Assessing Public Attention Towards 2022-2023 Mpox Outbreak Using Wikipedia

Steven Kerr

15 February 2024

Table of contents

Abstract (7) **1**

Introduction (6) **2**

Literature Review (5) **3**

Theory / Research Questions and Hypotheses (1) **5**

Data and Methods (2) **6**
Data sources 6
Mpox case data 6
Wikipedia data 8
Methods 9

Analysis (3) **11**
Descriptive analysis 11
Regression models 11

Conclusion (4) **11**
Discussion 11
Limitations 11
Policy Recommendations 11

References (8) **11**

Abstract (7)

[To be added last]

Introduction (6)

As the world becomes increasingly interconnected and climate change elevates the risk of zoonotic spillover events, the public becomes ever more susceptible to global-scale outbreaks.[1] In this context, traditional disease modeling techniques, while effective for well-documented diseases, may not be as applicable for emerging diseases with low case numbers and less understood epidemiological parameters. Critically, public attention plays a pivotal role in disease detection and response.

The advent of digital tools allows for the real-time tracking of this attention, offering a valuable complement to established disease surveillance methods. This is where public attention can serve as an innovative proxy for tracking disease spread. In this regard, the 2022-2023 multi-country mpox outbreak presents a recent case study, allowing us to examine the efficacy of real-time data in monitoring a relatively novel disease. By thoroughly exploring these non-traditional methods, we can more fully understand their advantages and limitations, thus providing crucial insights for decision-makers. This research not only enhances our understanding of digital epidemiology but also contributes to shaping more effective and timely responses to future infectious disease outbreaks.

This project[2] explores the potential for Wikipedia data to be utilized as an alternative metric for gauging public attention towards public health emergencies, specifically by focusing on the 2022-2023 multi-country mpox outbreak. The primary goal is to assess whether Wikipedia, with its vast user base and granular data, can effectively serve as a proxy for measuring public attention and information-seeking behavior in the context of global health crises. This work is motivated by the limitations of existing disease modeling techniques for emerging diseases, which may lack sufficient historical data for accurate forecasting. Public attention, as measured by digital resources like Wikipedia, could provide early indicators of disease outbreaks, thus aiding in more timely and effective public health interventions.

Data on the number of confirmed mpox cases is obtained from the World Health Organization (WHO), while page view statistics for Wikipedia articles related to mpox are sourced directly from Wikipedia. The analysis is restricted to the 10 countries most affected by the outbreak, using time series analysis and lag-correlation methods to examine the relationship between Wikipedia page views and mpox case numbers. The study aims to identify patterns and correlations that could support the use of Wikipedia data as a predictive tool for public health surveillance.

This work contributes to the broader academic discourse by exploring the potential of open-source data for enhancing disease surveillance and response strategies. It seeks to validate the efficacy of Wikipedia, as a non-traditional data source, in providing actionable insights to policymakers during public health emergencies. In doing so, this work addresses gaps in the literature regarding the effectiveness of digital tools in enhancing disease surveillance.

[1] https://www.nature.com/articles/s41579-021-00639-z

[2] Project GitHub: https://github.com/smkerr/Thesis_WorkInProgress

Literature Review (5)

Several key academic papers lay the groundwork for this approach. In "Assessing Public Interest Based on Wikipedia's Most Visited Medical Articles During the SARS-CoV-2 Outbreak: Search Trends Analysis" by Chrzanowski et al. (2021), the authors investigates how public interest in medical topics changed during the COVID-19 pandemic, as reflected by Wikipedia pageviews. The study conducted a retrospective analysis of access to medical articles across nine language versions of Wikipedia and correlated these patterns with global and regional COVID-19 deaths, comparing observed data to a forecast model trained on data from 2015-2019. This involved collecting daily page view statistics for 37,880 articles curated by the English Wikipedia Medicine Project from 1 July 2015 to 13 September 2020. The authors sourced page view statistics using ToolForge, a page view analytics tool) and sourced COVID-19 death statistics from Our World in Data. It found a correlation between the pandemic's severity and pageviews for COVID-19–related Wikipedia articles, concluding that changes in article popularity could serve as a method for epidemiological surveillance by reflecting public attention during disease outbreaks. Furthermore, it demonstrates the potential for using Wikipedia data for epidemiological surveillance and understanding public information-seeking behavior during disease outbreaks. While the paper focuses on the COVID-19 pandemic, similar methodologies can be applied to provide insights into public attention and information-seeking behavior during the Mpox outbreak.

