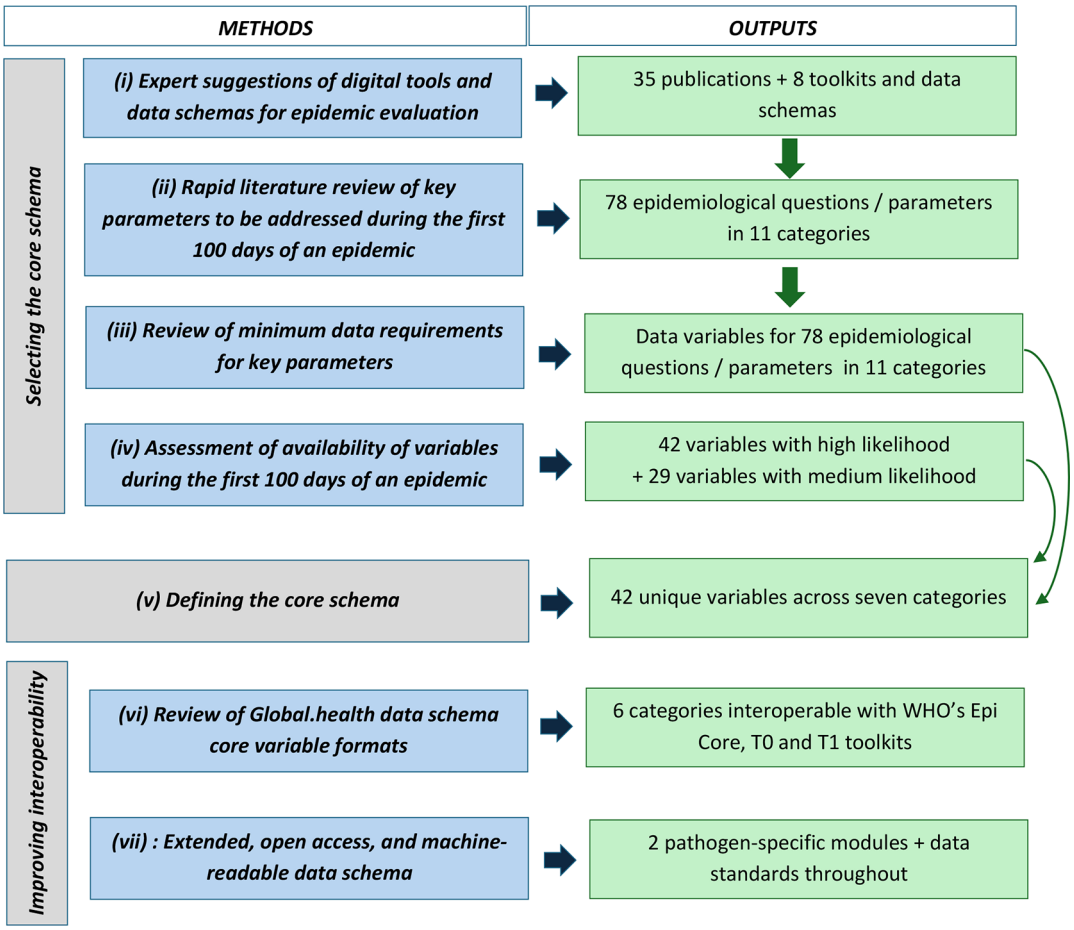

Figure 1. Conceptual framework of this study. Schematic illustrating the analyses and steps involved.

The purpose of the work described in this paper was to define the content of an early epidemic data schema to support timely, interoperable data collection and analysis during the initial phase of an outbreak. Central to the development of such a schema is the identification of key questions and parameters that are most important to estimate early during an epidemic. Here, we describe how these questions and parameters were determined through background research, literature review and expert interviews, and how they informed the development of the Global.health early epidemic data schema. We also describe the alignment of the Global.health data schema[9] with existing World Health Organization (WHO) toolkits (T0 and T1 forms) and define a minimum set of data required for epidemic investigation, particularly during the early phase (first 100 days) of an infectious disease outbreak. Ultimately, this work aims to ensure that epidemiological data curated to the schema can be shared in a standardized format that is readily ingested by existing and future analytical systems for epidemic analytics.

Methods

Our workflow and qualitative analysis were carried in multiple steps which are illustrated in Figure 1 and are described as follows:

(i) Selecting the core schema: Expert suggestions of digital tools and data schemas for epidemic evaluation. We first invited experts to identify existing toolkits and data schemas for epidemic data collection. A group of 22 scientists and academic researchers engaged in disease modelling, epidemiologists in national and international public health agencies (including three WHO offices) and outbreak specialists at humanitarian organizations were invited to share titles and links to relevant resources. Their suggestions included tools in the peer reviewed literature as well as additional resources from the grey literature.

(ii) Selecting the core schema: Rapid literature review of key parameters to be addressed during the first 100 days of an epidemic. Next, we conducted a literature review to identify key epidemiological parameters or questions that must be addressed during the first 100 days of an infectious disease epidemic. We searched the PubMed database was searched using the following terms: (((((epidemiology) OR (epidemiological)) AND ((((((outbreak) OR (outbreaks)) OR (disease outbreak)) OR (epidemic)) OR (epidemics)) OR (pandemic*))) AND ((((parameter*) OR (question*)) OR (data)) OR (information))) AND (((early) OR (initial)) OR (priority))) AND ((((((((((incubation period) OR (case fatality rate)) OR (case fatality rate (CFR))) OR (risk factors)) OR (basic reproduction number)) OR (R_0)) OR (effective reproduction number)) OR (serial interval)) OR (delay distributions)) OR (generation time)). The search was conducted on 23rd June 2024. The first 200 English language articles ranked by PubMed as most relevant to the search terms were combined with the resources identified through expert suggestion and deduplicated. The titles of all content were reviewed by two independent reviewers to identify those relevant to early outbreak response. Discrepant choices were resolved via discussion between the two reviewers. The full text of each selected title was reviewed to identify those with relevant data questions and parameters. Those selected by at least one reviewer were used for data extraction. Extracted content was categorized into types of parameters.

