Data Descriptor

A Dataset of Search Interests Related to Disease X Originating from Different Geographic Regions

Nirmalya Thakur *, Kesha A. Patel , Isabella Hall , Yuvraj Nihal Duggal , Shuqi Cui

Posted Date: 24 August 2023

doi: 10.20944/preprints202308.1701.v1

Keywords: disease X; big data; data science; data analysis; dataset development; database; google trends; data mining; healthcare; epidemiology

Data Descriptor

A Dataset of Search Interests Related to Disease X Originating from Different Geographic Regions

Nirmalya Thakur [1,*], Kesha A. Patel [2], Isabella Hall [3], Yuvraj Nihal Duggal [1] and Shuqi Cui [1]

[1] Department of Computer Science, Emory University, Atlanta, GA 30322, USA; nirmalya.thakur@emory.edu (N.T.) yuvraj.nihal.duggal@emory.edu (Y.N.D), nicole.cui@emory.edu (S.C.)
[2] Department of Mathematics, Emory University, Atlanta, GA 30322; kesha.patel@emory.edu
[3] Department of Computer Science, University of Cincinnati, Cincinnati, OH 45221; hallib@mail.uc.edu
* Correspondence: nirmalya.thakur@emory.edu

Abstract: The World Health Organization (WHO) added Disease X to their shortlist of blueprint priority diseases to represent a hypothetical, unknown pathogen that could cause a future epidemic. During different virus outbreaks of the past, such as COVID-19, Influenza, Lyme Disease, and Zika virus, researchers from various disciplines utilized Google Trends to mine multimodal components of web behavior to study, investigate, and analyze the global awareness, preparedness, and response associated with these respective virus outbreaks. As the world prepares for Disease X, a dataset on web behavior related to Disease X would be crucial to contribute towards the timely advancement of research in this field. Furthermore, none of the prior works in this field have focused on the development of a dataset to compile relevant web behavior data, which would help to prepare for Disease X. To address these research challenges, this work presents a dataset of web behavior related to Disease X, which emerged from different geographic regions of the world, between February 2018 to August 2023. Specifically, this dataset presents the search interests related to Disease X from 94 geographic regions. These regions were chosen for data mining as these regions recorded significant search interests related to Disease X during this timeframe. The dataset was developed by collecting data using Google Trends. The relevant search interests for all these regions for each month in this time range are available in this dataset. This paper also discusses the compliance of this dataset with the FAIR principles of scientific data management. Finally, a brief analysis of specific features of this dataset is presented to uphold the applicability, relevance, and usefulness of this dataset for the investigation of different research questions in the interrelated fields of Big Data, Data Mining, Healthcare, Epidemiology, and Data Analysis.

Dataset: https://dx.doi.org/10.21227/ht7f-rx42

Keywords: disease X; big data; data science; data analysis; dataset development; database; google trends; data mining; healthcare; epidemiology

1. Introduction

In the recent past, several viruses such as COVID-19 [1], the plague [2], Spanish Flu [3], HIV [4], and Ebola [5] have rampaged unopposed across different countries, infecting and leading to the demise of people, destruction of political regimes, affecting various sectors of the global economy, as well as causing financial and psychosocial burdens, the likes of which the world has not witnessed in centuries [6]. As a response to this, various organizations and policy-making bodies on a global scale have begun investigating approaches to learn from such virus outbreaks with an aim to not repeat the mistakes of the past during future virus outbreaks. "Disease X" is a placeholder name that was adopted by the World Health Organization (WHO) in February 2018 on their shortlist of blueprint priority diseases to represent a hypothetical, unknown pathogen that could cause a future epidemic [7,8]. The WHO used the placeholder term "Disease X" to make sure that its planning (such as relevant tests, expanded vaccinations, and production capabilities for vaccines) was robust,

versatile, and equipped to deal with an unidentified virus [9]. The idea of Disease X, according to Anthony Fauci (the director of the US National Institute of Allergy and Infectious Diseases at that time), was to motivate WHO's investigations on entire classes of viruses rather than just specific strains of certain viruses, with an aim to strengthen WHO's preparedness of dealing with such outbreaks [10].

Thus, it is crucial to plan and adopt a holistic approach to prevent and predict a new pandemic in the future. Prior works [11-14] in this field have discussed various means by which Disease X might start. For instance, the potential of deadly pathogens being released from melting glaciers could start a new pandemic. Alternatively, with the continual increase of global warming and climate changes, viruses dormant at present may become active and mutate and lead to the next pandemic. Furthermore, human and animal contact has become increasingly common, and the lack of proper protocols in this regard has led to the outbreak of zoonotic viruses in the past. A well-known example of this would be H1N1 which contained genetic material from human, avian, and swine origin, involving wildlife, pig farming, animal movement, and farm workers [15]. Therefore, in the last couple of years or so, works in this field have also focused on predicting what type of pathogen might be responsible for Disease X, with an aim to create, implement, and evaluate countermeasures that would help control the potential pandemic at a faster rate than previous pandemics such as COVID-19 [16-18]. Simpson et al. [19] stated that Disease X is likely to occur due to one or more of these risk factors - human interactions with wildlife, the production of goods derived from animals with minimal oversight of workers and an unclear supply chain, bug and tick vectors, extremely high population densities, and limited surveillance and laboratory capacities. This work by Simpson et al. [19] also states that Disease X will probably be caused by the zoonotic spread of a highly infectious RNA virus from a region where the confluence of risk factors and population dynamics will lead to prolonged person-to-person transmission.

A highly agreed upon aspect related to Disease X within the research community in this field is that the world is currently not prepared with the applicable countermeasures, policies, and procedures that would be necessary to control and contain this virus. There are multiple factors that we need to take into consideration when creating new response, control, and preparation measures, including vaccine development and distribution, country and state responses, political stances, as well as cultural and environmental factors. It is crucial that there is global preparation, coordination, and communication such that each of these factors is considered and managed in coordination with other factors to allow for controlling and containing a new pandemic [20].