"Association between public attention and monkeypox epidemic: A global lag-correlation analysis" by Yan et al. (2023) investigates the association between public attention, as measured by Google Trends Index (GTI), and the global mpox oubtreak. The authors use Google Trends data for information on internet search activity related to mpox as well as data on daily confirmed mpox cases from Our World in Data. It tests time-lag correlations between GTI and daily confirmed mpox cases across the 20 countries with the highest case numbers as of 20 September 2022 using the Spearman correlation coefficients, over a range of -36 to +36 days. Spearman correlation coefficients from these 20 countries were pooled to provide a combined correlation coefficient for each lag. To test whether the time series was stationary, an Augment Dickey-Fuller (ADF) test is applied. The study finds a strong positive correlation, particularly noticeable 13 days after a peak in public attention. The study also conducted meta-analyses and utilized vector autoregression (VAR) models to analyze the temporal relationship between GTI and daily confirmed mpox cases, and a Granger-causality test was employed to evaluate whether the GTI trend could predict daily confirmed mpox cases. The findings suggest that GTI could be a useful tool for early monitoring and prediction of mpox cases, highlighting the significance of digital epidemiology in infectious disease surveillance. The study emphasizes the potential of internet data like GTI in providing early warning signs for health outbreaks and aiding in rapid response strategies. While the study utilizes GTI data to public attention towards the mpox outbreak, however similar methods can be applied to Wikipedia pageviews data to test whether the same conclusions can be drawn using Wikipedia as a data source.

In "Trends in Online Search Activity and the Correlation with Daily New Cases of Monkeypox among 102 Countries or Territories" by Du et al. (2023), the authors investigate the relationship between online search activity related to mpox and the actual daily new cases of

mpox across 102 countries or territories. The study aims to understand how internet search trends can reflect public awareness towards mpox, potentially serving as an early indicator for outbreaks. Data on daily mpox cases from 1 May 2022 to 9 October 2022 was sourced from Our World in Data, while online search activity data related to mpox was sourced from Google Trends using the keyword "monkeypox".[3] Online search activity was expressed as relative normalized search volume numbers (RNSNs) ranging from 0 to 100 to reflect how many searches are performed for a keyword relative to the total number of searches on the internet over time where a value of 100 represents the time point at which the search term has reached its peak in popularity. Demographic data including total population, population density, average years of schooling, socioeconomic status, and public tourism were sourced from the United Nations and World Bank. Data on health status including HIV prevalence, sanitation levels, and health workforce densities were obtained from the 2019 Global Burden of Disease study. The authors use a segmented time-series analysis to estimate the impact of the PHEIC declaration on online search activity, adjusting for daily new cases across 194 countries or territories. Furthermore, the study tests time-lag correlations between online search activity and daily new cases, specifically considering lags of -21, -14, -7, 0, +7, +14, and +21 days. Next, a general linear regression model (GLM) is used to explore influencing factors on the relationship between online search activity and daily new cases. The authors find a significant correlation between online search activity and daily mpox cases, with online searches often preceding reporting of new cases. This study highlights the value of integrating internet search data into public health surveillance for emerging infectious diseases. Similar to the paper by Yan et al., this study utilizes Google Trends data, however similar methodologies could be applied towards Wikipedia pageviews data.

Beyond these studies, several other academic papers have contributed analysis to various aspects of the relationship between online information-seeking behavior and public health.