(iii) Selecting the core schema: Review of minimum data requirements for key parameters. The categorized parameters and questions extracted from the toolkits and literature in (i) and (ii) above were evaluated to identify the data variables required for their calculation or estimation. Variables were listed per parameter, then grouped using categories defined in prior work by Perrocheau et al.[21] Ten of the 22 individuals who participated in the expert suggestions provided iterative review of the resulting list to agree on the completeness and suitability of the variables and categories for addressing questions during the early phase of a public health response. Iteration resulted in consensus on a draft list of variables. The list was reviewed against the data reporting requirements of the International Health Regulations (IHR, 2005) to ensure that the variables covered the IHR reporting requirements.

(iv) Selecting the core schema: Assessment of availability of variables during the first 100 days of an epidemic. Variables defined in (iii) were identified, were available, within the Global.health day 0 schema and assessed for availability based on their presence or absence from previous Global.health data sources[8] during the first 100 days of an epidemic. Availability was determined as the count or frequency of records containing a variable of interest within 100 days of the first identified case. Data from epidemics including Ebola (2018–2020), Marburg (2024), COVID-19 (2019–2023) and Mpox (2023–2024) were assessed. Each variable was assigned a rating of 'high', 'medium', or 'low', indicating the mean availability of the variable in early epidemic data as ≥80%, ≥50% to <80% and < 50%, respectively. Those with low availability were excluded from the early outbreak variable selection.

(v) Defining the core schema. The outputs of (iii) and (iv) above were used to define a subset of the Global.health data schema as a 'core' schema for early outbreak assessment and response. First, the Global.health variables that matched the minimum data variables for key parameters were selected. This list was then reduced to include only those variables evaluated as high or medium availability in early epidemic data.

(vi) Improving interoperability: Review of Global.health data schema core variable formats. We compared the format of the variables included in the newly defined Global.health core data schema with the formats used in three of the key existing toolkits identified amongst the expert suggestions. Global.health formats were adjusted to align with WHO's Epi Core, T0 and T1 formats to ensure that data will be interoperable across these tools.

(vii) Improving interoperability: Extended, open access, and machine-readable data schema. Additional modular schemas were subsequently defined using the WHO T0/T1 and disease-specific toolkits identified in (i) above. The additional variables were selected to capture pathogen specific detail that were missing from the core schema. This group of variables were defined as a 'modular' schema and includes variables related to exposure, treatments and vaccination. The format of all data variables was reviewed against the data standards in use by WHO. Appropriate controlled terminologies were applied as relevant.

Results

Data schemas for epidemic evaluation

Experts identified 35 publications as relevant to the project aims and eight unique toolkits and data schemas (see Table 1). Six of the eight resources were designed to ingest data from epidemic field investigations.

Table 1. Toolkits. Summary of existing toolkits and digital resources for infectious disease data and information collection (WHO: World Health Organization; CDC: Centers for Disease Control).

Digital tool	Brief description of purpose, source and design
WHO T0 case investigation form[19]	The initial generic form designed to help outbreak field investigators rapidly understand an epidemic and propose initial control measures. It supports the collection of the minimum Epi Core[20] variables for epidemic investigation,[21] and was designed to help describe the epidemic over time, geographical spread and persons affected to guide decisions regarding the first measures to control the epidemic. Data collected from the WHO T0 case investigation form[19] can also be used to generate hypotheses about the case, source, and transmission mode of a pathogen. The additional variables collected on the T0 form outside of the specified Epi Core variables were collated from the most common variables indicated on the WHO Disease outbreak toolkits,[22] which were created to enable rapid specification of a case definition, collection of appropriate data for the disease and for field workers to have a resource for tools and training.
WHO T1 case investigation form for outbreaks of unknown cause[19]	Used to collect detailed information on any situation involving an epidemic of unknown origin. It is designed for initial data collection to identify the epidemic source, mode of transmission or agent involved, and is composed of 462 variables organized in five categories, many of which are conditional or symptom-based checklists. The clinical variables in the WHO T1 form are grouped per functional body system (e.g., neurological, digestive, cutaneous), which is preferred as opposed to the syndromic approach when the disease is unidentified.[19] The exposure variables in the WHO T1 form support the investigation of risks related to transmission, food or water contamination with pathogens and environmental toxins and hazards (e.g., heavy metals) that can lead to an outbreak.
The European Surveillance System (TESSy)	The TESSy portal facilitates collection, analysis, and dissemination of indicator- and event-based surveillance data in infectious disease and associated health issues. The TESSy metadata set lists the collected data for a wide range of pathogens, from both aggregate and case-based data. The European Surveillance System (TESSy), is now integrated into EpiPulse with the five Epidemic Intelligence Information System (EPIS) platforms and the Threat Tracking Tool (TTT).[23–26] TESSy's standardised data formats and validation rules are provided within the metadata to improve data quality, but the rationale for the collection of each variable is not provided. Data sources include those that are open access such as FluTrackers, monitoring websites and trusted media sites, as well as WHO sources and other access-restricted sources. Furthermore, access to EpiPulse is restricted.
The article 'Key data for outbreak evaluation: building on the Ebola experience'[27]	This draws from the authors prior experiences in outbreak data collection and analysis and provides a checklist of data needed to quantify severity and transmissibility, characterize heterogeneities in transmission and their determinants, and assess the effectiveness of different interventions. The article recommends that data are differentiated into individual-level data, exposure data, and population-level data, and the checklist also highlights the potential issues and biases that can be present in these data. The setting for data capture in the article is also primarily field investigations, like the WHO toolkits.

