scientific data

OPEN

DATA DESCRIPTOR

Constructing a global human epidemic database using open-source digital biosurveillance

Rinette Badker ✉, Naama Kipperman, Benjamin Ash, Nita K. Madhav, Ben Oppenheim, Patrick Savage, Nicole Stephenson & Chris Pardee

We developed a dataset consisting of outbreak data collected from official, open-source surveillance reports representing more than 170 pathogens, 237 countries and territories, and more than 3300 events that occurred primarily between 1963 and 2023. Here we present and analyze a subset of these data, comprising a dataset of human epidemic events with onset between 2015 and 2020. Structuring of epidemiological data in the dataset follows a specific methodology to ensure consistency across all events. This methodology has been designed to produce the most reliable spatiotemporal view of an outbreak as possible. To ensure data are true-to-source, the structured data undergoes multiple rounds of both manual and automated review and validation. The extensive and standardized nature of the dataset makes it well-suited for both descriptive epidemiology and exploring outbreak dynamics and disease emergence.

Background & Summary

A historical perspective on epidemic frequency and intensity is essential to inform public health action and provide a baseline understanding of infectious disease outbreak trends. Datasets of infectious disease outbreaks are increasingly used to analyze the changing disease landscape, train artificial intelligence (AI) and forecasting models, and guide surveillance programs[1–3]. However, many existing public health datasets do not provide comprehensively and methodologically-consistent coverage of historical outbreaks, in an easily accessible format.

Currently, infectious disease datasets are limited to data that cover only a limited number of pathogens (e.g., nationally notifiable diseases), are exclusive to a specific location or population (e.g., hospitalized patients or at risk groups), or lack historic coverage (i.e., covering only recent outbreaks)[1,4,5]. Existing outbreak datasets that span a wide range of pathogens at a global scale include the Global Infectious Disease Epidemiology Online Network (GIDEON) database (https://www.gideononline.com/) and the World Health Organization's Disease Outbreak News (WHO DON) reports[6]. While these datasets contain valuable information, they either require large amounts of additional processing to extract important outbreak data due to their format, or they are limited in the detail (especially spatiotemporal granularity) they offer about epidemic spread and impact.

Using publicly available outbreak reports, we have constructed a Human Epidemic Database (HED), including infectious disease events that could pose a significant risk to public health and/or societal, economic, or political stability. We defined an event according to the causative pathogen, country of origin, and the year in which the first case occurred. Events can reflect both discrete epidemiologically linked outbreaks, as well as broader annual or multi-year epidemics caused by endemic diseases. This methodology has been developed and refined over eight years to produce the most reliable dataset possible on the spatiotemporal distribution of reported cases and deaths.

The HED includes temporal and geographic data collected from over 500 distinct reporting sources comprising data from more than 170 unique pathogens, 237 countries (and territories/areas), and over 3300 distinct events primarily spanning 1963 to 2023 as of 12 January 2024. A limited number of epidemiologically significant events starting before 1963 with available reporting sources, including the 1918 influenza pandemic, were also structured. Over the course of eight years, our team has carefully structured an extensive dataset which connects disease activity geospatially and with thorough consideration of available spatiotemporal resolution and the distinction between confirmed, probable, or suspected cases and deaths.

Ginkgo Bioworks Inc., Boston, USA. ✉e-mail: rbadker@ginkgobioworks.com

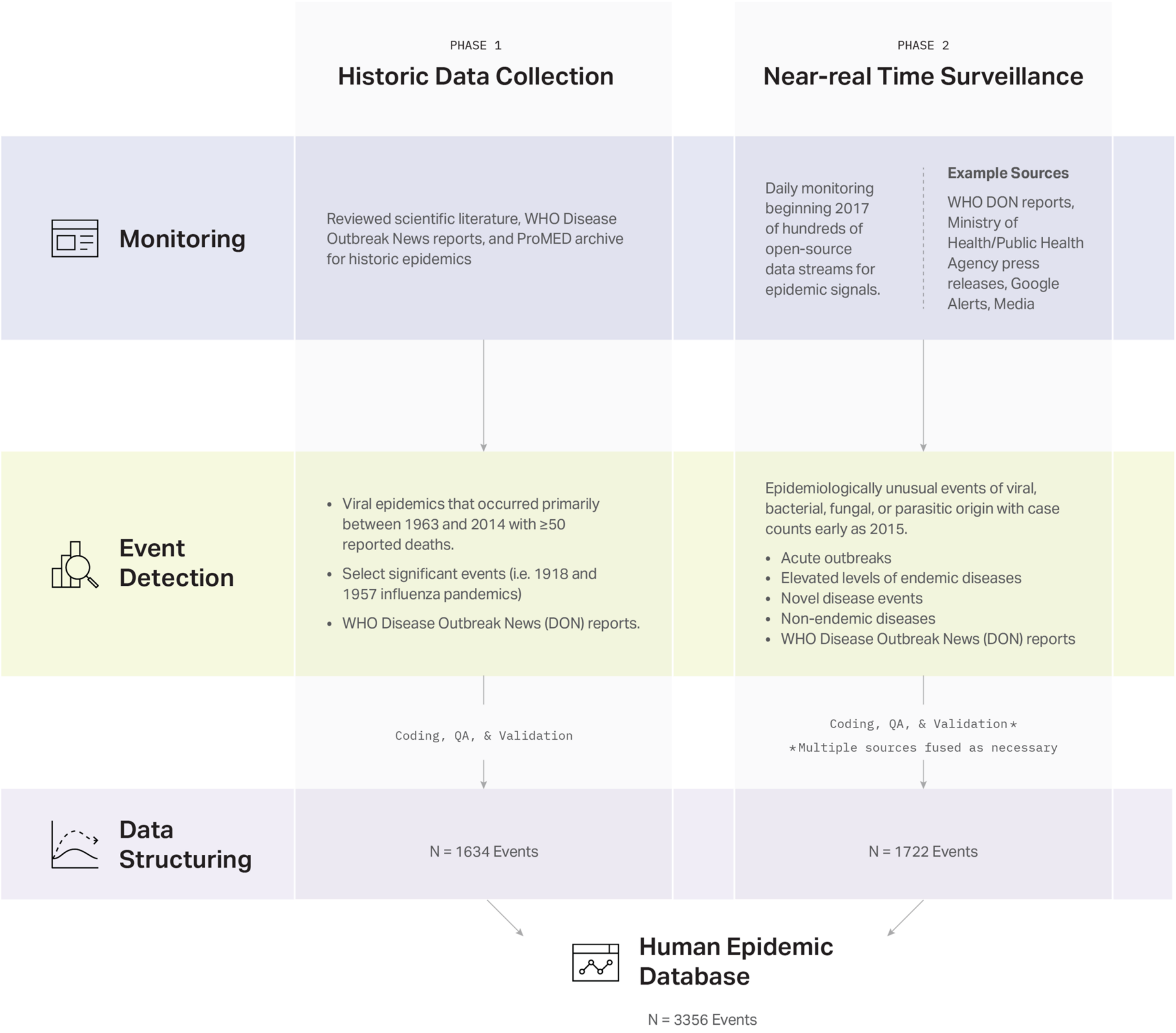
PHASE 1
Historic Data Collection
PHASE 2
Near-real Time Surveillance
Monitoring
Reviewed scientific literature, WHO Disease Outbreak News reports, and ProMED archive for historic epidemics
Daily monitoring beginning 2017 of hundreds of open-source data streams for epidemic signals.
Example Sources
WHO DON reports, Ministry of Health/Public Health Agency press releases, Google Alerts, Media
Event Detection
Viral epidemics that occurred primarily between 1963 and 2014 with ≥50 reported deaths.
Select significant events (i.e. 1918 and 1957 influenza pandemics)
WHO Disease Outbreak News (DON) reports.
Epidemiologically unusual events of viral, bacterial, fungal, or parasitic origin with case counts early as 2015.
Acute outbreaks
Elevated levels of endemic diseases
Novel disease events
Non-endemic diseases
WHO Disease Outbreak News (DON) reports
Coding, QA, & Validation
Coding, QA, & Validation*
*Multiple sources fused as necessary
Data Structuring
N = 1634 Events
N = 1722 Events
Human Epidemic Database
N = 3356 Events