- García-Gavilanes et al. (2016) use Wikipedia page view and page edit statistics to investigate public attention towards airline crashes.
- In their analysis of page views for COVID-19-related Wikipedia pages, Gozzi et al. (2020) find that page views were mainly driven by media coverage, declined rapidly, even while COVID-19 incidence remained high, raising questions about the impacts of attention saturation in disease outbreak settings.
- Bhagavathula and Raubenheimer (2023) conducted a joinpoint regression analysis to measure hourly percentage changes (HPC) in search volume in the hours immediately preceding and following WHO's determination to assign PHEIC status to mpox, finding an overall increase in information-seeking behavior, although results varied by country. This study revealed a 103% increase in public interest in top five Mpox-affected countries immediately following the WHO PHEIC announcement. However, search interest waned after the announcement, so that search interest appeared to reflect media attention more than disease spread.

[3]Note that this study was conducted prior to WHO's recommendation that monkeypox instead be referred to as "mpox" on 28 November 2022: https://www.who.int/news/item/28-11-2022-who-recommends-new-name-for-monkeypox-disease

- Gong et al. (2022) use the Baidu Index (BDI) and Sina Macro Index (SMI) to investigate the association between public attention towards the COVID-19 pandemic and new cases using Spearman correlation.
- Abbas et al. (2021) analyze associations between Google Search Trends for symptoms of COVID-19 and confirmed cases and deaths within the United States, demonstrating abilitiy to predict cases up to three weeks prior.
- Hickmann et al. (2015) demonstrate that it is possible to use Wikipedia page view statistics and CDC influenza-like illness (ILI) reports to create a weekly forecast for seasonal influenza, finding that that Wikipedia article access are highly correlated with historical ILI records, allowing for highly accurate disease forecasts several weeks before case data is available.

Given the existing literature, this thesis aims to fill the research gap by investigating whether Wikipedia can be predictive of mpox cases during the 2022-2023 multi-country outbreak. While previous studies have evaluated Wikipedia data for COVID-19 and others have utilized Google Trends data for mpox, this thesis represents the first attempt, as far as I am aware, to explore the relationship between Wikipedia page view statistics and mpox cases.

Theory / Research Questions and Hypotheses (1)

To what extent can Wikipedia data be effectively utilized as an alternative method for measuring public attention and information-seeking behavior during the 2022-2023 multi-country mpox outbreak?

This research question engages the following current and relevant conversations within the literature:

- *Open Source Data for Public Health Surveillance*: Examining the utility and limitations of using public data sources like Wikipedia to monitor and assess public health events, contributing to ongoing discussions about their reliability and relevance.
- *Information Dissemination and Public Awareness*: Investigating the extent to which public awareness of outbreaks is shaped by the impact (number of cases and/or deaths), connecting with debates about information ecosystems and their impact on public health communication.
- *Policy Implications*: Discussing the potential policy recommendations and interventions that can arise from a better understanding of public attention and information-seeking behavior during outbreaks on digital platforms.

My thesis also contains several sub-questions to be investigated in support of the main research question:

- Which medical articles saw traffic volume increase significantly after the start of the mpox outbreak?

- To what extent do the number of mpox cases correlate with the traffic volume of mpox-related Wikipedia articles?
- How effective is Wikipedia analytics data compared to other data sources (e.g., Google Trends) when it comes to gauging public attention towards the mpox outbreak?

Data and Methods (2)

Data sources

To answer my research question, I will rely on two main data sources. First, country-level data on the weekly number of mpox cases is sourced from WHO.[4] Second, Wikipedia analytics data on daily page view volume by article is sourced directly from Wikipedia.[5]

My topic focuses on assessing public attention towards the 2022-2023 multi-country mpox outbreak using Wikipedia data. As such, this analysis relies on two main data sources. First, country-level data on the weekly number of mpox cases is sourced from the World Health Organization (WHO). Second, Wikipedia analytics data on page view volume is sourced directly from Wikipedia.

