Pre-Analysis Plan

Assessing Public Attention Towards 2022-2023 Mpox Outbreak Using Wikipedia

Steven Kerr

31 January 2024

Summary

This project[1] 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] Project GitHub: https://github.com/smkerr/Thesis_WorkInProgress

Motivation & Background

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.[2] 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. For these reasons, I find the research question to be of particular interest.

This research topic closely connects to my academic and professional background. Throughout my studies, I have taken opportunities to apply my data science skills towards projects related to epidemics and infodemics (primarily COVID-19). During my Professional Year, I worked with CPC Analytics, a Berlin-based consultancy focused on global health topics, where I directly contributed to projects for the World Health Organization's (WHO) Pandemic Hub and the WHO Health Emergencies Programme's Mpox Data Analytics Unit. This experience not only provided me with a comprehensive understanding of the datasets related to mpox but also led to my contribution to two separate publications: "Description of the first global outbreak of mpox: an analysis of global surveillance data" and "Mpox in Children and Adolescents during Multicountry Outbreak, 2022-2023". Currently, I am contributing towards an evidence synthesis of global health literature at the Mercator Research Institute on Global Commons and Climate Change (MCC), further deepening my expertise in epidemiology and fueling my interest in global health issues. These cumulative experiences have not only expanded my knowledge in this field but have also solidified my aim to continue contributing to the filed of global health.

Through this project, I hope to enhance a range of critical skills. First, this project involves time series analysis as it tracks changes in public interest and confirmed cases over time, thus deepening my knowledge of statistical techniques involved in working with time series data. In terms of domain knowledge, this project requires an in-depth understanding of public health dynamics, particularly in the context of disease outbreaks, allowing me to further build on my existing experience. Lastly, this project will enhance my project management skills, as

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

it demands careful planning, organization, and time management to successfully conduct an independent research project within a set timeframe, thus preparing me for managing future data science projects effectively.

Introduction

Mpox (formerly known as monkeypox) is an infectious disease caused by the monkeypox virus (MPXV).[3] It was first identified in laboratory monkeys in 1958 and first recorded in humans in 1970.[4] While mpox transmission has been well-documented in parts of western and central Africa for several decades,[5] beginning in May 2022, a high proportion of cases were reported in countries which had not previously observed sustained chains of transmission.[6] The unexpected appearance of mpox in several regions in the initial absence of epidemiological links to areas in western and central Africa, suggests that there may have been undetected transmission for some time, especially concerning given that the World Health Organization (WHO) considers the confirmation of one case of mpox, in a country, to be an outbreak.[7] [8] In light of the rapidly increasing case numbers, WHO declared the multi-country mpox outbreak to be a Public Health Emergency of International Concern (PHEIC), the first such declaration since the COVID-19 Pandemic.[9] As of 31 December 2023, 117 WHO Member States across all six WHO regions have reported mpox cases, amounting to 93,030 confirmed cases, 652 probable cases, and 176 deaths.[10]

While sophisticated models exist for common diseases such the flu and COVID-19, outbreaks of relatively novel diseases or large-scale outbreaks of diseases which have previously had very limited transmission may lack sufficient prior data on which to build initial models. This highlights the need for real-time forecasting methods, even if only in the initial periods until sufficient data are available for more robust models. Due to its relatively low number of cases and lack of attention from the global research community prior to the 2022-2023 multi-country outbreak, mpox represents a potential case study to test whether these sorts of methods might apply to these types of outbreak. To that end, public attention can play a pivotal role in the detection and response to infectious diseases, having been used by some scholars to predict infectious disease outbreaks.[11] With modern advancements in digital tools, it is now possible to track this attention in real-time. Understanding real-time health information-seeking behavior can act as a proxy indicator of public health needs.

[3] https://www.who.int/health-topics/monkeypox/

[4] https://www.who.int/europe/health-topics/monkeypox

[5] https://www.who.int/health-topics/monkeypox/

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

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

[8] https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON393

[9] 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

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

[11] https://doi.org/10.1371/journal.pcbi.1004239

Internet searches have become a critical source of health information, with Wikipedia representing a widely used online resource for health information.[12] In fact, with regard to medical information, studies have shown that Wikipedia's popularity exceeds that of the National Health Service, WebMD, Mayo Clinic, and WHO websites combined.[13] As such, Wikipedia can be harnessed as a potential tool for infectious disease outbreak surveillance. Wikipedia page views have already been used to study and predict the spread of other infectious diseases.[14] The analysis of Wikipedia medical article popularity could be a viable method for epidemiological surveillance, as it provides important information about the reasons behind public attention and factors that sustain public interest in the long term.

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

[12] https://doi.org/10.1371/journal.pone.0228786

[13] https://europepmc.org/article/MED/19390105

[14] https://dx.plos.org/10.1371/journal.pcbi.1004239

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".[15] 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.

[15] 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

- 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.
- 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.

Research Question

My research question is as follows:

> 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 & Methods

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.[16] Second, Wikipedia analytics data on daily page view volume by article is sourced directly from Wikipedia.[17] For more details on these data sources, please refer to the Data Report.

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

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

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

(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[18] (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

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.[19]
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.

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

[19] 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.

- 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.
   - 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.