Fig. 1 Schematic of Human Epidemic Database (HED) monitoring, event detection, data collection, and data structuring workflow (data as of 1/12/24).

The process of identifying and structuring events for inclusion in the HED has undergone two phases (Fig. 1). Phase 1 of HED assembly occurred in 2015-2016 and focused on completed (historic) epidemics meeting a relatively narrow set of inclusion criteria. A more comprehensive Phase 2 has been ongoing since 2017 and focuses on contemporary (near real-time) structuring of data on epidemics. When presenting the methodology used to create and expand the HED, we have focused on data collected during the first years of Phase 2 data collection. These data cover events with epidemiological start dates between 2015 and 2020. The COVID-19 pandemic significantly impacted the public health reporting landscape, and any discussion of public health reporting during this period deserves its own investigation[7]. As we are presenting a subset of Phase 2 events, the dataset contains a smaller number of events compared to Fig. 1.

Data Collection Methods

Phase 1 of the HED is primarily comprised of events that occurred between 1963 and 2016, as well as some earlier significant events such as the 1918 and 1957 influenza pandemics. The narrow criteria that guided data collection for this phase means that only a subset of reported epidemics from this period have been included. Events collected during Phase 1 were limited to epidemics that were caused by a viral pathogen and resulted in 50 or more reported deaths. Furthermore, all public health events reported by the WHO DON were also included. Certain pathogens – such as seasonal influenza, malaria, or HIV – were excluded as the initial focus was on collecting data for newly emergent epidemic events.

As an evolution from Phase 1, Phase 2 HED collection identifies and structures data on a much broader range of epidemics. The restrictions to viral pathogens and high mortality events were removed to allow for the inclusion of a larger variety of medium- and high-consequence epidemics. Phase 2 events include any epidemiologically noteworthy increases in cases or deaths caused by a bacterial, viral, fungal, or parasitic pathogen. Epidemics within this framework include novel or emerging infectious disease events, elevated transmission of endemic diseases, and acute outbreaks of high consequence pathogens. All events reported in the WHO DON

Component	Values	Definition
Pathogen	Group 1	Every observed pathogen has previously been assessed individually based on its primary transmission route, capacity for human-to-human transmission, capacity for vector-to-human transmission, treatment availability, capacity for animal-to-human transmission, vaccine availability, case fatality rate, frequency of outbreaks, and Select Agent status. For events caused by unknown or not yet diagnosed pathogens, we use broad pseudo-pathogens based on symptomology (e.g., uncertain low-mortality gastrointestinal or uncertain high-mortality febrile). These variables determine which prioritized group each pathogen fits into for event scoring purposes, with Group 1 representing the most significant human pathogens and Group 10 representing the least significant human pathogens. A sample of pathogens for each group is available at https://doi.org/10.6084/m9.figshare.28187936.
	Group 2	
	Group 3	
	Group 4	
	Group 5	
	Group 6	
	Group 7	
	Group 8	
	Group 9	
	Group 10	
Epidemiological Characteristics	Novel/Emerging	The pathogen is new or emerging on a global scale.
	Not Endemic	The pathogen is not endemic to the impacted location but is not considered a novel or emerging pathogen.
	Epidemic	The pathogen is endemic to the impacted location, and activity has reached epidemic levels as defined as two times the 5-year average of cases in that location. Alternatively, the pathogen is endemic to the impacted location, but authorities have defined a unique outbreak associated with shared exposure mechanisms.
	Endemic: Increase	The pathogen is endemic to the impacted location, and activity is above expected levels but not yet above epidemic levels.
	Endemic: No change	The pathogen is endemic to the impacted location, and activity is in line with expected levels or information is not available to determine what expected levels would be.
	Imported	The pathogen may or may not be endemic to the impacted location, but the reported cases were actually exposed in a different location.
	Endemic: Decrease	The pathogen is endemic to the impacted location, and activity is below expected levels.
Geographic Scale	Global	Local transmission of cases in two or more World Bank Regions[13].
	Regional	Local transmission in two or more countries within a single World Bank Region or the world.
	National	Local transmission in two or more states/provinces within a single country.
	Unclear	Reported information is not sufficient to determine the geographic scale of the event.
	Local	Local transmission in a single state/province.
	Imported	No local transmission; reported cases were exposed in a different location.
Case Scale	$\geq$10	At least 10 cases have been reported in the location in question.
	$<$10	Fewer than 10 cases have been reported in the location in question.

