A Text-Based Analysis of Measles Outbreak Narratives in CDC Reports and Their Associations With State Demographic Patterns

Group 4: Jamila Sani and Kehinde O. Adeniyi

2025-12-10

GitHub: https://github.com/kadeniyi-bot/Adeniyi-Jsani-727.git

Background

Over the past year, measles outbreaks have increased in the United States, and public debates about these outbreaks and the vaccines needed for prevention have grown more divisive. States like Texas have seen a rise in measles-related deaths, especially among children. According to the Harvard Health Publishing report, there have been 1,753 confirmed cases across 42 states, many involving children, partly due to political pressure that has led to declines in vaccination rates in some states. This is the largest number of cases reported in a single year since 2000, when measles was declared eliminated in the United States. Given this growing concern, it is important to examine how measles outbreaks are described in Centers for Disease Control and Prevention (CDC) reports and how these narratives may relate to demographic patterns across states.

This project aims to analyze how measles outbreaks are described in official CDC reports and explore how these outbreak patterns relate to the demographic characteristics of U.S. states. Instead of studying public sentiment, the focus is on public health reporting patterns, extracting state-level outbreak mentions from CDC texts and comparing their geographic distribution with demographic data from the American Community Survey (ACS). The goal is to identify which states are most frequently mentioned in outbreak reports, understand the themes that emerge in CDC reporting, and determine if demographic factors help explain the distribution of outbreaks across states.

Text analysis techniques, including keyword frequency and topic modeling, were applied to identify predominant outbreak themes. Topic modeling was used to explore thematic patterns within comments scraped from CDC measles outbreak reporting webpages. Keyword frequencies were obtained from state outbreak mentions extracted and quantified to identify which states appeared most frequently in CDC communications. These state-level counts were then

merged with ACS demographic data to examine relationships between education, population characteristics, and outbreak representation. Visualizations created with ggplot2 highlighted geographic and thematic patterns across states.

Research Questions

We addressed three questions of interest for this study:

1. Which states were most frequently mentioned in CDC measles outbreak reports?
2. What major themes appear in CDC measles outbreak narratives based on text analysis?
3. How do state-level demographic characteristics from ACS data relate to outbreak representation in CDC reports?

Data Structure and Description

Our data were obtained primarily from two independent sources:

- CDC measles outbreak webpages: measles outbreak texts containing detailed descriptions of confirmed cases, affected states, risk factors, and temporal trends were web-scraped from relevant CDC webpages.
- ACS API data: state-level demographic variables, specifically education and population, were retrieved from the U.S. Census American Community Survey (ACS) through the Census API.

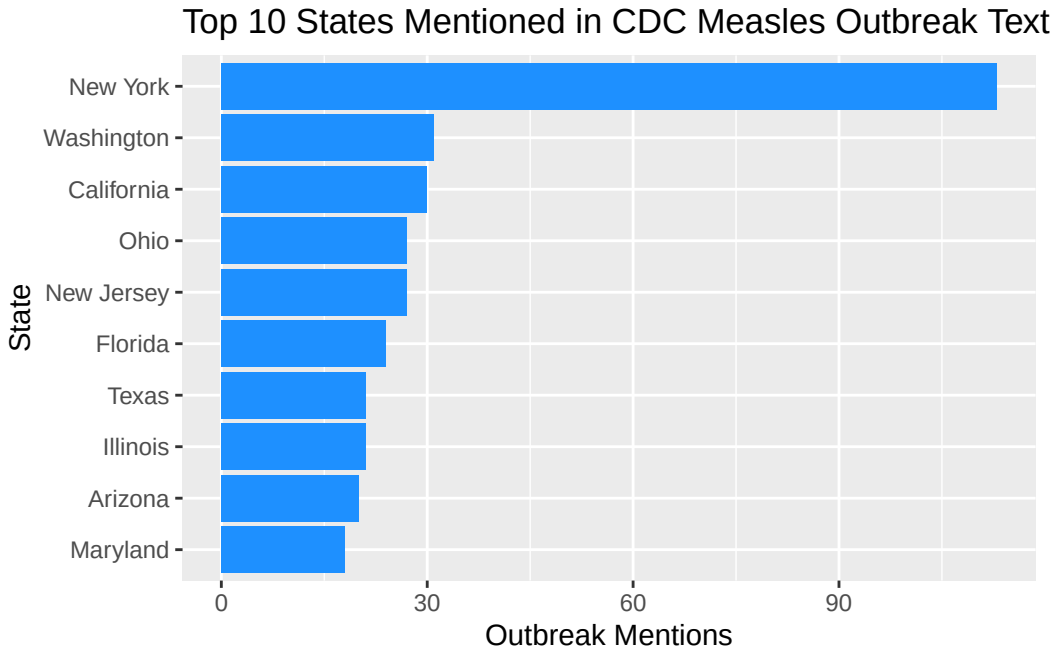
The scraped web pages (URLs) and the corresponding number of rows retrieved are as follows:

