AI-based surveillance of Depression in South Africa using Google Trends data

Annotated Bibliography

Tristan Dos Remendos (2465830)
University of Witswatersrand

March 26, 2024

References

[1] C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh, "Using web and social media for influenza surveillance," in *Advances in Experimental Medicine and Biology*, vol. 680, 2009, pp. 559 – 564.

Aim

To explore using web and social media data as a way to detect and track increases in influenza-like illness (ILI) in the population.

Style

Conference Paper. Experimental and statistical.

Cross-references

While focused on influenza rather than mental health conditions like depression explored in [2], [5], and [6], this study shares the underlying approach of using internet data for disease surveillance. The methodology of analyzing frequencies of relevant blog posts aligns with the principles of the Google Trends analyses in those other works, and is explored more extensively in sources [2], [3], [4], and [5]. The high correlation found between blog posts and CDC influenza data provides early evidence supporting the utility of this digital surveillance approach, consistent with the strong associations reported in some of the other studies. However, the limitations discussed around isolating true cases and the narrow 20-week timeframe suggest this work has a more limited scope compared to some of the larger-scale national studies.

Summary

Traditional influenza surveillance relies on extrapolating from diagnosed cases at healthcare providers, which likely misses many undiagnosed cases. Ideally, there may be an approach to surveying various diseases using web-based data which can produce valid or preliminary results.

The approach taken was to analyze the frequency of blog posts mentioning influenza keywords like “flu” or “influenza” over a 20-week period from October 2008 to January 2009. The researchers extracted relevant English blog posts from a large web crawling database. They calculated the number of “flu-content” posts per week and compared the trends to CDC ILI surveillance data from healthcare providers over the same time period.

The key finding was a high correlation (Pearson r = 0.626) between the weekly frequency of flu-content blog posts and the CDC’s ILI data, suggesting blogs could be a useful complementary data source for influenza surveillance.

A major benefit of this approach is that it can potentially detect ILI cases that go undiagnosed at healthcare providers. Limitations include that not all flu-content posts may actually reflect real ILI cases, and isolating relevant self-reported ILI cases from other types of flu posts is challenging. Additionally, the study period of 20 weeks is relatively short. It is also important to note that the study was conducted over a decade ago, meaning newer media sources may be a more accurate substitution for disease surveillance using this approach.

Overall, the study provides initial evidence that trends in flu-related web and social media content may have utility for augmenting traditional disease surveillance approaches.

[2] A. Fulk, D. Romero-Alvarez, Q. Abu-Saymeh, J. S. Onge, A. Peterson, and F. Agusto, “Using google health trends to investigate covid-19 incidence in africa,” *PLOS One*, June 2022.

Aim

To investigate the utility of Google Health Trends (GHT) in monitoring and predicting COVID-19 incidence in Africa.

Style

Research Article. Scientific and analytical.

Cross-references

Employing a similar regression-based methodology to [4] and [5] for analyzing search data, this study focuses specifically on using Google Trends for COVID-19 surveillance across multiple African countries. In contrast to the successful mapping of depression demonstrated in [5], the very weak correlations found here between search volumes and cases highlight the limitations of this approach for an emerging epidemic. The exploration of explanatory factors like demographics provides insights relevant to other works aiming to account for differences in digital access and disease awareness, which could impact search patterns. The authors’ suggestion that established diseases may be better suited aligns with the influenza findings in [1] compared to a novel outbreak.

Summary

The report explores whether Google Health Trends (GHT) search data can be used to help predict or track COVID-19 incidence across 54 African countries. The researchers collected COVID-19 case and death data for these countries across four time periods in 2020-2021. They also obtained GHT search data for four COVID-related terms like “coronavirus” and “COVID19”.

They performed regression analyses to assess the correlation between the GHT search data and reported COVID-19 cases for each country. Overall, they found very weak correlations, with adjusted R^2 values below 0.4 for all countries. The top three countries in terms of correlation were Algeria, Ethiopia and Kenya.

The researchers evaluated many potential factors like internet access, demographics, economics and health indicators to see if they could explain the patterns in the GHT data. A few variables like average weekly cases, total deaths, broadband subscriptions and volatility of the case data showed some association with GHT performance.

Ultimately, while this approach offers a free way to anticipate outbreaks, identify disease hotspots, and understand disease surveillance patterns, the study demonstrated little applicability of using GHT data for COVID-19 surveillance in most African countries studied. The authors suggest GHT may work better for tracking diseases with more consistent diagnosis and interest levels from the public. For emerging epidemics like COVID-19, GHT appears to have limited utility based on this analysis.