One of the overarching issues that were observed on a global scale when attempting to handle the COVID-19 pandemic was the lack of efficiency, coordination, agreement, and organization related to the production and distribution of vaccines and COVID-19 tests in a timely manner [21]. While various organizations and labs were working to create a vaccine, it seemed that different countries were scrambling to even put up testing centers and mass produce enough COVID-19 tests. It took far longer than ideal to ensure easy access to COVID-19 tests which allowed the COVID-19 virus to continue to spread at an alarming rate because symptoms were not guaranteed to be noticeable in all population groups [22]. Testing is one of the first lines of defense against viruses because the threshold of the spread can be determined, and suitable actions can be taken depending on the positive cases that are reported. This was an issue with the supply chain and communication across agencies during the outbreak and rapid spread of COVID-19. Those same supply chain issues were reported when trying to roll out the COVID-19 vaccines at an even slower rate than the tests. Research labs in different geographic regions seemed less prepared to mass produce and distribute the vaccines, which slowed down response rates and did not contain the spread of COVID-19 in a timely manner [23, 24]. Another issue associated with the COVID-19 pandemic was the lack of coordination and cooperation between countries in their responses [25]. During the outbreak of COVID-19, some countries implemented measures (such as partial or complete lockdowns) immediately, while others did not implement such measures at the same pace [26,27]. Finally, a major issue specifically seen during the COVID-19 pandemic was political stances standing in the way of scientific progress. There was a lot of misinformation surrounding the entirety of the pandemic. That ranged from the

 doi:10.20944/preprints202308.1701.v1

effectiveness of vaccines, the safety of the vaccines, the accuracy of the test results, approaches for treatment, and the severity of the virus. [28, 29].

During the outbreak of COVID-19 and similar virus outbreaks of the past, Google Trends attracted a significant amount of attention from researchers across different disciplines, such as Big Data, Data Mining, Healthcare, Epidemiology, Information Retrieval, and Data Analysis, as Google Trends helps to mine, analyze, and obtain real-time insights related to web behavior and the features of Google Trends surpass traditional surveys [30]. In the last few years, Google Trends has been highly popular for researchers in Healthcare for analysis of different patterns of web behavior related to different virus outbreaks [31-37]. Ginsberg et al. [38] discussed the significance of seasonal influenza and the potential threat of a pandemic caused by a new strain of the influenza virus using Google Trends. The work proposed a method to enhance early disease detection by monitoring Google search queries, which reflected health-seeking behavior. By analyzing Google search queries, the researchers accurately estimated weekly cases of influenza in different regions of the United States, allowing for rapid detection and response to influenza with only a one-day reporting lag. The work by Kapiány-Fövény et al. [39] focused on analyzing Google search volumes using Google Trends to forecast Lyme disease incidences. By integrating Google Trends data into a seasonal autoregressive moving average (SARIMA) model, the researchers compared their predictions with the actual reported values for Lyme disease incidence in Germany. The objective of the work done by Verma et al. [40] was to assess the potential of using Google Trends data for predicting disease outbreaks. Focusing on diseases like malaria, dengue fever, chikungunya, and enteric fever in two regions in India - Chandigarh and Haryana, the research compared Google Search trends with Integrated Disease Surveillance Programme (IDSP) data. The analysis revealed a temporal correlation between the two datasets, particularly with a lag of 2 to 3 weeks for chikungunya and dengue fever, indicating the feasibility of utilizing Google Trends for predicting disease outbreaks at both local and regional levels. Young et al. [41] explored the potential of using relevant Google Search queries from Google Trends to monitor and predict syphilis cases at a state level. The study investigated the relationship between weekly reported syphilis cases and online search activity related to risk factors. By employing linear mixed models, the study established associations between search query data and syphilis cases, achieving accurate predictions for a significant number of weeks. The results indicated a strong correlation between web behavior and reported syphilis cases, suggesting the feasibility of integrating such data into public health monitoring systems for disease surveillance and prediction. Another work by Young et al. [42] focused on utilizing Google search data to monitor and predict new HIV diagnosis cases in the United States. The researchers collected HIV-related search volume data and state-level new HIV diagnoses data using Google Trends. They developed a predictive model using significant predictor keywords identified through LASSO and combined this data with actual HIV case reports from the CDC. The model demonstrated strong predictive capabilities, achieving an average R^2 value of 0.99 and an average root-mean-square error (RMSE) of 108.75 when comparing predicted and actual HIV cases. Morsy et al. [43] focused on predicting Zika virus cases using Google search queries from Google Trends. The researchers developed a prediction model based on time-series regression (TSR) that utilized Zika search volume from Google Trends to anticipate confirmed Zika cases in Brazil and Colombia. The model with a 1-week lag of Zika query and a 1-week lag of Zika cases as a control for autocorrelation was found to be the most effective in predicting Zika cases. The results demonstrated the potential to forecast Zika cases a week ahead of outbreaks, offering healthcare authorities an early indicator for outbreak evaluation and precautionary measures. Using Google Trends, Ortiz-Martínez et al. [44] showed that there was a high correlation between the COVID-19 incidence in Colombia and Google searches on COVID-19 in Colombia ($R^2 = 0.8728$ and $p < 0.0001$). Therefore, it may be concluded that prior works in this field have focused on using Google Trends related to mining, analysis, and investigation of multimodal components of web behavior during various virus outbreaks. However, these works have two major limitations. First, none of these works focused on Disease X, which features in the shortlist of blueprint priority diseases of WHO. Second, these works focused on the analysis of the relevant data from Google Trends from a very limited number of regions. To address these limitations and to

contribute to the timely advancement of research and development in this field, this work presents a dataset that comprises web behavior data related to Disease X that emerged from 94 regions from February 2018 to August 2018. These 94 regions were selected for the development of this dataset as all these regions recorded a significant level of interest towards Disease X during this timeframe. This dataset was developed by collecting this data from Google Trends. The rest of this paper is organized as follows. Section 2 presents the detailed methodology which was followed for the development of this dataset. The dataset is described in Section 3. Section 3 also presents a brief analysis of specific features of this dataset to uphold its applicability, relevance, and significance for the investigation of different research questions. Section 4 concludes the paper, which is followed by references.