Table 1. *Continued*

Digital tool	Brief description of purpose, source and design
WHO PISA (Pandemic Influenza Severity Assessment)	PISA details the requirements for influenza surveillance and how to assess the impact of a potential outbreak.[28] It focuses on assessment of (i) transmissibility using case incidence and positive test rates, (ii) severity using deaths and hospitalizations, and (iii) impact using incidence, excess pneumonia/influenza or all-cause mortality, confirmed cases, hospital admissions, absenteeism in school and workplaces, and healthcare resources use such as beds occupied.
CDC framework for assessing epidemiological effects of influenza epidemics and pandemics[29]	Similar to the WHO's PISA risk assessment on transmissibility and severity. Transmissibility is assessed using R_0, the serial interval and the attack rate, while severity is assessed from case fatality, hospitalization records and genetic markers of virulence. This document also lists the strengths and limitations of each parameter and provides considerations towards evaluation of data quality. The WHO PISA suggests that parameters should be reviewed by age groups, and severity should consider presence or absence of underlying chronic diseases.
Key indicators for WHO priority pathogens with a confirmed outbreak	WHO's key epidemiological indicators and recommended analyses for surveillance of priority pathogens with a confirmed outbreak have been published for various diseases including cholera,[30] measles, yellow fever and meningitis.[31,32] They focus on incidence, attack rate, severity (case fatality) and test positivity. Only for cholera have the indicators been linked to specific data points and rationale provided.
Global.health day 0 core schema	The Global.health day 0 core schema[9] uses a common data model to standardize key epidemiological epidemic data that includes details for case status, location, age, sex, symptoms, transmission, hospitalization, treatment, vaccination, outcome, contacts, travel history, occupation, genomics and more. Each individual case is assigned a unique identification number when added to a line list and data are de-identified to protect individual privacy. The Global.health data sources are primarily those that are publicly available. Originally designed in the context of COVID-19 data curation[33] and most recently updated for mpox reporting (covering the 2022 global epidemic) where it was based heavily on the WHO mpox case report form. It collects data and metadata from official and non-official online public sources and provides the original source(s) of information for each case to support transparency and data sharing. These variables are specified by minimum case requirements and are pathogen-agnostic.

Key parameters to be addressed during the first 100 days of an epidemic

Our search terms identified 142,563 potentially relevant articles. The 200 titles ranked by PubMed as most relevant were combined with the 35 expert identified publications. Following removal of duplicates, the titles and abstracts from the remaining 235 were screened to identify 63 of potential relevance to the study. Full text review of these resources resulted in data extraction from 35 found to contain parameters relevant to understanding the first 100 days of an outbreak (Figure 2). We identified and extracted 78 unique epidemiological questions or parameters from the selected literature. Parameters were categorized into eleven categories as described in Table 2.

Minimum data reporting requirements for key parameters

Data variables identified by the project team and expert reviewers for each key parameter and question are listed in Table 2. Variables included in the table were found to cover all IHR reporting requirements. The final variables were considered those important to early epidemic investigation.

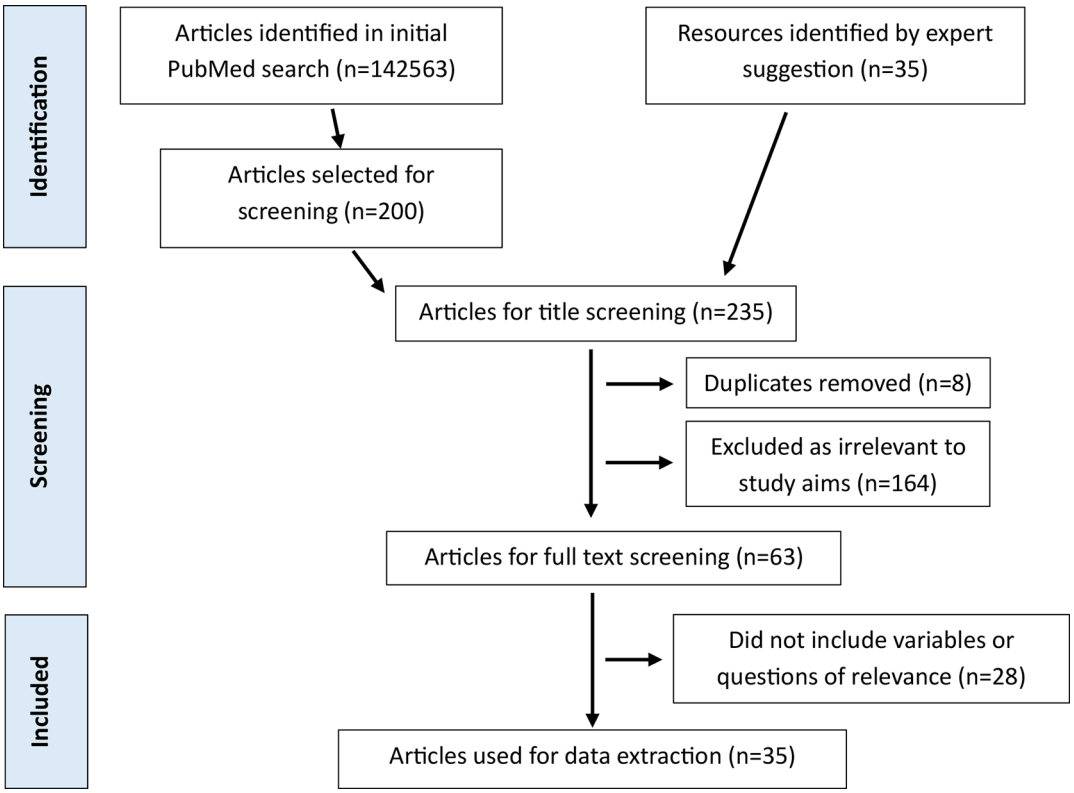

Figure 2. Literature search (PRISMA diagram of scoping review). Literature search workflow to identify and extract key questions and parameters analyzed to guide public health response during the early stages of an epidemic or outbreak.

Table 2. Key categories and parameters. A summary of categories and parameters or variables considered important for infectious disease epidemic investigation.

Category	Description and some of the parameters included
Epidemiological parameters	Used to model transmission and predict epidemic growth rate. Include basic reproduction number, incidence rate, effective reproduction number, incubation period, serial interval, generation time and secondary attack rate.
Delay distributions	Describe delays between events in the infection history and reporting, which affect transmission and can have implications for epidemic control.
Disease severity factors	Include case fatality rate, hospital and intensive care unit admission rates.
Risk factors for epidemic growth, susceptibility to infection and poor outcomes	Include individual level demographic factors, medical factors (e.g., pre-existing conditions, vaccination history, reinfections), occupational factors (e.g., health care worker status or other risks from specified occupations depending on the pathogen), behavioral factors (e.g., details of exposure and travel history to determine settings and risk factors of transmission).
Where/Who factors	Important for inferring transmission heterogeneities or risk of infection across different locations or population groups.
Clinical features	Case defining or prognostic features, laboratory results and vital signs.