[3] A. K. Johnson, R. Bhaumik, I. Tabidze, and S. D. Mehta, “Nowcasting sexually transmitted infections in chicago: Predictive modeling and evaluation study using google trends,” *JMIR PUBLIC HEALTH AND SURVEILLANCE*, 2020.

Aim

To develop a predictive model using Google Trends data to improve surveillance and forecasting of sexually transmitted infections (STIs) like chlamydia, gonorrhea, and syphilis.

Style

Journal Article. Experimental and statistical.

Cross-references

Similar to [2], [4], and [5], this work explores using Google Trends data for disease surveillance, but focuses on sexually transmitted infections (STIs) in Chicago. The predictive modeling approach aligns with the machine learning techniques employed in [4] and [5], providing a contrasting example for a different disease domain. The ability to "nowcast" emerging trends complements the motivations for early detection discussed in works like [2] and [4]. The subgroup analyses by demographics highlight considerations around generalizability and representation, an issue also raised in the other studies. The limitations regarding API access and media influence are relevant constraints applicable to all Google Trends analyses. However, by concentrating on a single city, the findings may have limited geographic scope compared to national-level studies in [2] and [5].

Summary

Traditional surveillance systems suffer from data quality issues, underreporting, and reporting delays, missing opportunities for timely interventions.

The authors obtained STI case data from 2011-2017 for Chicago from the city's health department. They identified the top 100 Google search terms correlated with "STD symptoms" and collected the search volume trends for those terms using the Google Health API. They applied elastic net regression modeling using the search data as predictors and the STI case counts as the outcome to develop predictive models.

The models showed moderate to high correlations between predicted and actual STI case counts across diseases and years, performing best for gonorrhea. Subgroup analyses by race, sex, and age improved model fit. Cross-correlation analyses helped identify search terms that preceded changes in STI trends, allowing for earlier prediction.

A key benefit is the ability to "nowcast" emerging STI trends in near real-time using search data, allowing for timely public health responses and targeted interventions. Limitations include the study being limited to one city, lack of other demographic/location data for subgroups, and reliance on an API that may have restricted access. The use of Google trends data is can also be volatile due to the aforementioned issues that come with the use of Internet-based data. (Susceptibility to media influence, misrepresentation of actual trends etc.)

Overall, integrating Google Trends with surveillance activities shows promise for enhancing disease monitoring, outbreak detection, and tailoring interventions to impacted subpopulations for STIs.

[4] L. G. Santos, "Surveillance of tuberculosis by analysing google trends," *Católica Lisbon School of Economics and Business*, June 2023.

Aim

To investigate the feasibility of using Google Trends (GT) data to predict the incidence of tuberculosis (TB) in Portugal.

Style

Research Report/Thesis. Scientific, statistical and analytical.

Cross-references

Paralleling the Google Trends analysis approach used in [2], [3] and [5] for different diseases, this study focuses specifically on tuberculosis surveillance in Portugal using search data and machine learning models. The exploration of model performance metrics like mean absolute errors provides methodological insights relevant to predictive accuracy evaluations done in similar studies like [3]. The promising results in forecasting tuberculosis cases align with the successful depression mapping achieved in [5] and STI nowcasting in [3], suggesting this approach may be better suited for more established diseases. This was hypothesized in by the authors in [2]. However, the study's limited geographic scope contrasts once again with the larger-scoped studies in [2] and [5].

Summary

The author aims to develop an approach by analyzing search patterns and employing machine learning models to forecast monthly Tuberculosis incidence, ultimately providing real-time information for early detection and intervention by public health authorities. The benefits of this approach include improved accuracy and timeliness of surveillance, alignment with seasonal patterns, and contributions to medical nowcasting literature. However, limitations such as restricted scope, lack of specificity in Google Trends data, susceptibility to media influence, and assumptions about seasonality pose challenges that need to be addressed in future research to fully leverage the potential of Google Trends for Tuberculosis surveillance.

The author collected historical TB incidence data from the European Centre for Disease Prevention and Control, as well as GT search volume data for 19 TB-related terms in Portugal from 2007-2023. Different machine learning models were trained on this data to forecast monthly TB incidence from 2021-2023. Model performance was evaluated using metrics like mean absolute error, root mean squared error, and the Akaike information criterion.

The results showed that the Partial Least Squares (PLS) model achieved the best predictive performance across all metrics compared to other models like ordinary least squares, LASSO, support vector machines, and deep neural networks. The author argues that this GT-based surveillance system could provide public health authorities with real-time projections to identify potential TB outbreaks earlier and more cost-effectively than traditional methods.