Table 1. Event inclusion score components collected during event detection.

continue to be included for consistency. Many Phase 2 events undergo data collection from multiple sources to present a more comprehensive spatiotemporal dataset when available. Additionally, when no single reporting source sufficiently summarizes an entire event, multiple structured sources may be fused into a more comprehensive view of the event as previously described in Badker *et al.*[8].

Phase 2 data collection efforts officially began in January 2017. However, there was a period of overlap between the end of Phase 1 collection and beginning of Phase 2 collection. Events that occurred during 2015 and 2016 – that is, events that emerged and continued during the period in which our team were collecting historical data, such as the emergence of Zika virus in Brazil during 2015 – were structured more similarly to other Phase 2 events and are therefore included in our summary and use-case analysis of near-real-time events. Phase 2 data collection efforts have been more sophisticated than Phase 1 efforts and have resulted in a more comprehensive dataset of events beginning in 2015 and later. The Phase 2 event detection and data collection processes are described here.

Event Detection

The Phase 2 event detection system aims to quickly identify and prioritize epidemics for ongoing monitoring and risk assessment. Digital surveillance experts conduct daily monitoring to identify newly available infectious disease data. Such monitored sources include official public health reports, unofficial outbreak aggregators (for example ProMED mail or Outbreak News Today), news media, and trusted social media accounts. Each event is qualitatively and quantitatively assessed based on the best available geographic and epidemiological details, and only events that meet an established inclusion score are structured in the HED.

Each event's inclusion score is based on specific information collected from identified epidemic reports. Such scoring components include the pathogen responsible for the event, the epidemiological characteristics of the pathogen in question at the location(s) being impacted, and the geographic scale of the event (Table 1). Other key details collected but not incorporated into the event's inclusion score include the earliest and latest known dates of cases and deaths for the event in question, the date on which the information being collected was published, and the location(s) being impacted by the event.

A team of subject matter experts (SMEs) meets daily to discuss newly monitored events, agree on tracked details, and review the current inclusion score for each event. First, each event is checked for consistent processing and analysis, and the specific details are discussed to ensure accuracy of the summary data collected from each signal source. This summary data is collected in a shared dataset, where each row of data represents summary information for a single location, impacted by a single event, over a single period of time. As such, each event is likely to contain multiple rows in the dataset to incorporate multiple national or subnational locations and multiple date ranges.

When collecting and reviewing an event's summary data, the SME team populates the tracker dataset with data providing the necessary inclusion score components listed in Table 1. Each component has its own field in the tracker, and each field contains a formula to score the individual component in question. For example, the Pathogen component scores higher for events caused by an etiology in Group 1 than in Group 10 (see the pathogen dataset in figshare). The tracker dataset automatically calculates each event's inclusion score by weighting and combining each component score. Although each component score contributes to the overall inclusion score, the scoring algorithm weights the Pathogen component highest, followed by the Epidemiological Characteristics and Geographic Scale components, followed by the Case Scale component. Because the tracker dataset typically contains multiple rows for each event, each of which likely contains different details leading to different local inclusion scores for that particular signal, the dataset also includes a function to visualize the highest of all local inclusion scores for each event on the tracker. This event level inclusion score is what determines whether a given outbreak or epidemic scores high enough for inclusion in the HED.

Events that score below the inclusion score threshold are typically excluded from further structuring, and those scoring above the threshold typically move on to data structuring in the HED. One example of an event that did not meet the inclusion score is Cryptosporidium sp._France_2019. The event was identified in November 2019 and scored below the threshold due to the "Local" geographic scale, the low pathogen score for *Cryptosporidium*, and the "Endemic: No change" epidemiological characteristics (https://www.ouest-france.fr/sante/cryptosporidiose-cinq-questions-sur-cette-maladie-qui-se-propage-dans-le-sud-est-de-la-france-6620419). It should be noted, though, that all events reported in a WHO DON are structured regardless of their inclusion score, reflecting the WHO's assessment that the event is epidemiologically significant. During this process, the SME team also discusses event definitions to distinguish similar events from one another and to allow for cross-jurisdictional mapping of related events.

Furthermore, events which are initially excluded from structuring in the HED are frequently reevaluated as new information becomes available and their inclusion score updated when needed. For example, the epidemiological characteristics for an event may change from "endemic: no change" to "epidemic" when the case count exceeds the impacted location's 5-year average. The increase in the component score due to this development may change the event inclusion score enough to raise it above the inclusion threshold. At this time the event would then be structured for inclusion in the HED. One example of this occurring is Coccidioides sp._United States_2019 which was initially excluded from structuring. When the event was first identified the geographic scale was "Local" and the epidemiological characteristics were "Endemic: No change". The next time the event was scored it met the inclusion threshold after the geographic scale rose to "National" and the epidemiological characteristics changed to "Endemic: Increase" in relation to the calculated 5-year average.

Data Structuring

Events that meet the inclusion criteria move on to the data structuring phase. For each such event, the SME team evaluates potential official sources to choose the most reliable source or sources of spatiotemporal data. When necessary, multiple sources may be structured and combined for a more complete spatiotemporal representation of the event[8]. Data sources are evaluated using qualitative and quantitative metrics encompassing source reliability, availability/permanence, consistency, structure, clarity, and the spatiotemporal resolution of provided data. On rare occasions when no official source could be identified for the source, reliable "Media" sources may be structured in order to capture the event. The methodology used to structure these sources resembles the creation of multisource fusion files.