Table 2. *Continued*

Category	Description and some of the parameters included
Contextual and external factors	Broader population-level indicators for surveillance bias (how an infected case was identified to the healthcare system), impact of human mobility patterns, climate and other drivers of transmission, as well as knowledge of pathogen genomic lineage for detailed analyses.
Diagnostic factors	Knowledge of testing methods and strategies across different locations to interpret incidence and other epidemiological aspects listed above.
Surveillance parameters	High-level estimates of the proportion of infections reported, asymptomatic cases as well as peak-time forecasts of incidence, useful to model the impact of the epidemic. These could be a function of, or derived from, the contextual and external factors.
Impact of pharmaceutical interventions	Medical treatments and vaccinations, and their impact on incidence and individual outcomes.
Non-pharmaceutical interventions	Non-pharmaceutical interventions, or "Public Health and Social Measures",[34] such as isolation procedures, and other behavioral interventions (e.g. hand-washing, mask-wearing) and their impact on incidence and individual outcomes.

Availability of variables during the first 100 days of an epidemic

Of the current 140 variables in the Global.health data schema, 42 had a high likelihood of being captured in a line list in early epidemic data (**Supplementary Table S1**). These variables included parameters that describe pathogen name, case status, location of the case report, list of symptoms, date of case report, outcome, date of case confirmation, date of outcome evaluation, occupation, where contact with infected or suspected cases occurred, travel history and sources of information. Where travel history is mentioned, the travel location is indicated.

There were 29 variables with a medium likelihood of availability in epidemic data sources including pre-existing conditions, vaccination information, date of symptom onset, pathogen strain subtype, diagnostic test, contact with animal or insect bites, date of entry into the country, type of medical treatment and exposure to water or chemical agents (**Supplementary Table S1**). Other variables (n = 56) were rated as low likelihood of availability. These included confirmed prior infection, co-infection with other pathogens, number of vaccine doses, date of first vaccination, admission to intensive care unit, date of isolation, vital measurement results, dates of medical treatment, type of specimen and mode of travel.

Evaluation of the Global.health data also had a low availability of data on whether an individual visited a healthcare facility several days prior to symptom onset, or how a case was found (which might help to assess surveillance bias), as well as reliably tell the date of the healthcare visit. The reports cannot reliably differentiate between primary and secondary cases within contact exposure, but where death and treatment are mentioned, the date of death and type of treatment are often indicated. Other things to note: (i) vaccination information is available depending on the causative pathogen and hence the low likelihood of availability as noted above, (ii) 'death' is the most commonly reported outcome while recovery or date of recovery is not always mentioned, (iii) where symptoms may not be mentioned, an asymptomatic case might be assumed, (iv) home monitoring may not be distinguished from the general 'isolation' variable, and (v) pre-existing conditions are reported more frequently than previous infections. The current Global.health schema variables on mass gatherings were specific to COVID-19 and therefore unlikely to be available for other epidemics.

The core schema

Selection of the data variables for key parameters and reduction of this list to those with high or medium availability resulted in a list of 42 unique variables across seven categories. These variables, featured in Table 3, comprise the Global.health core schema.

Table 3. Minimum data requirements**.** This table highlights variables or information that would be considered the minimum requirements for various categories and parameters.

Category	Parameter(s)	Minimum variables or information required
Epidemiological parameters	Incubation period[a]	Date of symptom onset Exposure date or date of last contact to the suspected source (e.g., animal, infectious individual, toxin, etc.) Travel dates (e.g., exposure interval at the source and symptom onset in other location)
	Basic reproduction number (R0)[b] Effective reproduction number (Rt)[c]	Date of symptom onset of confirmed cases Date of case confirmation (Additional information on vaccination is recommended for *Rt* where suitable)
	Incidence rate[d], attack rate[e], growth rate[f]	Case status (e.g., confirmed, probable, suspected) Case location Date of case report
	Secondary attack rate	Date of symptom onset in *infector* Date of symptom onset in *infected*
	Serial interval distribution[g]	Household size (Other supplementary data elements include travel history location(s) and setting where contact occurred with the confirmed/suspected case)
Delay distributions	Measures of delays between events in the infection history and reporting (e.g., Date of symptom onset to date of reporting)	Date of symptom onset Date of case report Date of isolation Date of hospitalization Date of hospital discharge Outcome (e.g., hospitalization, death, recovery) Date of outcome evaluation Date of specimen collection Date of case confirmation
Disease severity factors	Measure of severity e.g., Infection fatality ratio/Infection fatality rate, case-hospitalization ratio	Case status (e.g., confirmed, probable, suspected) Disease outcome (hospitalization, death, recovery) Date of outcome evaluation Age Race and/or ethnicity Clinical vulnerability Occupation (e.g., healthcare worker) Date of case report Intensive care treatment
Risk factors for epidemic growth, susceptibility to infection and individual outcomes	Demographic, medical and occupation risk factors	Travel history location(s) Date of country entry Date of travel start Date of symptom onset Age Sex at birth Date of hospitalization Pre-existing conditions including pregnancy Presenting symptoms Presenting treatments Vaccination (including timing) Occupation (e.g., healthcare worker) Contact status with infected or suspected case

Table 3. *Continued*

Category	Parameter(s)	Minimum variables or information required
		Close proximity with animals or insect bites prior to symptom onset Exposure to harmful water or chemicals Visiting a healthcare facility days prior to symptom onset Infection history (previous similar infection)
Diagnostic factors	Diagnostic performance	Diagnostic test Results from negative and positive control samples Type of specimen
Impact of pharmaceutical interventions	Impact on incidence and individual outcome	Medical treatments (including timing) Vaccination (including timing) Outcome (e.g., hospitalization, death, recovery)
Impact of non-pharmaceutical interventions	Impact on incidence and individual outcome	Type of interventions and date Date of symptom onset Date of isolation Outcome (e.g., hospitalization, death, recovery) Time series of suspected, probable or confirmed cases (population level)

[a]The interval between exposure and initial occurrence of signs and symptoms.
[b]The expected number of cases generated by a single case in a population where all individuals are susceptible to infection.
[c]The number of cases generated in the current state of a population.
[d]The number of new cases in a population within a specified period of time.
[e]The proportion of an at-risk population that contracts the disease during a specified time interval.
[f]How quickly the numbers of infections are changing over a period of time.
[g]The interval between symptom onset in an index case and in a secondary case.