2. Methodology

Google Trends [45], a tool developed by Google, allows the mining and analysis of real-time and historical information associated with Google search queries, enabling researchers to uncover valuable insights into the interests of individuals across different domains and topics [46]. Google Trends analyzes aggregate search behavior by considering searches on Google and can thus provide unique insights associated with web-behavior. This feature is particularly valuable in health informatics, where understanding public engagement and interests in health-related topics and predicting disease outbreaks is of paramount importance [47].

The real-time data availability of Google Trends makes it superior to traditional survey methods, and it is also far less time-consuming. Additionally, as the web behavior data available via Google Trends is anonymous, it allows researchers to explore different forms of data analysis that might have been otherwise difficult due to privacy concerns of the general public [47]. Google Trends presents several significant advantages over traditional survey methods, positioning it as a potent tool for research and analysis of multimodal characteristics of web-behavior. The foremost advantage lies in the cost-effectiveness of utilizing Google Trends. Unlike traditional surveys, which frequently entail significant expenses for participant recruitment, data collection, and analysis, Google Trends operates as a cost-free resource. This financial flexibility allows researchers to channel resources into more focused areas of investigation or allocate them toward enhancing the research process itself, promoting greater flexibility in research endeavors. Another key advantage centers around the breadth and diversity of the data captured by Google Trends. Conducting regular surveys on a global scale is a logistical challenge, often constrained by geographic and demographic limitations. However, Google Trends seamlessly aggregates web behavior data on a global scale which can be used for in-depth study and analysis. This global perspective of Google Trends enhances the generalizability of findings and facilitates cross-cultural comparisons, making it a valuable resource for understanding the intricacies of web behavior across different geographic regions. Moreover, the near real-time nature of data availability on Google Trends is a game-changer. Google Trends offers almost immediate access to search trends as they unfold, providing researchers with timely access to evolving interests and trends. This swift access to information enables timely analysis, decision-making, and trend detection, making it particularly advantageous in fields that require quick response, such as public health and policy formulation. In contrast, traditional surveys often grapple with time delays, influenced by the labor-intensive nature of participant recruitment and adherence to inclusion criteria. The delays inherent in survey-based research can hinder the ability to capture real-time insights, potentially impacting the accuracy and relevancy of the findings. The instant accessibility of Google Trends data addresses this limitation, empowering researchers with the agility to adapt and react promptly to emerging trends or shifts in user interests related to a topic as evidenced by relevant web-behavior.

Google Trends presents the frequency at which a specific search term is input into Google's search engine relative to the overall search volume on the site during a specific time frame. Mathematically, if $n(q, l, t)$ represents the number of searches for the query q in the location l during the period t, the relative popularity (RP) of the query is computed as shown in Equation (1). In Equation (1), $Q(l,t)$ is a set of all the queries made from location l at time t, $\Pi(n(q,l,t)>\tau)$ is a dummy variable with value 1 when $n(q, l, t) > \tau$ (Query is popular) and 0 otherwise. The resulting numbers

 doi:10.20944/preprints202308.1701.v1

are then scaled within the range of 0 to 100 based on the proportion of the topic relative to the total number of search topics. This defines the Google Trends Index (GTI) as shown in Equation (2).

$$RP_{(q,l,t)} = \frac{n_{(q,l,t)}}{\Sigma_{q\epsilon Q(l,t)} n(q,l,t)} \times \Pi_{(n(q,l,t)>\tau)} \quad (1)$$

$$\boldsymbol{GTI_{(q,l,t)}} = \frac{\boldsymbol{RP(q,l,t)}}{\boldsymbol{max\{RP(q,l,t)_{t\epsilon 1,2,\ldots,T}\}}} \times \mathbf{100} \quad (2)$$

These index values can be generated by Google Trends starting from January 1, 2004, up to 36 hours prior to the present search. Google Trends excludes search data from very limited users and highlights popular search topics while assigning 0 to terms with low search volumes [48]. The following is an overview of the features of Google Trends:

- Search Term Trends: This feature allows users to see how the popularity of a specific search term or keyword has changed over time. Google Trends provides a graphical representation to highlight these trends.
- Related Queries: Google Trends displays related queries that are frequently searched alongside the user's primary search term. This can help identify related topics or terms relevant for data analysis.
- Regional Interest: Users can view the geographical regions where a specific search term is most popular using Google Trends. Google Trends provides insights into regional differences in search interest for search terms.
- Trending Searches: This feature of Google Trends highlights the current and popular search queries or topics, providing real-time insights into what people are searching for on Google.
- Year in Search: Google Trends often releases a "Year in Search" report summarizing the top search queries from the past year. In this report, it offers an overview of significant events and trends.
- Category Comparison: Users can compare the search interest of different categories or topics on Google using Google Trends. This can be useful for understanding the relative popularity of various topics.
- Time Period Selection: Google Trends allows users to specify the time period for which they desire to query and analyze the data. This can range from a few hours to multiple years.
- Data Visualization: Google Trends provides interactive charts and graphs to visualize search data.
- Real-Time Data: Google Trends often updates in near real-time, making it valuable for tracking ongoing events.
- Data Export: Google Trends allows different options to export data related to search interests, related queries, and related topics for a search term on Google for further analysis.

For developing this dataset, the web behavior data in terms of search interests related to Disease X (as a topic) was collected using Google Trends from February 2018 to August 2023. February 2018 was selected as the start time as WHO added Disease X to their shortlist of blueprint priority diseases in February 2018. August 2023 was the most recent month at the time of data collection. First, the global search trends related to Disease X (as a topic) during this timeframe (February 2018 to August 2023) were analyzed using Google Trends. The result provided by Google Trends is shown in Figure 1. Thereafter, by using the "Regional Interest" feature of Google Trends, the list of regions that recorded significant search interests related to Disease X was compiled and exported. This list of regions is shown in Table 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1701.v1

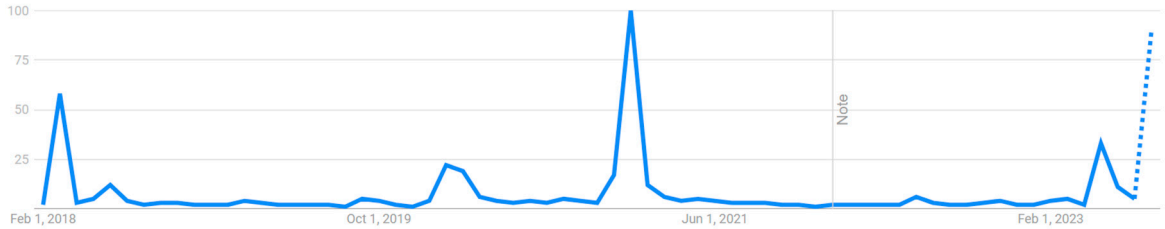

Figure 1. Trends in Search Interests related to Disease X (as a topic) on a Global Scale between February 2018 to August 2023.