Mpox case data

Daily aggregated numbers of mpox cases by country correspond with the date on which cases were reported to public health authorities. One advantage of this dataset is that it is considered to be largely complete since it comprises every confirmed and probable case reported to the national public health authorities.[6] While this still leaves room for cases to go underreported in instances where an individual does not seek medical attention (e.g., for fear of stygmatization) or for asymptomatic cases, it still represents the most comprehensive view of the outbreak's scale. Aggregated data is available for all reported cases as of 31 December 2023.

The top 10 countries by number of confirmed cases are the United States of America (31,246), Brazil (10,967), Spain (7,752), France (4,171), Colombia (4,090), Mexico (4,078), the United Kingdom (3,875), Peru (3,812), Germany (3,800), and China[7] (2,025).

As of 31 December 2023, mpox cases have been reported by 117 WHO Member States across all six WHO regions. The dataset contains cases reported between 7 January 2022 to 31 December 2023.

For this analysis, I will only consider confirmed cases. However, I do not expect this decision to greatly impact the results considering that probable cases only make up 0.7% of overall

[4] https://worldhealthorg.shinyapps.io/mpx_global/

[5] https://wikitech.wikimedia.org/wiki/Data_Engineering/Systems/AQS

[6] For more information on what constitutes a confirmed or probable case, please refer to WHO's mpox case definitions: https://www.who.int/emergencies/outbreak-toolkit/disease-outbreak-toolboxes/mpox-outbreak-toolbox

[7] Cases shown include those reported in mainland China (1,611), Taiwan (333), Hong Kong (80), and Macao (1).

cases (652/93,682) and only eight countries report any probable cases. While Puerto Rico reports the highest proportion of probable cases at 42% of total cases (150/361), probable cases make up less than 5% of the remaining countries' total cases.

While WHO collects aggregated data on a daily basis, nearly all countries' public health authorities report cases at a weekly frequency. As a result, it is more useful to aggregate cases by epidemic week. The epidemic curve below depicts the aggregated weekly number of cases by week reported.

While the global trend appears to have been quite coherent with the number of weekly cases peaking in July 2022, this disguises the fact that the trends in cases looked quite distinct at the WHO region-level.

Here we observe that there is substantial variation between the six WHO regions, with the Eastern Mediterranean Region, European Region, and Region of the Americas following a similar trend with cases peaking in summer/fall 2022, while cases in the South-East Asia Region and Western Pacific Region peak in summer/fall 2023. In contrast, cases reported by the African Region appear to be more uniformly distributed, reflecting the fact that mpox is endemic to certain areas of western and central Africa.

In addition to aggregated case data, WHO also collects line list data where each row corresponds with an individual case and contains information on demographics, clinical presentation, epidemiological exposure factors, and laboratory testing.[8] Due to privacy concerns, line list data is stripped of all personally identifiable information and aggregated by country and date before being made available by WHO.[9] In contrast to aggregated case data, the `date` variable of the detailed case dataset corresponds with either the date of symptom onset, the date of diagnosis (if date of symptom onset is not available), or the date of reporting (if date of symptom onset and date of diagnosis are not available).[10] This difference in how cases are assigned to a date grants us a much more granular view of countries' epidemic curves, as shown below. Dates are aggregated at the weekly level to further protect individual cases' privacy.

A significant disadvantage of this detailed dataset is that it only contains information 63% for all reported mpox cases (58,883/93,030). This is driven almost entirely by the fact that WHO no longer includes cases from the United States in this dataset, despite the fact that the United States represents 34% of global cases (31,246/93,030). That said, this data can still be useful for analysis of other countries. As such, detailed dataset can complement the aggregated data presented above.

While WHO collects line list data on a daily basis, this data is aggregated by epidemic week to safeguard the privacy of individual cases. The epidemic curve below depicts the weekly number of cases. Compared to the epidemic curves produced using aggregated data, the detailed data allows us to plot much smoother curves which seem to adhere more closely to the trend we might expect of an infectious disease outbreak.