Improving interoperability: Review of Global.health data schema core variable formats

The structure of the core schema was aligned with the structure of the WHO's Epi Core, T0 and T1 toolkits, ensuring that variables in both systems are aligned and data are captured similarly. The categories were structured as follows:

I. *Demographics*: includes variables with responses such as age, gender, sex at birth, case status (e.g., confirmed, suspected, probable), locality or residence location and job occupation. The residence query results in a location structure that is pre-filled by mapbox with latitude and longitude, ISO code for country and geographical administrative information.

II. *Medical History*: this category contains variables on pre-existing conditions including pregnancy, co-infections, history of the same infection prior to the current diagnosis and vaccination information.

III. *Clinical Presentation*: includes variables on reported symptoms, date of symptom onset, hospitalization, intensive care treatment, outcome of illness (e.g., death, recovered, post-acute sequelae), vital measurement results, pharmaceutical and non-pharmaceutical treatments.

IV. *Laboratory Information*: including genomic information, testing information (e.g., type of specimen, method of testing date of sample collection, pathogen strain/subtype) and date of case confirmation.

V. *Exposure*: contains variables indicating likely sources of exposure e.g., contact with suspected or confirmed cases or infected animals/animal products, or through environmental exposure (water sources, chemicals, exposure).

VI. *Source Information*: contains variables on data origin (e.g., government bulletins, ministry of health reports, media platforms, etc.), date of entry and/or modification and curator's details and comments.

Table 4. Exposure and intervention modules. Each column represents the two Global.health schema modules. In each column we show the representative data elements and examples of specific variables related to these elements. Currently, the intervention module has only three data elements.

Exposure module	Intervention module
Contact with case (e.g., contact ID, contact setting, date of last contact)	Pharmaceutical interventions – vaccinations (e.g., number of doses, vaccine name, vaccination date, side effects)
Mass gathering (e.g., mass type, date of event, location of event)	Pharmaceutical interventions – treatments (e.g., type and name of treatment, route, start and end date, daily dose, traditional treatment)
Animal contact (e.g., animal species, date and location of contact with animal, insect bites or stings, contact with skinned wild game, raw animal meat of blood)	Non-pharmaceutical interventions (e.g., face mask, social distancing, hand washing, school closure)
Travel history (e.g., date of entry into country, location of travel, mode of travel)	
Treatment (e.g., visit to healthcare facility, type and location of facility, visit to a traditional healer)	
Water source (e.g., type of drinking water source, contact with flood water)	
Chemical source (e.g., potential source of chemical exposure, place and duration of exposure, suspected chemical product)	

Several variable responses in the Global.health schema e.g., symptoms, specimen type, pre-existing medical conditions and type of medical treatment can be selected from predefined lists with an option for free text entry.

Improving interoperability: Extended open access and machine-readable data schema

Based on the results above, we updated the Global.health 'core' schema to capture variables relevant to all epidemic types and added a 'modular' schema. The latter contains two modules with variables related to exposure and interventions e.g., treatments and vaccination, which can be adjusted to become pathogen specific (Table 4). The variables in both schema setups were reviewed against existing toolkits in terms of data type, question text, and response options to ensure consistency and interoperability with other data capture specifications. The schemas specify syntactic and semantic standard codes for each variable according to the WHO SMART guidelines,[10] to allow for interoperability of shared data. These codes were prepared from standardised dictionaries including SNOMED-CT (International),[11] SNOMED-GPS (Global Patient Set),[12] LOINC,[13] ICD-11,[14] ICHI[15] and ICF,[16] and align with existing codes used in other WHO digital adaptation kits for interoperability within WHO systems. For each variable, multiple codes were specified where applicable to allow better interoperability among users or systems working with the resulting curated dataset. Collectively, the core and modular schema form a digital adaptation kit, which as part of the WHO SMART guidelines can be adapted to contain additional variables capturing exposure types, treatments, and vaccinations.

Discussion

Infectious disease surveillance requires timely and reliable data ingestion processes which are often complex and fragmented across institutions and governmental entities. Analytical methods, tools and resources for epidemic investigation have grown significantly over the last two decades, however, data acquisition and interoperability have remained a challenge and are difficult to standardize across pathogens and diseases. We have developed a unified data ingestion framework for descriptive mapping of variables collected during an epidemic response, which has a wide scope of parameters and variables preset with a uniform coding system for optimal interoperability. We examined the data collection variables in the Global.health data schema to assess interoperability with WHO systems and ensure that the output of data curated to the Global.health schema can be shared in a standard format readily ingested to support key epidemiological tasks identified by the WHO.

Currently, the occupational risk factors in the Global.health schema will differ depending on the type of epidemic, but it would be desirable to have agnostic occupation categories applicable to all epidemics. In our assessment of data availability for the minimum reporting requirements, we did not evaluate the likelihood of availability by geographical origin. This would be useful to identify gaps in data collection methods or surveillance system structures that could potentially challenge harmonizing surveillance data across countries.

The low availability of some data variables is likely to impact epidemic and outbreak response and epidemiological parameter estimation. For instance, inadequate information on vaccination might hinder evaluation of coverage and impact immunization services on disease burden. Similarly, inadequate information on hospital discharge could hinder predictive models for estimating the discharge probability of acute care patients. We found that the date of specimen collection was lacking, which would also impact understanding of when a disease process was present in a patient. The ongoing efforts for improved data collection should emphasize such variables, and the several others as they play an important role in monitoring disease response and management.

In future work, we will further assess whether the key parameters identified here can be adapted to the field resources available during an epidemic investigation through a user survey, the outcome of which will be used to update the core Global.health schema and the minimum data requirements for epidemic evaluation. This will lead to improved data acquisition and better reporting guidelines during epidemic response. We also acknowledge that data representation might vary by location and that future work should focus on developing specific guidelines by region for improved early data availability and accessibility.