Table 1. List of 94 regions that recorded significant search interests related to Disease X (as a topic) between February 2018 to August 2023.

List of Regions
Singapore, Haiti, Honduras, El Salvador, Madagascar, Panama, Bolivia, Reunion, Guatemala, Cuba, United Arab Emirates, Paraguay, Nicaragua, Hong Kong, Macao, Qatar, United Kingdom, Brunei, Ecuador, Uruguay, Oman, Bahrain, Ireland, Kuwait, Costa Rica, Argentina, India, Puerto Rico, Venezuela, France, St. Helena, Brazil, Mexico, Côte d'Ivoire, Peru, Canada, Australia, Zimbabwe, Colombia, United States, Luxembourg, Lebanon, Ghana, Algeria, New Zealand, Portugal, Malaysia, Myanmar (Burma), Ethiopia, Dominican Republic, China, Chile, Nepal, Belgium, Iraq, Taiwan, South Africa, Tunisia, Sri Lanka, Thailand, Switzerland, Spain, Bangladesh, Saudi Arabia, Kenya, South Korea, Germany, Norway, Pakistan, Indonesia, Hungary, Morocco, Austria, Israel, Nigeria, Bulgaria, Philippines, Netherlands, Denmark, Greece, Italy, Jordan, Egypt, Sweden, Finland, Czechia, Romania, Poland, Iran, Türkiye, Russia, Vietnam, Ukraine, Japan

Thereafter, by utilizing Google Trends as the data source, search interests related to Disease X (as a topic) for all these 94 regions between February 2018 and August 2023 were collected and exported as .CSV files. To consolidate the 94 .CSV files into one workbook on Microsoft Excel, the Power Query interface on Excel was employed. The Power Query tool uses each individual file as a data source and imports each file's data onto the Excel Workbook. Each region's search interest for "Disease X" is present on distinct sheets in this file which was uploaded to IEEE Dataport [49] as a dataset. The flowchart in Figure 2 shows the step-by-step process that was followed for the development of this dataset. This dataset is described in Section 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1701.v1

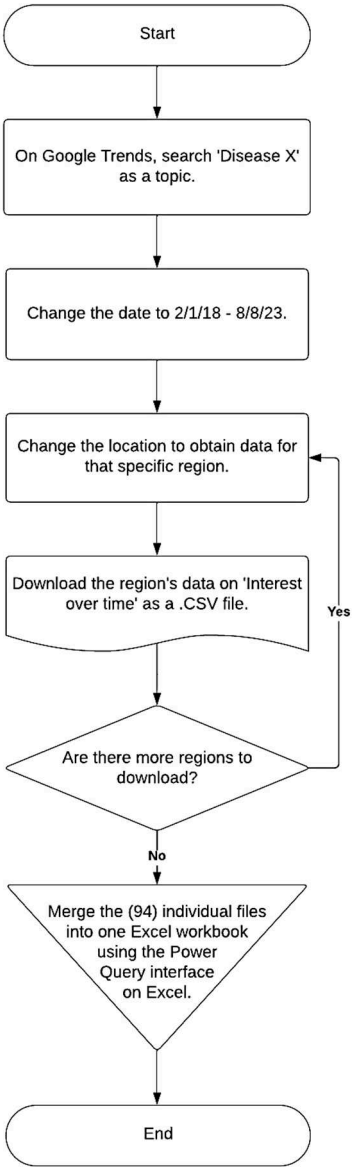

Figure 2. A flowchart to represent the step-by-step process of the development of this dataset.

3. Data Description and Analysis

This section describes the dataset, which is available at https://dx.doi.org/10.21227/ht7f-rx42. This dataset contains one Microsoft Excel workbook that comprises 94 different sheets where each sheet presents the search interests related to Disease X (as a topic) between February 2018 to August 2023 for a different region. The search interest data for all the regions stated in Table 1 is available in this dataset. For each region, this dataset presents the search interests related to Disease X (as a topic) for each month in this timeframe, i.e., from February 2018 to August 2023. This data can be analyzed to obtain the trends in search interests during this timeframe for each of these 94 regions. For instance, the analysis of this data for the United States is presented in Figure 3. In this Figure, the X-axis represents the months, and the Y-axis represents the search interest related to Disease X on a scale of 0 to 100.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1701.v1

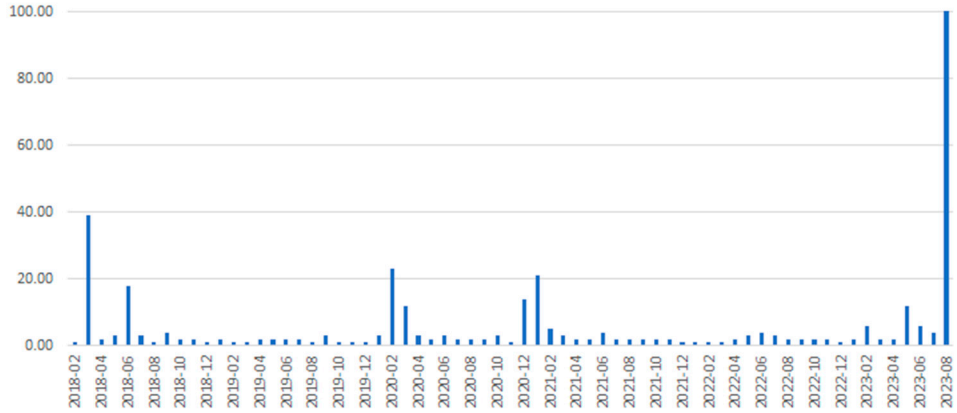

Figure 3. Trends in Search Interests related to Disease X (as a topic) for the United States between February 2018 to August 2023.