[8] https://www.who.int/publications/m/item/monkeypox-minimum-dataset-case-reporting-form-(crf)
[9] https://worldhealthorg.shinyapps.io/mpx_global/
[10] https://worldhealthorg.shinyapps.io/mpx_global/

Again, while the global trend appears to have been quite coherent with the number of weekly cases peaking in July 2022, the trends vary at the WHO region-level.

Wikipedia data

The Wikimedia Foundation makes it straightforward to access various analytics data related to its projects, including Wikipedia, by providing the Wikimedia Analytics Query Service (AQS) REST API. AQS offers a range of analytics data, such as page view statistics, editor activity levels, and other traffic data from as far back as 1 August 2015. The REST API facilitates the retrieval of analytics data from Wikipedia in a structured way. Given that Wikipedia data is abundant, publicly accessible, and commonly used, many resources exist to easily access this data.[11] {waxer} is one such package which serves as a Wikimedia API wrapper that facilitates querying for traffic (pageviews, unique devices), user (e.g. active editors), and content-based metrics (e.g. edits counts, pages counts) from Wikimedia Analytics Query Service with R.

Since this project is concerned with assessing public attention, Wikipedia page view statistics will serve as our primary measure for online information-seeking behavior. We query page view statistics from the AQS REST API using the following specifications:

We start exploring the dataset by examining the absolute and relative frequency of different values within the `project` and `page_name` variables. As expected, all page views are for the English Wikipedia project. Notably, the "Mpox" article has a substantially higher traffic volume than the "Monkeypox virus" article (71% vs. 29%).

Next, we explore the missingness of each of variable in the `pageviews` dataset. We find that `redirect_name` is the only column missing values. For the specified Wikipedia project and articles, 27% of page views were redirected from other search terms.

Seeing as a substantial number of page views are driven by redirects, it is important to understand whether it is valid to include these redirects in our analysis. If search terms appear related to mpox or the monkeypox virus, then it is safe to assume that this represents online information-seeking behavior towards our topic. Our table shows that all search appear to directly relate to mpox or the monkeypox virus, so we conclude that is valid to include redirects in this analysis. This will be reevaluated and handled as more projects and articles are added to the analysis.

Considering that mpox case data is available at a weekly level of granularity, we aggregate page view data by week and plot the results below. We observer two large spikes, with the first centered on May 2022 when non-endemic countries began reporting mpox cases[12] and the second centered on late July 2022 when WHO declared mpox to be a Public Health Emergency of International Concern (PHEIC).[13] Other smaller peaks can be observed, although

[11] Mikhail Popov, Data Science Manager in Product Analytics at the Wikimedia Foundation, has published a list of R packages related to or affiliated with the Wikimedia Foundation here: https://people.wikimedia.org/~bearloga/notes/r-pkgs.html

[12] https://www.who.int/emergencies/situations/monkeypox-oubreak-2022

[13] https://www.who.int/europe/news/item/23-07-2022-who-director-general-declares-the-ongoing-monkeypox-outbreak-a-public-health-event-of-international-concern

the general trend indicates that public attention decreases over time.

Another element of this project involves developing a baseline for what online information-seeking behavior may have looked like in the absence of the mpox outbreak, as measured by Wikipedia article access. For this, we will use overall project views for the respective Wikipedia projects to detect seasonality and long-term trends in Wikipedia search trends. To illustrate what this looks, we examine the Wikipedia projects corresponding with the most common languages in the top 10 countries by number of mpox cases.

The plot below shows that English Wikipedia predominates overall Wikipedia traffic, with weekly traffic levels for the other language projects clustered together well below.

While this Data Report showcases the underlying data structure of Wikipedia page view statistics using as "Mpox" and "Monkeypox virus" articles as mere examples, there are many other articles related to mpox and its symptoms[14] that may also contribute to this analysis. Supplemental article titles are listed below. Since article titles have been recorded in English, the next step would be to cross-reference them with their corresponding titles for other Wikipedia projects.