Data structuring moves cyclically through three stages, which results in newly available data being integrated into the HED. Once an event reaches the structuring phase, it is followed until it is declared over by an official body (such as the WHO) or its sources discontinue coverage (i.e., 90 days has passed since the last reported case). Each time new information is reported by a structured source, the event cycles through the three stages again to ensure the most up-to-date information has been included in the HED. While the goal is to have a complete overview of each identified epidemic, only information provided by official reporting sources such as national public health agencies, the WHO, or a regional disease control agency, is structured for the HED. Since not all reporting agencies publish updates consistently or reliably across events, HED data may be limited for some events.

The first stage in the data structuring cycle is data coding and quality assurance (QA). Each structured dataset represents a single source reporting on a single event, but multiple published versions of the source may be combined into a unified dataset, with careful consideration of data conformity. New information is added to the dataset by a data analyst and then passed to a different data analyst for review and QA. Since not all data is reported equally, some data interpretation may be made during coding and subsequently agreed on or revised during QA. For example, cases and deaths can be classified as confirmed, suspected, or probable depending on available case definitions and laboratory results. Every attempt is made to be consistent with the reporting source, but on occasion subject matter experts are consulted to review the context. For example, during a multi-year outbreak Dengue virus_Reunion_2017, a reporting source switched to reporting cases for the current year rather than the entire outbreak. Context in other reports confirmed that, despite the change in reporting, the outbreak was ongoing, so we started adding the previous year's case and death counts to the current ones.

During QA, the coded data is reviewed against the original source material to check for coding errors, methodological errors, or interpretation differences. A dataset may go back and forth during this process until the coder and reviewer both agree with any decisions that were made when converting the source material to a dataset. Once the QA process has been completed, the data is considered "gold standard", that is, to be the most accurate possible encoding of the source material.

Quantitative metric	Definition
Frequency	the number of reports over the duration of the event
Consistency	the standard deviation of inter-report intervals
Timeliness	the number of days between the first case occurring and the first official report
Case and death completeness	the percentage of cases and deaths in the data that were reported directly by the source as opposed to derived based on reported new and cumulative cases or deaths
Subnational completeness	the proportion of cases allocated to the subnational level
Accuracy	the proportion of total cases that have been classified as confirmed

Table 2. Quantitative metrics of official reporting sources in the HED.

Data Elements	Variable Name	Definition
Event Identifier	Event name	A unique name given to each event which consists of the pathogen, country of origin, and the start year of the event in question.
Pathogen Fields	Pathogen	The disease-causing organism of the event.
Date Fields	Start Date	The earliest available event date for the location in question.
	End Date	The latest available event date for the location in question.
Location Fields*	Country/Territory/Area	The country name for the impacted location.
	ISO 2	ISO country code
Case and Death Fields	Cumulative Cases	The total number of reported cases for the affected location during the event timeframe identified by the start and end dates.
	Cumulative Confirmed Cases	The total number of reported laboratory confirmed cases for the location in question during the event timeframe.
	Cumulative Deaths	The total number of reported deaths for the location in question during the event timeframe identified by the start and end dates.
	Cumulative Confirmed Deaths	The total number of reported laboratory confirmed deaths for the location in question during the event timeframe.
Source Fields	Source Name	The reporting source the data was retrieved from.
	Source URL	The URL associated with the last update for a given event/country at the time the update was made.

Table 3. Data elements included in the dataset. *Location fields are blank for rows that provide an event wide total.

Data validation and ingestion are the second stage of the HED structuring process. After QA is completed, a subject matter expert prepares the gold standard data for ingestion into the HED. This data is processed by an automated validation tool to check for logical inconsistencies and perform predefined spatiotemporal transformations as needed. Such transformations ensure that reported cases and deaths accurately roll up through space and time. For example, these transformations ensure that cases reported in California are also reflected in the United States as a whole, or that deaths reported in January are accurately incorporated into cumulative totals reported as of February. Once a dataset successfully passes validation, it can be ingested into the HED.

During the ingestion process, quantitative metrics are calculated for each event based on certain qualities of the reporting source (Table 2). The overall quantitative score represents a weighted aggregation of each individual metric score. Scores are calculated individually for every source structured in each event, and these scores are normalized against the maximum possible quantitative score. This allows each event-source's scores to be compared so that users can select sources that fit their use-cases best.

During the final stage of data structuring, a subject matter expert reviews and graduates the dataset after it has been ingested into the HED. The gold standard file is once again reviewed against the source material to confirm accuracy and data interpretations, with greater scrutiny of broader contexts within and without the event in question. The gold standard file is then reviewed against a visual preview of the validated and ingested data. This ensures the automated geotemporal transformations conducted during validation have not caused any errors in the data. As in the QA stage, a dataset may go back and forth during this process until the SME and another senior team member agree on its accuracy and completeness. Once this is done, the dataset is officially incorporated into the HED.

For large events or those which cross borders, it is rare for a single reporting source to provide comprehensive surveillance data[8]. Therefore, some Phase 2 events may require multiple datasets from complementary data sources be fused to provide a full overview of the event. The source for these events is described as 'Multisource Fusion'. The process to create such a fusion requires a senior team member to align all structured sources for the event according to date range and location. They then combine these sources into a single dataset, with special attention given to ensure that individually-reported cases and deaths are not double-counted and that sudden increases or decreases in the fused epicurve reflect actual reporting rather than discrepancies between sources. For example, a Multisource Fusion for the event Zaire ebolavirus_Democratic Republic of the Congo_2018b combines data from the independently structured sources "Democratic Republic of the Congo MOH Communique de Presse", "Uganda Ministry of Health Press Releases", "WHO AFRO External Sitreps", "WHO AFRO Weekly Bulletin on Outbreaks and Other Emergencies", and "WHO DON". Similarly, an example of a rare Media fusion can be found for the event Dengue virus_Philippines_2018, which incorporates articles published by the media outlets Manilla Bulletin, Outbreak News Today, Philippines News Agency, and Philstar.