This analysis of this data for the United States shows that the search interest related to Disease X has been the highest in August 2023. Similar trends and insights associated with search interests for Disease X emerging from different geographic regions can be obtained from analysis of the search interest data for that region as available in this dataset. Figure 4 shows a world map-based analysis of the search interests related to Disease X for all 94 regions during this timeline. The intensity of the color in Figure 4 represents the value of search interest related to Disease X from a certain region. So, a region that recorded a very high value of search interest related to Disease X during this timeframe is indicated by a darker shade of the color blue as compared to a region that recorded a very low value of search interest related to Disease X during this timeline. This analysis shows that the top 10 regions that recorded the highest search interests related to Disease X during this timeframe are Singapore, Honduras, Haiti, Nicaragua, Guatemala, El Salvador, Brunei, Panama, Cuba, and the United Arab Emirates. Furthermore, this analysis also helps to infer the list of regions that recorded the least (but significant) search interests related to Disease X during this timeframe. The list of 10 regions that recorded the least (but significant) search interests related to Disease X during this timeframe is Finland, Romania, Czechia, Ukraine, Poland, Türkiye, Vietnam, Iran, Russia, and Japan.

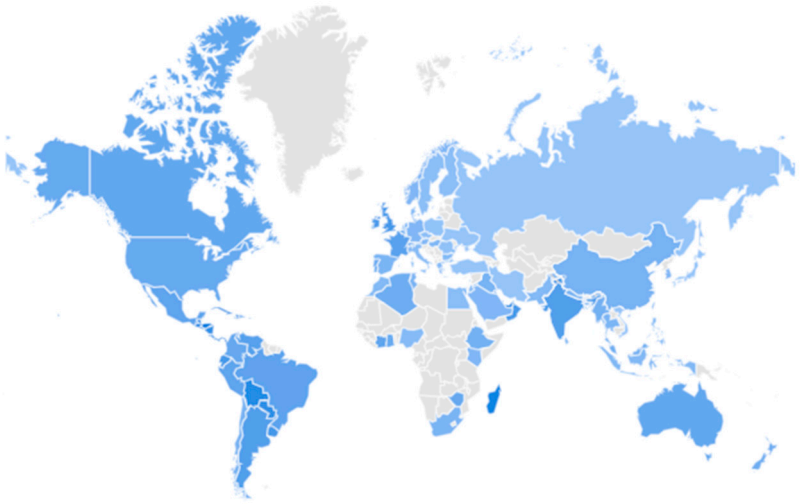

Figure 4. A world map-based analysis of the search interests related to Disease X for different regions from February 2018 to August 2023.

During the development of this dataset, it was observed that online searches on Google related to Disease X during this timeframe (February 2018 to August 2023) had several related queries. The 'rising' keywords associated with these related queries were collected by using the "Related Queries"

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 August 2023 doi:10.20944/preprints202308.1701.v1

feature of Google Trends, as described in Section 2. Figure 5 shows a word cloud-based representation of these queries related to Disease X during this timeframe. In this context, it is worth mentioning that the mining of the data from Google Trends for the development of this dataset was performed on August 8, 2023. Google Trends provided the search interest for August 2023 for each of the 94 regions by taking into account the relevant Google Searches recorded from August 1, 2023, to August 8, 2023. So, if the data collection is performed once again at the end of August 2023 or at a later date using Google Trends, it is possible that the search interest for August 2023 for some of these regions might change as Google Trends would then report the search interest value for August 2023 by taking into account all relevant Google Searches recorded from August 1, 2023, to August 31, 2023.

Figure 5. A word cloud-based representation of 'rising' queries related to Disease X from February 2018 to August 2023.

In the remainder of this section, the compliance of this dataset with the FAIR principles of Scientific Data Management [50] is explained. The FAIR principles represent a vital framework crafted to amplify the accessibility and utility of scientific data and research outcomes. The acronym "FAIR" encapsulates four fundamental principles of scientific data management: Findability, Accessibility, Interoperability, and Reusability. These principles underscore the significance of making data effortlessly discoverable, openly accessible, compatible with other datasets, and comprehensively documented for the sake of reproducibility. Essentially, the FAIR principles endeavor to cultivate a more cooperative and transparent research landscape, facilitating the exchange of knowledge and bolstering the lasting influence of scientific investigations related to database development and database management. Several prior works in the field of dataset development have discussed how the developed datasets such as - the human metabolome database for 2022 [51], WikiPathways dataset [52], datasets of Tweets about COVID-19 [53,54], a dataset of Tweets about MPox [55], computational 2D materials database (C2DB) [56], the open reaction database [57], RCSB Protein Data Bank [58], and the PHI-base: pathogen–host interactions database [59], just to name a few, complied with the FAIR principles of scientific data management.

This dataset, available at https://dx.doi.org/10.21227/ht7f-rx42, is findable as it has a unique and permanent DOI assigned by IEEE Dataport. This DOI can be used by researchers from any discipline to find this dataset online. This dataset satisfies the accessibility property as it can be accessed by any user on the internet using any device via the DOI, as long as the user's device is connected to the internet and is operating in a desired manner. The dataset is interoperable as the data in this dataset is available in a standard format (.xlsx file) that can be downloaded, read, and analyzed across

different computer systems, frameworks, and applications. Finally, this dataset satisfies the reusability property as the data can be re-used any number of times for the study and investigation of different research questions that focus on the analysis of search interests related to Disease X.