While we acknowledge the significance of such a framework, we recognize its limitations. Firstly, despite our efforts, the Global.health core schema may not be comprehensive with respect to all pathogen groups as 'Disease X' may present with unrepresented risk factors or clinical characteristics. Second, the schema cannot ensure fidelity of the data collected. Third, currently the schema is in English, which may challenge its accessibility for non-native speakers. Fourth, depending on the dynamics and nature of an epidemic, national or governmental policies may not always support external data collection products. Barriers may and evolving data formats might occur as the epidemic develops, which could impact the scale and quality of the data. For example, the first hundred cases might have detailed information (or vice versa, depending on the response team's readiness), but less available with increasing number of cases.[17] Or, as the epidemic grows, public health agencies might discontinue reporting individual-level case data and instead switch to reporting total numbers (or estimates thereof) of confirmed or suspected cases.[17] For this, we may consider developing other augmented data ingestion frameworks.

It is also worth noting the initial challenge users might initially encounter adapting to datasets containing coded information (SNOMED, LOINC, etc.) rather than the usual simple database coding and text data format. These limitations notwithstanding, the updated Global.health schema has already been used in collaboration with national and international health authorities during public health emergencies, demonstrating practical feasibility in field settings. The digitization of WHO T0/T1 toolkits has also benefited the use of standardized data schema. We hope our work will encourage relevant and directed data capture across the infectious disease research community, ensuring standardized and efficient data collection for timely and informative decisions around appropriate public health responses.

Ethics and consent
Ethical approval and consent were not required.

Data availability
No underlying data were associated with this study.

Extended data
OSF: Unified framework for the ingestion of early epidemic data for downstream data analytics. Dataset. DOI 10.17605/OSF.IO/AKQ53[18]

This project contains the following extended data:

- Supplementary Table 1 Table of assessment and rating of the likelihood of data availability.

Acknowledgements
The authors acknowledge Jacqueline Powers for contributions in the early stages of the project.

References

1. Heymann DL, Chen L, Takemi K, *et al.*: **Global health security: the wider lessons from the west African Ebola virus disease epidemic.** *Lancet.* 2015; **385**: 1884–1901.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Lee ACK, Iversen BG, Lynes S, *et al.*: **The state of integrated disease surveillance in seven countries: a synthesis report.** *Public Health.* 2023; **225**: 141–146.
PubMed Abstract | Publisher Full Text

3. Quiroga Gutierrez AC, Lindegger DJ, Taji Heravi A, *et al.*: **Reproducibility and Scientific Integrity of Big Data Research in Urban Public Health and Digital Epidemiology: A Call to Action.** *Int J Environ Res Public Health.* 2023; **20**: 1473.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Fairchild G, Tasseff B, Khalsa H, *et al.*: **Epidemiological Data Challenges: Planning for a More Robust Future Through Data Standards.** *Front Public Health.* 2018; **6**: 336.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Kucharski AJ, Hodcroft EB, Kraemer MUG: **Sharing, synthesis and sustainability of data analysis for epidemic preparedness in Europe.** *Lancet Reg Health Eur.* 2021; **9**: 100215.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Fallatah DI, Adekola HA: **Digital epidemiology: harnessing big data for early detection and monitoring of viral outbreaks.** *Infect Prev Pract.* 2024; **6**: 100382.
Publisher Full Text

7. Liverani M, Teng S, Le MS, *et al.*: **Sharing public health data and information across borders: lessons from Southeast Asia.** *Global Health.* 2018; **14**: 94.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Global.health.
Reference Source

9. Global.health outbreak schema.
Reference Source

10. WHO: WHO SMART Guidelines.
Reference Source

11. SNOMED-CT.
Reference Source

12. SNOMED-GPS.
Reference Source

13. LOINC.
Reference Source

14. ICD-11.
Reference Source

15. International Classification of Health Interventions (ICHI).
Reference Source

16. International Classification of Functioning, Disability and Health
Reference Source

17. Xu B, Gutierrez B, Mekaru S, *et al.*: **Epidemiological data from the COVID-19 outbreak, real-time case information.** *Sci Data.* 2020; **7**: 106.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Kamau E: **Dryad: Unified framework for the ingestion of early epidemic data for downstream data analytics. Dataset.**
Publisher Full Text

19. WHO: Case investigation form for outbreaks of unknown cause (T1).

20. EpiCore.
Reference Source

21. Perrocheau A, Brindle H, Roberts C, *et al.*: **Data collection for outbreak investigations: process for defining a minimal data set using a Delphi approach.** *BMC Public Health.* 2021; **21**: 2269.
PubMed Abstract | Publisher Full Text | Free Full Text

22. WHO Disease outbreak toolboxes.
Reference Source

23. The European Surveillance System (TESSy).
Reference Source

24. EpiPulse.
Reference Source

25. Epidemic Intelligence Information System (EPIS) platforms.
Reference Source

26. Threat Tracking Tool (TTT).
Reference Source

27. Cori A, Donnelly CA, Dorigatti I, *et al.*: **Key data for outbreak evaluation: building on the Ebola experience.** *Philos Trans R Soc Lond B Biol Sci.* 2017; **372**: 20160371.
PubMed Abstract | Publisher Full Text | Free Full Text

28. WHO PISA (Pandemic Influenza Severity Assessment).
Reference Source

29. CDC framework for assessing epidemiological effects of influenza epidemics and pandemics.
Reference Source

30. WHO: Cholera.
Reference Source

31. WHO: Measles.
Reference Source

32. WHO: YFV.
Reference Source

33. Kraemer MUG, Scarpino SV, Marivate V, *et al.*: **Data curation during a pandemic and lessons learned from COVID-19.** *Nat Comput Sci.* 2021; **1**: 9–10.
PubMed Abstract | Publisher Full Text

34. Majeed A, Abbasi K: **Medical journals should use the term "public health and social measures".** *BMJ.* 2025; **388**: r409.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ? ✓

Version 2

Reviewer Report 08 June 2026

https://doi.org/10.21956/wellcomeopenres.29012.r155933

✓ **Lilia Perfeito**
LIP–Laboratory for Instrumentation and Experimental Particle Physics, Lisboa, Portugal

I was happy to review this paper as it addresses a key problem in epidemiology, particularly in new outbreaks, which is the difficulty in finding standardized data across multiple countries. Global.Health seems like a promising tool to do that, if it is adopted by health authorities.