Pathogen	Number of unique outbreak events
Zika virus	117
Measles virus	98
Dengue virus	90
Salmonella	82
Shiga toxin-producing *E. coli*	43
Poliovirus	38
Yellow fever virus	34
Vibrio cholerae	32
Listeria monocytogenes	27
Chikungunya virus	21
Lassa virus	21

Table 4. Top 10 pathogens with the highest number of unique outbreak events from 2015–2020, as of 1/12/24.

Country of origin	Number of unique outbreak events	Country affected*	Number of outbreak events
United States	130	United States	145
Canada	36	Canada	52
China	36	China	37
Nigeria	21	France	35
Brazil	20	Germany	29
France	20	Japan	29
Pakistan	20	United Kingdom	28
Mexico	17	Sweden	27
United Kingdom	17	Brazil	24
Democratic Republic of the Congo	16	Nigeria	23
Sweden	16		

Table 5. Top 10 countries of origin and countries affected by outbreaks from 2015–2020, as of 1/12/24. *Some events affect multiple countries and may therefore be counted more than once. Country counts should not be added together for regional totals.

Country origin-pathogen combination	Number of unique outbreak events	Country affected-pathogen combination*	Number of outbreak events
United States - Salmonella	43	United States - Salmonella	46
United States - Shiga toxin-producing *E. coli*	21	Canada - Salmonella	22
Canada - Salmonella	14	United States - Shiga toxin-producing *E. coli*	22
United States - *Listeria monocytogenes*	10	Japan - Measles virus	14
China - Novel Influenza A (H5N6)	7	South Korea - Measles virus	10
China - Novel Influenza A (H9N2)	7	United States - *Listeria monocytogenes*	10

Table 6. Top 5 country-pathogen combinations 2015–2020, as of 1/12/24. *Some events affect multiple countries and may therefore be counted more than once. Country counts should not be added together for regional totals.

Data Records

Here we present a dataset consisting of HED events with start dates between 2015 to 2020. These data have been made available in a public data repository as a comma separated value (CSV) file[9]. A breakdown of the included variables are described in Table 3. Also of note, a more granular COVID-19 dataset is hosted on the Humanitarian Data Exchange (https://data.humdata.org/dataset/2019-novel-coronavirus-cases), which includes some additional variables along with descriptions. Some of the events outlined in Fig. 1 are also available in a public repository as part of a publication assessing historical trends in zoonotic spillover events (https://github.com/concentricbyginkgo/zoonotic_spillover_trend)[10].

Events are given a unique identifier which uses the pathogen name, location where the outbreak is first identified, and the year the outbreak started. Occasionally additional text is added to the event name to differentiate discrete events with overlapping pathogen and spatiotemporal details. The HED includes a row for each country impacted by an event and an event wide row which includes the case and death counts for all locations impacted. For a small number of events, available information did not allow for all cases or deaths to be attributed to a specific country so the event wide row will exceed the sum of each impacted country. This event wide row has the location fields blank as they are not specific to an individual country. All data in the HED is specific to the event represented in the associated row.

Pathogen	Country of origin	Total countries spanned	Event start date	Event end date	Cumulative reported cases	Cumulative reported deaths	Implied Case fatality ratio (%)
SARS Coronavirus 2	China	234	12/12/2019	ongoing*	773,448,535	6,991,829	0.9
Zika virus	Brazil	55	1/4/2015	12/31/2016	749,694	45	0.01
Plasmodium sp.	Venezuela	4	12/30/2018	10/19/2019	326,659	100	0.03
Measles virus	Ukraine	6	1/1/2018	11/6/2019	112,867	41	0.04
Measles virus	Philippines	9	12/17/2018	12/13/2019	42,580	477	1.12
Dengue virus	Reunion	1	4/10/2017	12/31/2020	41,225	42	0.1
Measles virus	Venezuela	7	6/25/2017	12/31/2019	33,437	97	0.29
Dengue virus	Myanmar	1	1/1/2017	8/19/2017	21,288	131	0.62
Salmonella	Pakistan	13	11/1/2016	5/17/2023	21,276	—	—
Measles virus	Romania	2	1/1/2016	7/17/2020	20,358	64	0.31

Table 7. Profile of largest magnitude events (as measured by cumulative reported cases) in HED 2015–2020. *SARS-CoV-2 data as of 1/12/24.

Overview of Dataset

The standardization of open-source disease surveillance data into a single comprehensive dataset allows for a global view of outbreak characteristics and trends over time. Since Phase 2 captures the more comprehensive, near-real-time collection effort, we present events in the HED starting between 2015 and 2020. Summaries and visualizations were generated using RStudio version 2023.06.1 (https://posit.co/products/open-source/rstudio/). A breakdown of the countries and pathogens most represented in this phase of the HED can be seen in Table 5. The United States, Canada, and China are associated with the largest number of events during this time frame. These countries report both the largest number of events originating within their boundaries and the largest number of events with at least one case reported in their boundaries.

While HED events starting 2015 to 2020 represent approximately 120 different causative pathogens, the top 10 most represented pathogens make up over 50% of these events (Tables 4–6). Zika virus is the most represented pathogen (117 distinct events) followed by measles virus (98), and dengue virus (90). Several of the frequently reported pathogens in the HED align with the most frequently reported pathogens in an aggregated dataset of WHO DON reports from 1996 to 2022[6]. The most common countries and pathogens summarized in Tables 4–6 reflect epidemiological contexts, the global outbreak reporting ecosystem, and internal methodologies discussed in the "Data Collection Methods" section. To hone in more narrowly on epidemiological trends, additional exclusion criteria can be applied to minimize the confounding effects of reporting bias, as Meadows *et al.* did with zoonotic spillover events[10].

Table 7 provides a profile of the ten largest magnitude events (as measured by cumulative case counts per events) in the HED during this window. A more granular event-level view can answer questions about the features of outbreaks over time. For example, an event-level view allows us to see that large scale epidemics have diverse causative pathogens and geographic points of origin.