Methods

While mpox case data is available for 117 WHO Member States and Wikipedia page view statistics exist for nearly 300 languages, I will limit the scope of this analysis to the 10 countries with the most cumulative cases, including the United States of America (31,246), Brazil (10,967), Spain (7,752), France (4,171), Colombia (4,090), Mexico (4,078), the United Kingdom (3,875), Peru (3,812), Germany (3,800), and China[15] (2,025). Accordingly, I will limit the analysis to the Wikipedia projects for the languages that prominently feature in these countries, including English (the United States of America and the United Kingdom), Portuguese (Brazil), Spanish (Spain, Mexico, and Peru), French (France), German (Germany), Chinese (China). I will examine the time period from 1 January 2022 to 31 December 2023.

Country	**Number of cases**	**Wikipedia project**
United States of America	31,246	English
Brazil	10,967	Portuguese
Spain	7,752	Spanish
France	4,171	French
Colombia	4,090	Spanish
Mexico	4,078	Spanish
United Kingdom	3,875	English
Peru	3,812	Spanish
Germany	3,800	German
China	2,025	Chinese

[14] https://www.who.int/news-room/fact-sheets/detail/monkeypox

[15] Cases shown include those reported in mainland China (1,611), Taiwan (333), Hong Kong (80), and Macao (1).

My proposed methodology takes inspiration from th work of Yan et al. and Du et al. I will conduct an observational study to assess the lag-correlation between public attention and mpox cases for the 10 countries with the most cumulative cases.

1. Define collection of mpox-related Wikipedia articles
    - Identify articles directly related to mpox, including historical information, symptoms, treatment, and prevention.
    - Identify articles with a low degree of separation within network of Wikipedia articles.
    - Analyze Wikipedia page view statistics to identify medical articles that experienced significant increases in traffic coinciding with the timeline of the 2022-2023 mpox outbreak.[16]
2. Data preparation
    - Collect daily mpox case numbers from WHO.
    - Extract daily traffic volume data for the defined collection of Wikipedia articles using Wikipedia's API with the {waxer} package using R.
    - De-noise data by aggregating both mpox cases and Wikipedia page view statistics to the weekly level.
    - Standardize Wikipedia traffic volumes to be expressed as a percentage of total traffic volume for a given Wikipedia language version.
3. Statistical analysis
    - Perform Spearman correlation tests to examine the time-lag relationship between Wikipedia traffic volumes and mpox case numbers. The range of -21 to +21 days will allow analysis of lead and lag effects.
    - Use non-parametric methods, considering the non-normal distribution of the data.
4. Augmented Dickey-Fuller (ADF) test
    - Implement the ADF test to check for stationarity in both the Wikipedia traffic and mpox case series. Non-stationary data can lead to spurious results in subsequent analyses.
5. Vector autoregression (VAR) model
    - Develop a VAR model to understand the dynamic relationship between the two time series. This model will help in capturing the temporal interdependencies and feedback mechanisms between Wikipedia traffic and mpox cases.
    - Determine the optimal lag length for the VAR model based on information criteria like AIC or BIC.
6. Granger causality test
    - Apply the Granger causality test within the VAR framework to assess whether Wikipedia traffic volumes can be considered a predictor of mpox case trajectories.
    - This test will help determine if changes in Wikipedia page views precede changes in mpox cases, indicating a predictive relationship.
7. Validation and robustness checks
    - Conduct sensitivity analyses to test the robustness of the findings against different model specifications and subsets of data.

[16] An assessment would need to be made to determine that the increase in traffic volume of certain medical articles following the start of the outbreak is not spurious but substantial.

    - Validate the results through comparison with other studies or datasets.
8. Interpretation and implications
    - Interpret the results, while considering the limitations of observational data and the potential for confounding factors.
    - Discuss the implications for public health surveillance during health emergencies.

Analysis (3)

Descriptive analysis

Regression models

Conclusion (4)

Discussion

Limitations

Policy Recommendations

References (8)

[s/o Zotero]