https://www.cdc.gov/measles/data-research/index.html - 281 rows

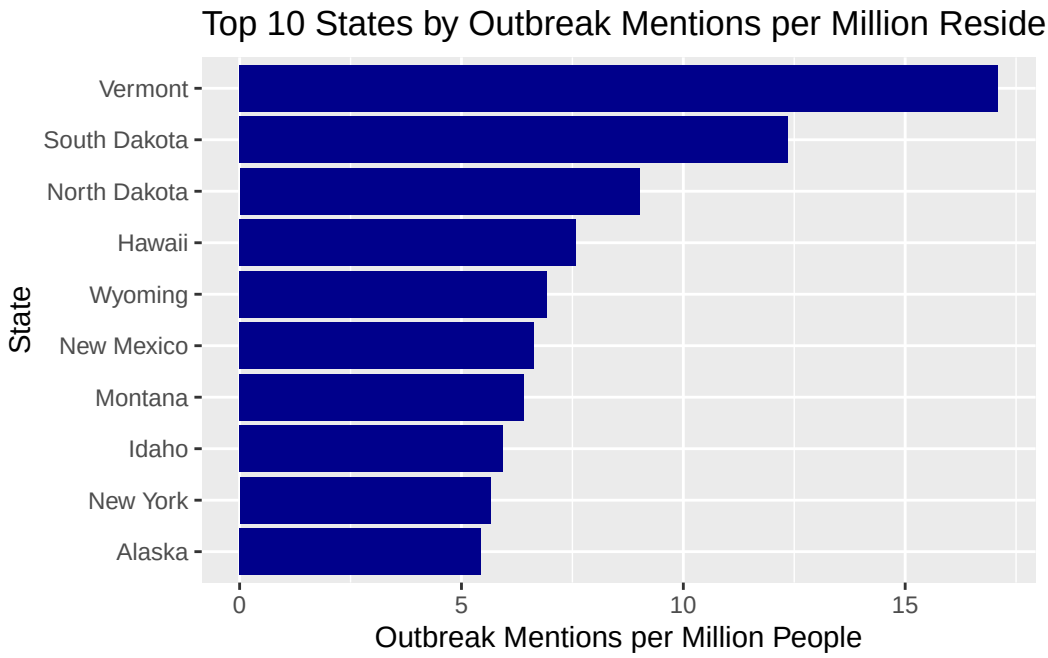
https://www.cdc.gov/measles/index.html - 86 rows

After scraping measles-related text from CDC webpages and retrieving demographic variables from the ACS, we carried out a series of data cleaning and preprocessing steps to ensure the final dataset was accurate, consistent, and meaningful for analysis. Our first step involved cleaning the CDC text data. The initial information we scraped from the CDC webpages included many irrelevant details.

Data cleaning and preprocessing steps to prepare the scraped text for analysis included removing stop words, such as the word "measles," which appeared in nearly every document and would otherwise dominate the topics during modeling. Additional preprocessing steps included converting text to lowercase, tokenization, removing punctuation and numbers, lemmatization or stemming, eliminating boilerplate website and navigation junk, and extracting relevant blocks.



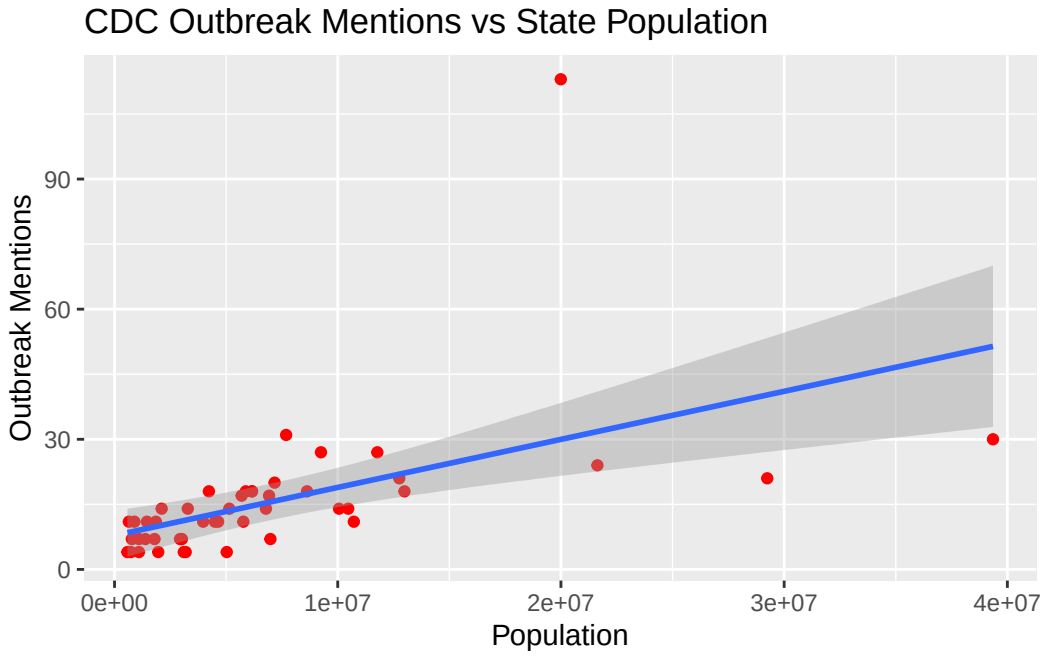
The above plot illustrates the top ten states mentioned in the CDC measles outbreak text from our scraped data. New York had the highest number of mentions (113), significantly surpassing all other states. The next most frequently mentioned states were Washington with 31 mentions and California with 30 mentions. Aside from New York, which had more than three times the mentions of any other state, the remaining states had relatively similar counts. Densely populated states like New York and larger states like Washington and California appear most often, reflecting repeated state level outbreaks reported across several years hence the likelihood of increasing its prominence in CDC reporting.