Figures 2 and 3 highlight the breadth, scope, and heterogeneity of events in the HED. Figure 2 shows a breakdown of outbreak events by magnitude, implied case fatality ratio, and primary pathogen transmission route. The number of structured events ranges from 76 events starting in 2015 to 273 events starting in 2019, with a median of 167 events starting each year. The years 2017 and 2019 saw the largest number of events, at 259 and 273 events respectively. In 2020, there was a notable decrease in the number of unique new events structured for the HED compared to the preceding years, at 135 events. This decrease is likely due to disruptions to routine surveillance activities globally during the first year of the COVID-19 pandemic, rather than any change in HED collection methodology[7]. Events that span multiple countries and or regions comprise less than 10% of all events but constitute the highest consequence events in terms of case counts and fatalities (Table 7).

These data allow for summarizing region-specific outbreak epidemiology. The quantity and types of events vary across regions (Table 8). The region affected by the largest number of events during this time frame is Sub-Saharan Africa (226 events), followed by Latin America & Caribbean (218) and Europe and Central Asia (183). The proportion of total outbreak cases attributable to each region also varies substantially by year (Fig. 3b/c).

Technical Validation

Prior to being incorporated into the HED, each event goes through a validation stage which includes both automated and manual evaluation of the dataset, as described under "Data Structuring" above. Each individual event is checked to ensure all required fields have been populated in accordance with established naming dictionaries. As the HED is being continuously updated, manual checks and automated validation are also regularly completed by the team.

Usage Notes

The HED contributes significantly to the ongoing study of global patterns of disease epidemics. The HED has been structured to ensure consistency across all events, which enables users to easily extract outbreak data according to epidemiological, spatial, or temporal criteria. The comprehensive scope of the dataset makes these data useful for both descriptive epidemiological purposes and for exploring research questions related to outbreak dynamics and disease emergence. Meadows *et al.* used the HED to analyze trends in viral zoonotic spillover events and characterize their frequency and severity over time[10]. The analysis applied strict exclusion criteria

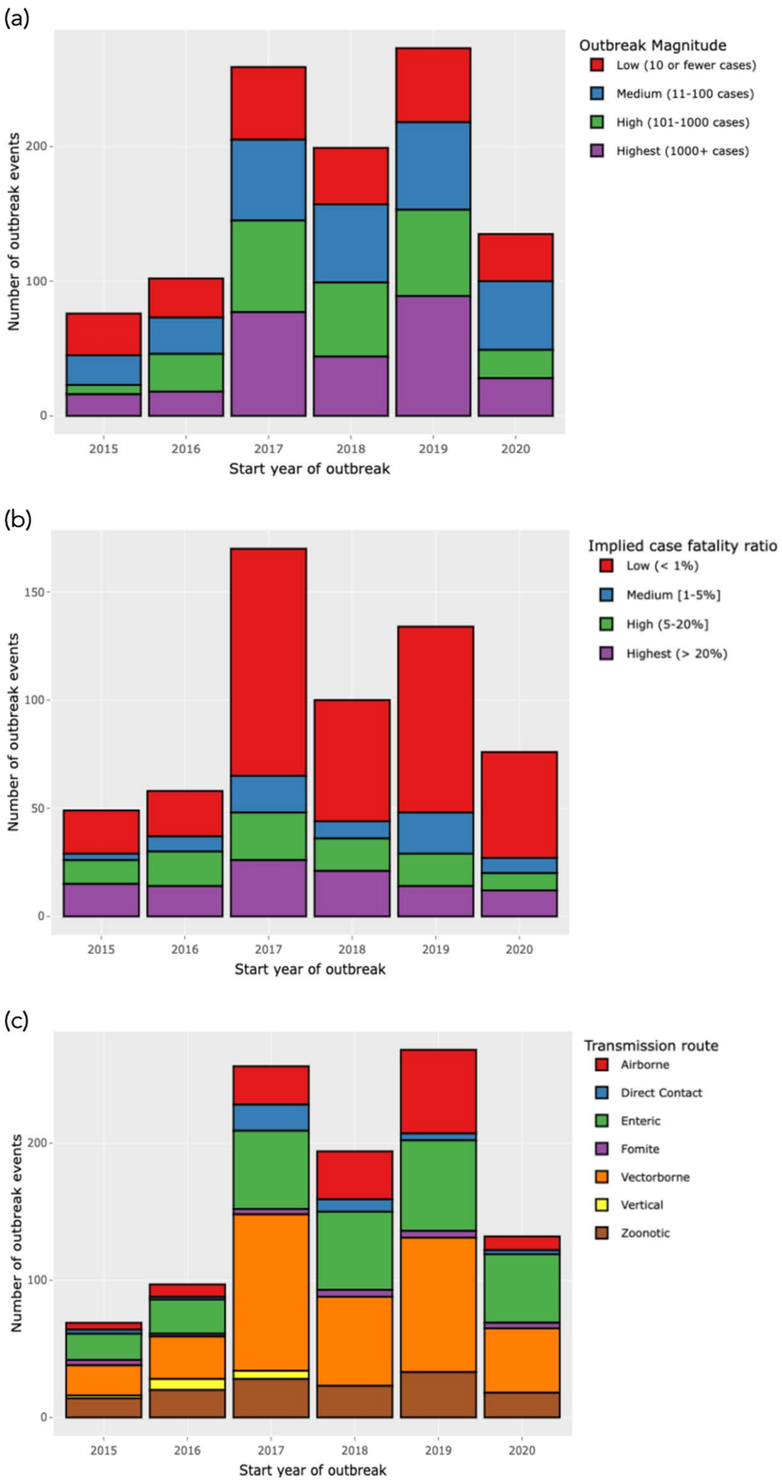

Fig. 2 The number of events over time by (**a**) case magnitude; (**b**) implied case fatality ratio; (**c**) transmission route.

to address potential confounding caused by changes in event detection and reporting over time, including the change in outbreak detection between Phase 1 and Phase 2.