[5] A. Wang, R. McCarron, D. Azzam, A. Stehli, G. Xiong, and J. DeMartini, "Utilizing big data

from google trends to map population depression in the united states: Exploratory infodemiology study," University of California, Tech. Rep., 2022.

Aim

To utilize big data from Google Trends to map and analyze population depression levels across the United States based on internet search queries related to depression.

Style

Research Report. Statistical and analytical.

Cross-references

Building on the fundamental approach explored in [1] for using internet data for disease surveillance, this study utilizes Google Trends analysis and regression modeling techniques similar to [2], [3] and [4], but for mapping depression prevalence across the United States instead of infectious diseases. The successful identification of trends and geographic/seasonal patterns demonstrates the potential value of this approach for mental health monitoring, a domain not covered in the other works focused on influenza, COVID-19, tuberculosis etc. The limitations regarding interpretation and generalizability are applicable considerations for all such digital surveillance efforts. The findings on aligning data with known risk factors reinforce the validity of this method against traditional epidemiological data sources.

Summary

The paper tackles the challenge of mapping population depression in the United States by leveraging big data from Google Trends, aiming to understand depression search intent and generate heat maps of depression prevalence. Using this approach offers cost-effective and easily accessible real-time tracking of mental health trends, enabling identification of seasonal patterns and geographic variations in depression search intent for informed public health planning. It complements traditional surveys and epidemiological studies, providing predictive insights. Though limitations include the issue of interpreting trends without clinical context, susceptibility to external influences on Google Trends data, and the potential exclusion of certain populations, affecting the generalizability of findings.

The key findings were:

1. Depression search interest grew 67% from 2010 to 2021 and is projected to grow another 7.4% by 2025, mirroring the rise in reported depression prevalence.
2. Search interest showed significant seasonal variation, peaking in spring and reaching its lowest in summer. This matches the pattern seen in seasonal affective disorder.
3. States with higher air pollution and those in the Northeast (further from the equator) had higher depression search interest, aligning with known environmental

and geographic risk factors.

4. Multivariable regression models confirmed time, seasonality, air quality, and geographic region as significant predictors of depression search interest levels.

The authors argue that Google Trends data can serve as a novel digital epidemiological tool to map depression prevalence and trends across populations at low cost and in real-time, overcoming limitations of traditional surveys. However, they caution against over-reliance on this data which may have biases and not fully capture severely depressed individuals. This study offers preliminary evidence that analyzing big data from search engines can complement surveys to better understand the burden of mental health conditions.

[6] A. C. Yang, S.-J. Tsai, N. E. Huang, and C.-K. Peng, "Association of internet search trends with suicide death in taipei city, taiwan, 2004–2009," *Journal of Affective Disorders Reports*, 2011.

Aim

To examine how internet search trends for suicide-related terms were associated with and potentially leading indicators of actual suicide rates in the population studied, specifically in Taipei City, Taiwan.

Style

Journal Article. Scientific and statistical.

Cross-references

While the studies in [1], [2], [4], and [5] used regression and machine learning approaches for analyzing search data, this work employs cross-correlation and stepwise regression methods to associate suicide-related search trends with actual suicide rates. The temporal lead-lag findings complement the seasonal patterns identified in [5] for depression searches, highlighting how search data could provide early indicators of concerning mental health conditions beyond just measuring prevalence. The authors' proposal to filter harmful suicide content also provides a pragmatic insight relevant to implementation of systems like those described in the other studies. However, the narrow geographic focus once again limits generalizability compared to national-level scoped works.

Summary

The paper focuses on investigating the relationship between Internet search trends for suicide-related terms and suicide death rates, aiming to identify significant factors associated with suicide through statistical analysis. This approach offers a source of information for understanding suicide risks and trends, potentially aiding in the development of targeted interventions to reduce suicide rates. It emphasizes the importance of search engine providers filtering harmful sources in keyword-driven search results. However, limitations include the inability to establish causality , the study's focus on Taipei City, Taiwan, which may limit generalizability, and the lack of consideration for other factors influencing suicide rates.

Using cross-correlation analysis, they found that search trends for terms related to suicide and depression coincided with suicide data trends. Other terms like "divorce", "complete guide of suicide", and "bipolar disorder" preceded suicide data trends by 1-2 months.

They then used stepwise multiple regression to identify which search trends were significantly associated with suicide data. Searches for "major depression" and "divorce" accounted for up to 30.2% of the variance in overall suicide data. When only considering leading search trends, "divorce" and the pro-suicide term "complete guide of suicide" accounted for 22.7% of variance.

The authors suggest their findings provide preliminary evidence that online searches for suicide information could indicate suicidal behavior. They propose filtering potentially harmful pro-suicide content in search results as a prevention strategy.