4. Conclusions

The World Health Organization (WHO) added "Disease X" to their shortlist of blueprint priority diseases to represent a hypothetical, unknown pathogen that could cause a future epidemic. Since then, several works in this field have analyzed virus outbreaks of the past to propose approaches, methodologies, principles, and guidelines for better awareness, preparedness, and response towards Disease X. Many of these works that focused on analyzing virus outbreaks of the past such as COVID-19, Influenza, Lyme Disease, and Zika virus, utilized Google Trends to mine and analyze multimodal components of web-behavior. However, two primary limitations exist in these works. First, these works did not specifically focus on Disease X. Second, many of these works focused on the analysis of Google Trends data originating from a very limited number of geographic regions. To address these limitations and to contribute towards the timely advancement of research in this field, this work presents a dataset of search interests related to Disease X (as a topic) originating from 94 regions of the world between February 2018 to August 2023. These 94 regions were selected for the development of this dataset as all these regions recorded a significant level of interest towards Disease X during this timeframe. The dataset is available at https://dx.doi.org/10.21227/ht7f-rx42. In this dataset, for every region, the search interest related to Disease X is available for each month during this timeframe. The dataset complies with the FAIR principles of scientific data management. This paper also presents a brief analysis of specific features of this dataset to uphold its relevance and usefulness for the investigation of different research questions in the interrelated fields of Big Data, Data Mining, Healthcare, Epidemiology, Information Retrieval, and Data Analysis. As per the best knowledge of the authors, no similar work in this field has been done so far. Future work in this area would involve analyzing the specific trends of search interests related to Disease X across different geographic regions to determine and investigate specific similarities or dissimilarities of those trends.

Author Contributions: Conceptualization, N.T.; methodology, N.T., K.A.P, and Y.N.D.; software, N.T., K.A.P, and Y.N.D.; validation, N.T.; formal analysis, N.T.; investigation, N.T.; resources, N.T.; data curation, N.T., and K.A.P; writing—original draft preparation, N.T., I.H., K.A.P, Y.N.D. and S.Q.; writing—review and editing, N.T. and I.H.; visualization, N.T.; supervision, N.T.; project administration, N.T.; funding acquisition, Not Applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This work resulted in the creation of a dataset that is available at https://dx.doi.org/10.21227/ht7f-rx42, as per the CC BY 4.0 License.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fauci, A.S.; Lane, H.C.; Redfield, R.R. Covid-19 — Navigating the Uncharted. *N. Engl. J. Med.* **2020**, *382*, 1268–1269, doi:10.1056/nejme2002387.
2. Prentice, M.B.; Rahalison, L. Plague. *Lancet* **2007**, *369*, 1196–1207, doi:10.1016/s0140-6736(07)60566-2.
3. Aassve, A.; Alfani, G.; Gandolfi, F.; Le Moglie, M. Epidemics and Trust: The Case of the Spanish Flu. *Health Econ.* **2021**, *30*, 840–857, doi:10.1002/hec.4218.
4. Joint United Nations Programme on HIV/AIDS.; World Health Organization *2008 Report on the Global AIDS Epidemic*; World Health Organization: Genève, Switzerland, 2008; ISBN 9789291737116.
5. Jacob, S.T.; Crozier, I.; Fischer, W.A., II; Hewlett, A.; Kraft, C.S.; Vega, M.-A. de L.; Soka, M.J.; Wahl, V.; Griffiths, A.; Bollinger, L.; et al. Ebola Virus Disease. *Nat. Rev. Dis. Primers* **2020**, *6*, 1–31, doi:10.1038/s41572-020-0147-3.