The HED can also be combined with other types of geospatial data using ISO country codes to match locations. This opens diverse possibilities for exploration of regional, country-level, or sub-national factors related to infectious disease burden. Given the multidisciplinary nature of epidemic risk, a comprehensive global epidemiological dataset that can be analyzed alongside ecological, political, economic, and other indicators is valuable. The dataset has already been used to support different analyses for policy fora and standards-setting. Policy-relevant analysis and engagement supported by the HED includes Oppenheim *et al.*'s

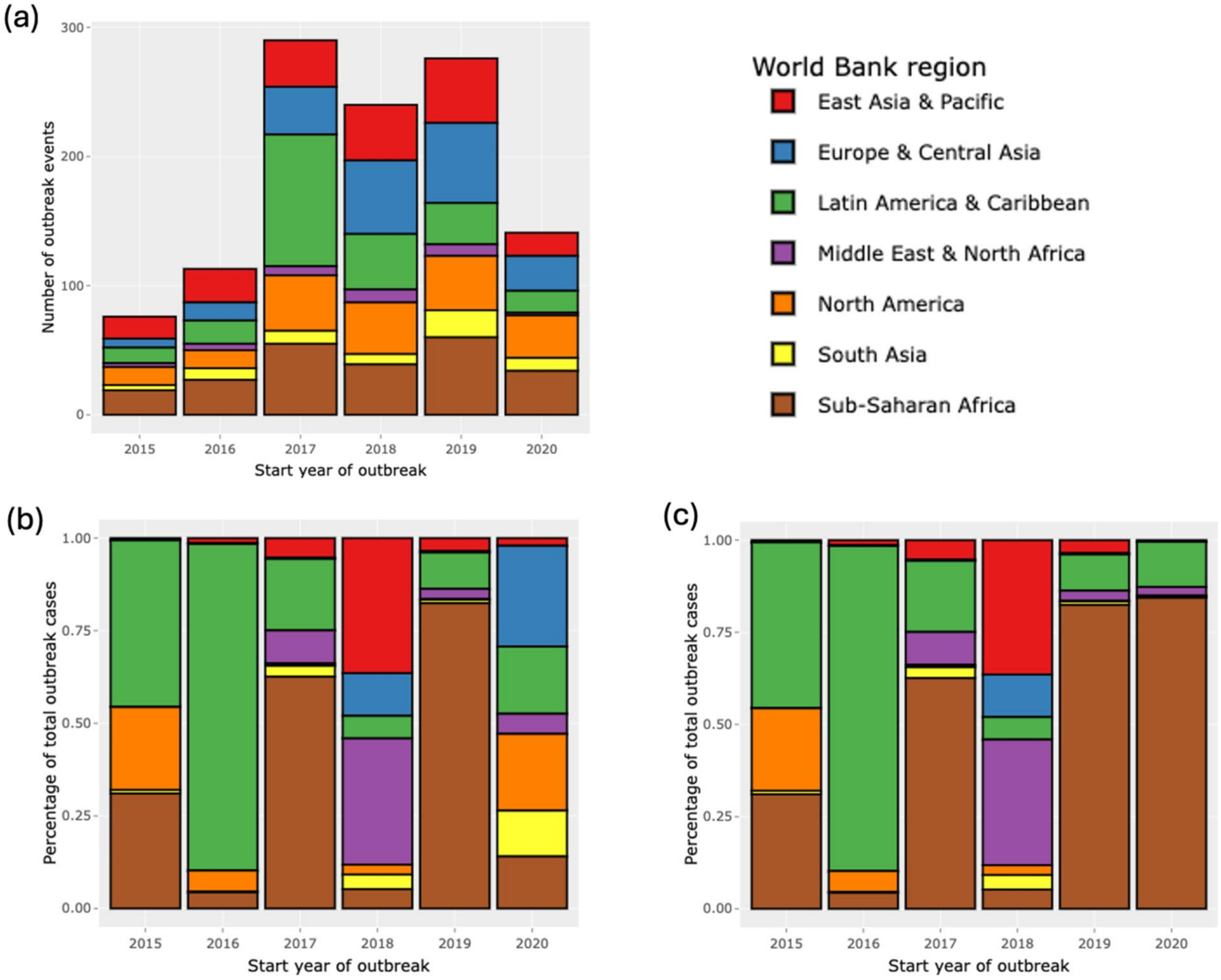

Fig. 3 The number of events and reported outbreak cases 2015–2020 by World Bank region over time (**a**) total events; (**b**) proportion of reported outbreak cases; (**c**) proportion of reported outbreak cases excluding SARS-CoV-2.

World Bank Region[13]	Total events per region	Top five pathogens
Sub-Saharan Africa	226	*Vibrio cholerae*, measles virus, Lassa virus, yellow fever virus, poliovirus
Latin America & Caribbean	218	Zika virus, dengue virus, viral conjunctivitis, yellow fever virus, *Plasmodium* spp.
Europe & Central Asia	183	Measles virus, salmonella, Shiga toxin-producing *E. coli*, *Listeria monocytogenes*, *streptococcus pyogenes*
North America	182	Salmonella, Shiga toxin-producing *E. coli*, *Listeria monocytogenes*, measles virus, West Nile virus
East Asia & Pacific	179	Measles virus, dengue virus, Zika virus, Severe fever with thrombocytopenia syndrome virus, Novel Influenza A (H9N2), poliovirus
South Asia	62	Dengue virus, poliovirus, Nipah virus, Crimean-Congo hemorrhagic fever virus, Human immunodeficiency virus
Middle East & North Africa	36	Measles virus, Crimean-Congo hemorrhagic fever virus, dengue virus, poliovirus, *Vibrio cholerae*, West Nile virus

Table 8. Total number of outbreaks in HED by World Bank Region[13] affected and most common outbreak pathogens 2015–2020.

development of an Epidemic Preparedness Index (EPI)[11] – a quantitative metric for capturing a country-level pandemic preparedness – and a 2021 report of the G20 High Level Independent Panel about Pandemic Preparedness and Response (https://pandemic-financing.org/wp-content/uploads/2021/07/G20-HLIP-Report.pdf). More recently, the HED has informed a working paper prepared for the Disease Control Priorities Project (4) about drivers of pandemic risk (https://www.cgdev.org/sites/default/files/estimated-future-mortality-pathogens-epidemic-and-pandemic-potential.pdf). The data collection process of the HED has also informed scientific engagement about the challenges of open source disease surveillance data and avenues for technical improvement and standard-setting[8].