I believe the clarifications made after version 1 have made the paper stronger and I do not have any other major comments; just 3 minor ones:
1 - If possible, it would be helpful for the community if the authors made available the 78 key parameters that they identified in the literature (as a supplementary table). Since only 42 were included, I am curious about what we may be missing out and whether in the future those may be included.
2 - In figure 2, the PRISMA diagram, the authors start with 35 expert-curated resources and end up with 35 articles used for data extraction. Is this just a coincidence or were these the same?
3 - in page 6 of the pdf, we can find the following excerpt:
“(iv) Selecting the core schema: Assessment of availability of variables during the first 100 days of an epidemic. Variables defined in (iii) were identified, were available, within the Global.health day 0 schema and assessed for availability based on”
Should the second “were” be “where”?

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

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

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

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

Not applicable

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

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Complex Systems

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.

Version 1

Reviewer Report 13 January 2026

https://doi.org/10.21956/wellcomeopenres.27296.r140042

?

Geneva Tamunobarafiri Igwama
University of Akron, Akron, Ohio, USA

This manuscript discusses a framework for gathering and standardizing early epidemic data to assist with epidemiological analysis. The authors introduce the Global. health data schema, which aims to enhance the compatibility and usability of varied outbreak data. They evaluate the framework against World Health Organization (WHO) tools and conduct a literature review to identify key epidemiological parameters relevant in the initial 100 days of an epidemic, alongside expert consultations and reviews of data from recent epidemics like Ebola, COVID-19, Mpox, and Marburg.

The authors compile 78 essential epidemiological parameters into 11 categories and define the minimum data necessary to estimate these parameters. They refine the Global. health schema into a "core" schema and adaptable modular components. The paper emphasizes that better standardization and prioritization of early outbreak data can improve real-time data analysis and public health decisions.

The manuscript presents relevant insights into public health, aligning well with international guidelines and demonstrating solid methodology through a clear workflow. It offers a practical, openly available schema already in use in some field settings. However, there are concerns regarding the lack of detail in the expert consultation process, which may hinder reproducibility.

The literature review could benefit from improved rigor, and availability assessments are noted to be somewhat subjective.

Major comments that require attention include improving transparency and rigor in expert selection, clarifying the literature review methods, and providing a validation example for the framework's practical benefits. Minor recommendations suggest addressing geographic disparities in data availability, discussing language accessibility, and improving the clarity of data tables.

The manuscript is a significant contribution to epidemic preparedness, but it needs major revisions for methodological transparency and validation before full acceptance.

References

1. Kamau E, Kelly S, Darji D, Baidjoe A, et al.: Unified framework for the ingestion of early epidemic data for downstream data analytics. *Wellcome Open Research*. 2025; **10**. Publisher Full Text

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

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

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

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

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

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Public health, digital health, Medicine, Nursing, Epidemology, 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.

Author Response 21 Mar 2026

Everlyn Kamau

Reviewer 2 This manuscript discusses a framework for gathering and standardizing early

epidemic data to assist with epidemiological analysis. The authors introduce the Global. health data schema, which aims to enhance the compatibility and usability of varied outbreak data. They evaluate the framework against World Health Organization (WHO) tools and conduct a literature review to identify key epidemiological parameters relevant in the initial 100 days of an epidemic, alongside expert consultations and reviews of data from recent epidemics like Ebola, COVID-19, Mpox, and Marburg.

The authors compile 78 essential epidemiological parameters into 11 categories and define the minimum data necessary to estimate these parameters. They refine the Global. health schema into a "core" schema and adaptable modular components. The paper emphasizes that better standardization and prioritization of early outbreak data can improve real-time data analysis and public health decisions.

The manuscript presents relevant insights into public health, aligning well with international guidelines and demonstrating solid methodology through a clear workflow. It offers a practical, openly available schema already in use in some field settings. However, there are concerns regarding the lack of detail in the expert consultation process, which may hinder reproducibility. The literature review could benefit from improved rigor, and availability assessments are noted to be somewhat subjective. **Thank you. The entire methods section has been rewritten to add detail, improve reproducibility and insure clarity of process.**
Major comments that require attention include improving transparency and rigor in expert selection, clarifying the literature review methods, and providing a validation example for the framework's practical benefits. Minor recommendations suggest addressing geographic disparities in data availability, discussing language accessibility, and improving the clarity of data tables. T**hank you. Disparities in data availability: - This was not a focus of this paper. We were interested in assessing the likelihood of data variables being present in the Global.health data sources. While geographic disparities may play a role in data availability, addressing that is beyond the scope of this work. We have highlighted this in the discussion for future work. Language accessibility: - While the schema is currently in English, accessibility in other languages is possible but not within scope of our current work. Clarity of data tables: - Thank you for identifying this lack of clarity. We have rewritten the results section to match the rewritten methods section. Tables are linked to specific stages of results to improve clarity of process.**

The manuscript is a significant contribution to epidemic preparedness, but it needs major revisions for methodological transparency and validation before full acceptance. **Thank you. We have added detail to the specific issues raised and look forward to further feedback.**

Competing Interests: No competing interests were disclosed.

Reviewer Report 11 November 2025

https://doi.org/10.21956/wellcomeopenres.27296.r135848

?

Juan Torres Munguía
Georg-August-Universität Göttingen, Göttingen, Germany

Dear Editorial Team, dear authors,

Thank you very much for the opportunity to review the paper entitled "Unified framework for the ingestion of early epidemic data for downstream data analytics." I would like to sincerely commend the authors for their effort and dedication. The topic addressed is highly relevant and impactful.

Overall, the paper presents a proposal for the collection of data during the early phase of a disease outbreak by establishing a unified schema that captures the most relevant indicators for decision-making. I believe the manuscript is suitable for indexing after a medium revision. In particular, I have the following comments:

Abstract
1. The goal of the paper is not mentioned. Please add it in the background subsection.