6. Sampath, S.; Khedr, A.; Qamar, S.; Tekin, A.; Singh, R.; Green, R.; Kashyap, R. Pandemics throughout the History. *Cureus* **2021**, *13*, doi:10.7759/cureus.18136.
7. Chatterjee, P.; Nair, P.; Chersich, M.; Terefe, Y.; Chauhan, A.; Quesada, F.; Simpson, G. One Health, "Disease X" & the Challenge of "Unknown" Unknowns. *Indian J. Med. Res.* **2021**, *153*, 264, doi:10.4103/ijmr.ijmr_601_21.
8. Prioritizing Diseases for Research and Development in Emergency Contexts Available online: https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts (accessed on 17 August 2023).
9. Barnes, T. World Health Organisation Fears New "Disease X" Could Cause a Global Pandemic Available online: https://www.independent.co.uk/news/science/disease-x-what-is-infection-virus-world-health-organisation-warning-ebola-zika-sars-a8250766.html (accessed on 17 August 2023).
10. Scutti, S. World Health Organization Gets Ready for 'Disease X' Available online: https://www.cnn.com/2018/03/12/health/disease-x-blueprint-who/index.html (accessed on 17 August 2023).
11. Adalja, A.A.; Watson, M.; Toner, E.S.; Cicero, A.; Inglesby, T.V. Characteristics of Microbes Most Likely to Cause Pandemics and Global Catastrophes. In *Current Topics in Microbiology and Immunology*; Springer International Publishing: Cham, 2019; pp. 1–20 ISBN 9783030363109.
12. Kreuder Johnson, C.; Hitchens, P.L.; Smiley Evans, T.; Goldstein, T.; Thomas, K.; Clements, A.; Joly, D.O.; Wolfe, N.D.; Daszak, P.; Karesh, W.B.; et al. Spillover and Pandemic Properties of Zoonotic Viruses with High Host Plasticity. *Sci. Rep.* **2015**, *5*, 1–8, doi:10.1038/srep14830.
13. Jones, K.E.; Patel, N.G.; Levy, M.A.; Storeygard, A.; Balk, D.; Gittleman, J.L.; Daszak, P. Global Trends in Emerging Infectious Diseases. *Nature* **2008**, *451*, 990–993, doi:10.1038/nature06536.
14. Carlson, C.J.; Albery, G.F.; Merow, C.; Trisos, C.H.; Zipfel, C.M.; Eskew, E.A.; Olival, K.J.; Ross, N.; Bansal, S. Climate Change Increases Cross-Species Viral Transmission Risk. *Nature* **2022**, *607*, 555–562, doi:10.1038/s41586-022-04788-w.
15. Peiris, J.S.M.; Tu, W.-W.; Yen, H.-L. A Novel H1N1 Virus Causes the First Pandemic of the 21stCentury. *Eur. J. Immunol.* **2009**, *39*, 2946–2954, doi:10.1002/eji.200939911.
16. Van Kerkhove, M.D.; Ryan, M.J.; Ghebreyesus, T.A. Preparing for "Disease X." *Science* **2021**, *374*, 377–377, doi:10.1126/science.abm7796.
17. Iserson, K. The next Pandemic: Prepare for "Disease X." *West. J. Emerg. Med.* **2020**, *21*, 756, doi:10.5811/westjem.2020.5.48215.
18. Tahir, M.J.; Sawal, I.; Essar, M.Y.; Jabbar, A.; Ullah, I.; Ahmed, A. Disease X: A Hidden but Inevitable Creeping Danger. *Infect. Control Hosp. Epidemiol.* **2022**, *43*, 1758–1759, doi:10.1017/ice.2021.342.
19. Simpson, S.; Kaufmann, M.C.; Glozman, V.; Chakrabarti, A. Disease X: Accelerating the Development of Medical Countermeasures for the next Pandemic. *Lancet Infect. Dis.* **2020**, *20*, e108–e115, doi:10.1016/s1473-3099(20)30123-7.
20. Simpson, S.; Chakrabarti, A.; Robinson, D.; Chirgwin, K.; Lumpkin, M. Navigating Facilitated Regulatory Pathways during a Disease X Pandemic. *NPJ Vaccines* **2020**, *5*, 1–9, doi:10.1038/s41541-020-00249-5.
21. Radanliev, P.; De Roure, D. Disease X Vaccine Production and Supply Chains: Risk Assessing Healthcare Systems Operating with Artificial Intelligence and Industry 4.0. *Health Technol. (Berl.)* **2023**, *13*, 11–15, doi:10.1007/s12553-022-00722-2.
22. Singh, R.; Sarsaiya, S.; Singh, T.A.; Singh, T.; Pandey, L.K.; Pandey, P.K.; Khare, N.; Sobin, F.; Sikarwar, R.; Gupta, M.K. Corona Virus (COVID-19) Symptoms Prevention and Treatment: A Short Review. *J. Drug Deliv. Ther.* **2021**, *11*, 118–120, doi:10.22270/jddt.v11i2-s.4644.
23. Kooli, C. COVID-19: Public Health Issues and Ethical Dilemmas. *Ethics Med. Public Health* **2021**, *17*, 100635, doi:10.1016/j.jemep.2021.100635.
24. Golan, M.S.; Jernegan, L.H.; Linkov, I. Trends and Applications of Resilience Analytics in Supply Chain Modeling: Systematic Literature Review in the Context of the COVID-19 Pandemic. *Environ. Syst. Decis.* **2020**, *40*, 222–243, doi:10.1007/s10669-020-09777-w.
25. Schuerger, C.; Batalis, S.; Quinn, K.; Adalja, A.; Puglisi, A. Viral Families and Disease X: A Framework for U.s. Pandemic Preparedness Policy Available online: https://cset.georgetown.edu/wp-content/uploads/CSET-Viral-Families-and-Disease-X-A-Framework-for-U.S.-Pandemic-Preparedness-Policy.pdf (accessed on 17 August 2023).
26. Fontanet, A.; Cauchemez, S. COVID-19 Herd Immunity: Where Are We? *Nat. Rev. Immunol.* **2020**, *20*, 583–584, doi:10.1038/s41577-020-00451-5.
27. Frederiksen, L.S.F.; Zhang, Y.; Foged, C.; Thakur, A. The Long Road toward COVID-19 Herd Immunity: Vaccine Platform Technologies and Mass Immunization Strategies. *Front. Immunol.* **2020**, *11*, doi:10.3389/fimmu.2020.01817.
28. Kiviniemi, M.T.; Orom, H.; Hay, J.L.; Waters, E.A. Prevention Is Political: Political Party Affiliation Predicts Perceived Risk and Prevention Behaviors for COVID-19. *BMC Public Health* **2022**, *22*, doi:10.1186/s12889-022-12649-4.