In a public health reporting landscape where data are highly fragmented and variably structured, a comprehensive dataset of high-quality, standardized epidemiological data holds distinct value. This unified geospatial view of epidemic burden and spread over time can support outbreak research and provide historic epidemiological context in an easy-to-analyze format, which few reporting entities offer. Integrating this dataset with animal and environmental health data could inform epidemiologic and ecological questions with a One Health perspective. This could encompass research on the drivers of disease transmission, the impacts of climate change, and the mapping of zoonotic spillover risk. In addition, this dataset could be combined with population mobility and flight connectivity data to model the spread of pathogens and quantify the risk of disease importation for different countries. The HED could also be combined with information on disease-specific morbidity to estimate the proportion of individuals likely to miss work or school due to illness.

Caution needs to be applied when using our dataset, as it has some limitations. Although the inclusion criteria were expanded in Phase 2, the types of endemic pathogens included in routine surveillance reports tend to score below our inclusion threshold due to low pathogen prioritization scores. The reliance on open-source disease surveillance reports also means that countries and entities that produce the highest quality data most consistently are necessarily more represented. The HED is also limited by information blind spots as it is dependent on publicly available disease reports. Public health infrastructure, politics, and pathogen characteristics are among the factors that shape country-level infectious disease reporting rates[12].

Code availability

No custom code was used. Software tools used for processing are mentioned in the Methods and Technical Validation sections.

Received: 8 October 2024; Accepted: 18 February 2025;
Published online: 26 February 2025

References

1. Thaivalappil, A., Mascarenhas, M., Waddell, L. A. & Young, I. A qualitative program evaluation of the Publicly Available International Foodborne Outbreak Database. *Can Commun Dis Rep* **47**(1), 59–65, https://doi.org/10.14745/ccdr.v47i01a09 (2021).
2. Smith, K. F. *et al*. Global rise in human infectious disease outbreaks. *J. R. Soc. Interface* **11**, 2014 0950, https://doi.org/10.1098/rsif.2014.0950 (2014).
3. Torres Munguía, J. A., Badarau, F. C., Díaz Pavez, L. R., Martínez-Zarzoso, I. & Wacker, K. M. A global dataset of pandemic-and epidemic-prone disease outbreaks. *Sci. Data* **9**, 683, https://doi.org/10.1038/s41597-022-01797-2 (2022).
4. Messina, J. P. *et al*. A global compendium of human Crimean-Congo haemorrhagic fever virus occurrence. *Sci. Data* **2**, 150016, https://doi.org/10.1038/sdata.2015.16 (2015).
5. Mylne, A. *et al*. A comprehensive database of the geographic spread of past human Ebola outbreaks. *Sci. Data* **1**, 140042, https://doi.org/10.1038/sdata.2014.42 (2014).
6. Carlson, C. J. *et al*. The World Health Organization's Disease Outbreak News: A retrospective database. *PLOS Glob. Public Health* **3**, e0001083, https://doi.org/10.1371/journal.pgph.0001083 (2023).
7. Contarino, F., Bella, F., Di Pietro Ermínio, E., Randazzo, C. & Contrino, M. L. Impact of the COVID-19 pandemic on infectious diseases reporting. *J Prev Med Hyg* **65**(2), E145, https://doi.org/10.15167/2421-4248/jpmh2024.65.2.3197 (2024).
8. Badker, R. *et al*. Challenges in reported COVID-19 data: best practices and recommendations for future epidemics. *BMJ Glob. Health* **6**, e005542, https://doi.org/10.1136/bmjgh-2021-005542 (2021).
9. Badker, R., Kipperman, N., Ash, B. & Pardee, C. Constructing a global human epidemic database using open-source digital biosurveillance. *Ginkgo Biosecurity Figshare* https://doi.org/10.6084/m9.figshare.28187936 (2025).
10. Meadows, A. J., Stephenson, N., Madhav, N. K. & Oppenheim, B. Historical trends demonstrate a pattern of increasingly frequent and severe epidemics of high-consequence zoonotic viruses. *BMJ Global Health* **8.11**, e012026, https://doi.org/10.1136/bmjgh-2023-012026 (2023).
11. Oppenheim, B. *et al*. Assessing global preparedness for the next pandemic: development and application of an Epidemic Preparedness Index. *BMJ Glob. Health* **4**, https://doi.org/10.1136/bmjgh-2018-001157 (2019).
12. Meadows, A. J. *et al*. Infectious Disease Underreporting Is Predicted by Country-Level Preparedness, Politics, and Pathogen Severity. *Health Secur* **20**, 331–338, https://doi.org/10.1089/hs.2021.0197 (2022).
13. World Bank. Atlas of Sustainable Development Goals 2017: From World Development Indicators. World Bank Atlas. Washington, DC: World Bank. http://hdl.handle.net/10986/26306 License: CC BY 3.0 IGO (2017).

Acknowledgements

The authors would like to thank Tanya Philippsen, Chris Ngoon, Sarah Barthel, Damien Joly, Kierste Miller, and Junhua Hu who contributed to the data collection and quality control of the HED and Swati Sureka, who provided editorial support.

Author contributions

All authors reviewed the manuscript at various points prior to submission. Conception and design: R.B., N.K., & C.P. Data acquisition, entry, coding, and analysis: B.A., R.B., N.K., C.P., & P.S. Reviewing analysis and editing: N.K.M., N.S., & B.O.

Competing interests

All authors are or have been employed by Ginkgo Bioworks. Inc and may have an equity stake; no other relationships or activities that could appear to have influenced the submitted work.

Additional information

Correspondence and requests for materials should be addressed to R.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.