Methods
2. The description of the methods could be improved. Briefly describe Figure 1, indicating that it represents a methodology based on different steps.
3. The indication that the early phase refers to the first 100 days should be introduced earlier in the text, ideally at the first instance it is mentioned, to ensure clarity for the reader.

Introduction
4. Please add a paragraph about the impact of diseases and the relevance of studying them. For instance, include the number of deaths from COVID-19 using data from the WHO COVID-19 dashboard or the number of disease outbreaks worldwide from this paper: https://doi.org/10.1038/s41597-022-01797-2
5. Define the concept of data interoperability.
6. Add a paragraph describing the goal of the paper and its structure.

Methods
7. Figure 1 (Conceptual framework of this study) can be improved by presenting it as a logical pipeline, numbering and naming each stage (e.g., Stage 1: Identification of relevant variables, Stage 2: Identification of the minimum set of indicators, etc.). Then, describe these stages in the text using exactly the same names as in Figure 1.

Results

8. Elaborate the results according to the stages defined in the Methods section. Follow the same structure and add a framework showing the outputs for each stage and step.
9. In the paragraph "Literature review of key parameters to be addressed during the first 100 days of an epidemic", the authors use both "35" and "Thirty-five". Please standardize the format: use numerals for numbers 10 and above, and words for numbers zero through nine.
10. In the paragraph "Global.health data schema core variables", the authors use inconsistent formatting for category names (e.g., Laboratory Information vs. Medical history). Please standardize the formatting.
11. Include more details about the Mpox data ingestion in the Democratic Republic of Congo. What was learned from that experience? Please add a paragraph discussing the practical feasibility.

Discussion
12. The authors mention that the Global.health core schema may not be comprehensive for all pathogen groups. Please provide an example of a pathogen for which the proposed schema might be insufficient.
13. The manuscript mentions that language could be a barrier for non-native English speakers. How was this challenge addressed in the Democratic Republic of Congo?

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

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

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

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: statistics, health data, development and humanitarian analysis

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.

Author Response 21 Mar 2026

Everlyn Kamau

Reviewer 1 Dear Editorial Team, dear authors,

Thank you very much for the opportunity to review the paper entitled "Unified framework for the ingestion of early epidemic data for downstream data analytics." I would like to sincerely commend the authors for their effort and dedication. The topic addressed is highly relevant and impactful.

Overall, the paper presents a proposal for the collection of data during the early phase of a disease outbreak by establishing a unified schema that captures the most relevant indicators for decision-making. I believe the manuscript is suitable for indexing after a medium revision. In particular, I have the following comments:
We thank the reviewer for their positive assessment of our work.
Abstract
1. The goal of the paper is not mentioned. Please add it in the background subsection.
Thank you. We have added the goal of the paper in the abstract and background section. Methods
2. The description of the methods could be improved. Briefly describe Figure 1, indicating that it represents a methodology based on different steps. **Thank you. The methods section has been reorganized and rewritten for clarity. Figure 1 has been revised and described for clarity.**
3. The indication that the early phase refers to the first 100 days should be introduced earlier in the text, ideally at the first instance it is mentioned, to ensure clarity for the reader.
Thank you. This has been clarified by moving the introduction of this term earlier in the introduction section and defining it at its first instance.

Introduction
4. Please add a paragraph about the impact of diseases and the relevance of studying them. For instance, include the number of deaths from COVID-19 using data from the WHO COVID-19 dashboard or the number of disease outbreaks worldwide from this paper: https://doi.org/10.1038/s41597-022-01797-2
Thank you. While useful information for any health-related research, the authors do not consider the additional value of such information in the context of the current work. 5. Define the concept of data interoperability. **Thank you. Data interoperability implies the seamless access, exchange, integration, and cooperative of use of data in different systems, applications, and organizations in a coordinated way, ensuring meaning, integrity, and usefulness are maintained. This definition has been added to improve the introductory section**.
6. Add a paragraph describing the goal of the paper and its structure. **Thank you. The last paragraph in the introduction now describes the goal and structure of the paper.**
Methods
7. Figure 1 (Conceptual framework of this study) can be improved by presenting it as a logical pipeline, numbering and naming each stage (e.g., Stage 1: Identification of relevant variables, Stage 2: Identification of the minimum set of indicators, etc.). Then, describe these stages in the text using exactly the same names as in Figure 1.

Thank you. Figure 1 has been updated to improve clarity per your suggestions.

Results
8. Elaborate the results according to the stages defined in the Methods section. Follow the same structure and add a framework showing the outputs for each stage and step. **Thank you. The results section has been restructured to match the sections of the methods. Outputs have been detailed in each Table to show the outputs of the steps.**
9. In the paragraph "Literature review of key parameters to be addressed during the first 100 days of an epidemic", the authors use both "35" and "Thirty-five". Please standardize the format: use numerals for numbers 10 and above, and words for numbers zero through nine. **Thank you. Use of words and figures for numbers has been standardised.**
10. In the paragraph "Global.health data schema core variables", the authors use inconsistent formatting for category names (e.g., Laboratory Information vs. Medical history). Please standardize the formatting. **This has now been addressed - thank you.**
11. Include more details about the Mpox data ingestion in the Democratic Republic of Congo. What was learned from that experience? Please add a paragraph discussing the practical feasibility. **Thank you. That information is beyond the scope of this paper and is under consideration for publication in a separate work. We removed mention of specific outbreaks.**

Discussion
12. The authors mention that the Global.health core schema may not be comprehensive for all pathogen groups. Please provide an example of a pathogen for which the proposed schema might be insufficient. **Thank you. We are not aware of an actual example of a pathogen for which the proposed schema might be insufficient. We have provided Disease X as an example and suggested ways in which it may require variables outside of the current schema.**
13. The manuscript mentions that language could be a barrier for non-native English speakers. How was this challenge addressed in the Democratic Republic of Congo? **Thank you. In the DRC the primary language for ingestion was French and we used translation services and large language models to translate from French to English. While we managed this situation, languages with more complex translation requirements or different characters may not be so manageable.**

Competing Interests: No competing interests were disclosed.