29. Rabin, C.; Dutra, S. Predicting Engagement in Behaviors to Reduce the Spread of COVID-19: The Roles of the Health Belief Model and Political Party Affiliation. *Psychol. Health Med.* **2022**, *27*, 379–388, doi:10.1080/13548506.2021.1921229.
30. Thakur, N.; Han, C.Y. Country-Specific Interests towards Fall Detection from 2004–2021: An Open Access Dataset and Research Questions. *Data (Basel)* **2021**, *6*, 92, doi:10.3390/data6080092.
31. Adalja, A.A.; Watson, M.; Toner, E.S.; Cicero, A.; Inglesby, T.V. Characteristics of Microbes Most Likely to Cause Pandemics and Global Catastrophes. In *Current Topics in Microbiology and Immunology*; Springer International Publishing: Cham, 2019; pp. 1–20 ISBN 9783030363109.
32. Kreuder Johnson, C.; Hitchens, P.L.; Smiley Evans, T.; Goldstein, T.; Thomas, K.; Clements, A.; Joly, D.O.; Wolfe, N.D.; Daszak, P.; Karesh, W.B.; et al. Spillover and Pandemic Properties of Zoonotic Viruses with High Host Plasticity. *Sci. Rep.* **2015**, *5*, 1–8, doi:10.1038/srep14830.
33. Jones, K.E.; Patel, N.G.; Levy, M.A.; Storeygard, A.; Balk, D.; Gittleman, J.L.; Daszak, P. Global Trends in Emerging Infectious Diseases. *Nature* **2008**, *451*, 990–993, doi:10.1038/nature06536.
34. Carlson, C.J.; Albery, G.F.; Merow, C.; Trisos, C.H.; Zipfel, C.M.; Eskew, E.A.; Olival, K.J.; Ross, N.; Bansal, S. Climate Change Increases Cross-Species Viral Transmission Risk. *Nature* **2022**, *607*, 555–562, doi:10.1038/s41586-022-04788-w.
35. Nuti, S.V.; Wayda, B.; Ranasinghe, I.; Wang, S.; Dreyer, R.P.; Chen, S.I.; Murugiah, K. The Use of Google Trends in Health Care Research: A Systematic Review. *PLoS One* **2014**, *9*, e109583, doi:10.1371/journal.pone.0109583.
36. Cook, S.; Conrad, C.; Fowlkes, A.L.; Mohebbi, M.H. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLoS One 2011, 6, e23610, doi:10.1371/journal.pone.0023610.
37. Bragazzi, N.L.; Alicino, C.; Trucchi, C.; Paganino, C.; Barberis, I.; Martini, M.; Sticchi, L.; Trinka, E.; Brigo, F.; Ansaldi, F.; et al. Global Reaction to the Recent Outbreaks of Zika Virus: Insights from a Big Data Analysis. PLoS One 2017, 12, e0185263, doi:10.1371/journal.pone.0185263.
38. Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature* **2009**, *457*, 1012–1014, doi:10.1038/nature07634.
39. Kapitány-Fövény, M.; Ferenci, T.; Sulyok, Z.; Kegele, J.; Richter, H.; Vályi-Nagy, I.; Sulyok, M. Can Google Trends Data Improve Forecasting of Lyme Disease Incidence? *Zoonoses Public Health* **2019**, *66*, 101–107, doi:10.1111/zph.12539.
40. Verma, M.; Kishore, K.; Kumar, M.; Sondh, A.R.; Aggarwal, G.; Kathirvel, S. Google Search Trends Predicting Disease Outbreaks: An Analysis from India. *Healthc. Inform. Res.* **2018**, *24*, 300, doi:10.4258/hir.2018.24.4.300.
41. Young, S.D.; Torrone, E.A.; Urata, J.; Aral, S.O. Using Search Engine Data as a Tool to Predict Syphilis. *Epidemiology* **2018**, *29*, 574–578, doi:10.1097/ede.0000000000000836.
42. Young, S.D.; Zhang, Q. Using Search Engine Big Data for Predicting New HIV Diagnoses. *PLoS One* **2018**, *13*, e0199527, doi:10.1371/journal.pone.0199527.
43. Morsy, S.; Dang, T.N.; Kamel, M.G.; Zayan, A.H.; Makram, O.M.; Elhady, M.; Hirayama, K.; Huy, N.T. Prediction of Zika-Confirmed Cases in Brazil and Colombia Using Google Trends. *Epidemiol. Infect.* **2018**, *146*, 1625–1627, doi:10.1017/s0950268818002078.
44. Ortiz-Martínez, Y.; Garcia-Robledo, J.E.; Vásquez-Castañeda, D.L.; Bonilla-Aldana, D.K.; Rodriguez-Morales, A.J. Can Google® Trends Predict COVID-19 Incidence and Help Preparedness? The Situation in Colombia. *Travel Med. Infect. Dis.* **2020**, *37*, 101703, doi:10.1016/j.tmaid.2020.101703.
45. Google Trends Available online: https://trends.google.com/trends/ (accessed on 18 August 2023).
46. Mavragani, A.; Ochoa, G. Google Trends in Infodemiology and Infoveillance: Methodology Framework. *JMIR Public Health Surveill.* **2019**, *5*, e13439, doi:10.2196/13439.
47. Arora, V.S.; McKee, M.; Stuckler, D. Google Trends: Opportunities and Limitations in Health and Health Policy Research. *Health Policy* **2019**, *123*, 338–341, doi:10.1016/j.healthpol.2019.01.001.
48. Mulero, R.; García-Hiernaux, A. Forecasting Spanish Unemployment with Google Trends and Dimension Reduction Techniques. SERIEs (Berl) 2021, 12, 329–349, doi:10.1007/s13209-021-00231-x.
49. IEEE DataPort Available online: https://ieee-dataport.org/ (accessed on 18 August 2023).
50. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 1–9, doi:10.1038/sdata.2016.18.
51. Wishart, D.S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B.L.; et al. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res.* **2022**, *50*, D622–D631, doi:10.1093/nar/gkab1062.
52. Slenter, D.N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; Mélius, J.; Cirillo, E.; Coort, S.L.; Digles, D.; et al. WikiPathways: A Multifaceted Pathway Database Bridging Metabolomics to Other Omics Research. *Nucleic Acids Res.* **2018**, *46*, D661–D667, doi:10.1093/nar/gkx1064.

53. Banda, J.M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; Artemova, E.; Tutubalina, E.; Chowell, G. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—an International Collaboration. *Epidemiologia. (Basel)* **2021**, *2*, 315–324, doi:10.3390/epidemiologia2030024.
54. Thakur, N. A Large-Scale Dataset of Twitter Chatter about Online Learning during the Current COVID-19 Omicron Wave. *Data (Basel)* **2022**, *7*, 109, doi:10.3390/data7080109.
55. Thakur, N. MonkeyPox2022Tweets: A Large-Scale Twitter Dataset on the 2022 Monkeypox Outbreak, Findings from Analysis of Tweets, and Open Research Questions. *Infect. Dis. Rep.* **2022**, *14*, 855–883, doi:10.3390/idr14060087.
56. Gjerding, M.N.; Taghizadeh, A.; Rasmussen, A.; Ali, S.; Bertoldo, F.; Deilmann, T.; Knøsgaard, N.R.; Kruse, M.; Larsen, A.H.; Manti, S.; et al. Recent Progress of the Computational 2D Materials Database (C2DB). *2d Mater.* **2021**, *8*, 044002, doi:10.1088/2053-1583/ac1059.
57. Kearnes, S.M.; Maser, M.R.; Wleklinski, M.; Kast, A.; Doyle, A.G.; Dreher, S.D.; Hawkins, J.M.; Jensen, K.F.; Coley, C.W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143*, 18820–18826, doi:10.1021/jacs.1c09820.
58. Goodsell, D.S.; Zardecki, C.; Di Costanzo, L.; Duarte, J.M.; Hudson, B.P.; Persikova, I.; Segura, J.; Shao, C.; Voigt, M.; Westbrook, J.D.; et al. RCSB Protein Data Bank: Enabling Biomedical Research and Drug Discovery. *Protein Sci.* **2020**, *29*, 52–65, doi:10.1002/pro.3730.
59. Urban, M.; Cuzick, A.; Seager, J.; Wood, V.; Rutherford, K.; Venkatesh, S.Y.; De Silva, N.; Martinez, M.C.; Pedro, H.; Yates, A.D.; et al. PHI-Base: The Pathogen–Host Interactions Database. *Nucleic Acids Res.* **2019**, *48*, D613–D620, doi:10.1093/nar/gkz904.