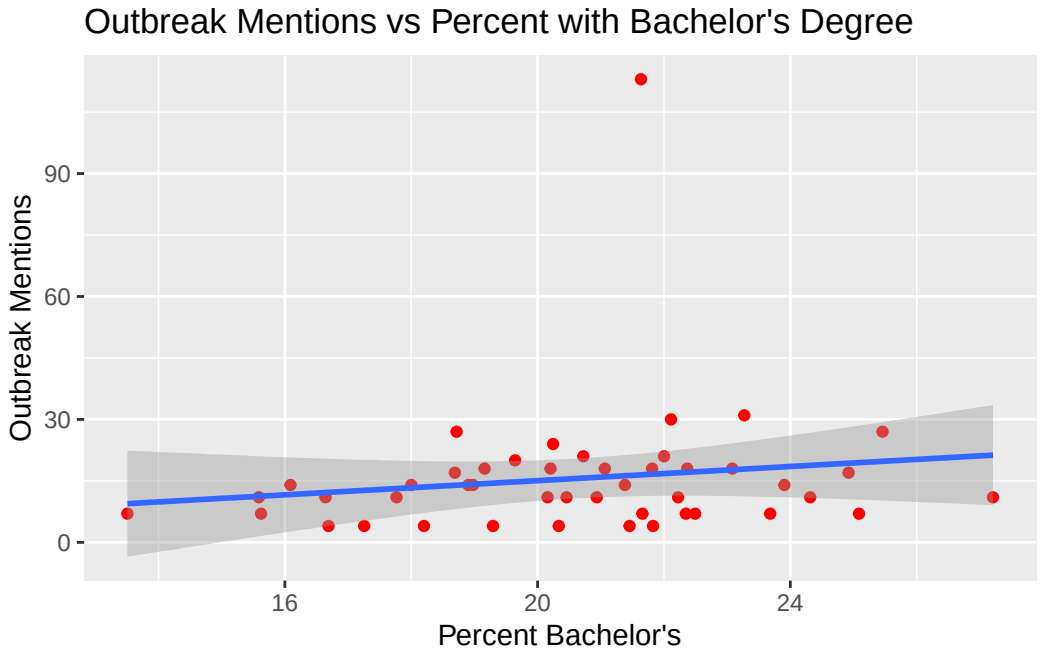
The graph above shows outbreak mentions per million residents. After adjusting for state population size, smaller states such as Vermont and South Dakota appear most frequently per capita. This shows that outbreak impact is proportionally larger in certain low population states even when total mentions are lower.

When outbreak mentions are scaled per million people, there is a dramatic shift in the top ten states. New York which was initially the top state now ranks near the bottom in ninth place. This indicates that, although these states have smaller populations, their reported outbreaks have a larger impact relative to their population size. In other words, outbreak activity is proportionally higher per capita in certain low-population states, even when their total number of mentions is lower compared to populous states.

Also, we created scatter plots to examine the relationships between CDC outbreak mentions and demographic characteristics of states.



The first scatterplot above displays the relationship between each state's population size (x-axis) and the outbreak mentions (y-axis). The blue regression line fitted to the data indicates a positive linear relationship, meaning that as a state's population increases, its number of outbreak mentions also tends to increase. This suggests that more highly populated states are mentioned more frequently in CDC measles updates. As expected, New York which had a substantially higher number of mentions compared to other states, appears as the outlier observed in the plot.



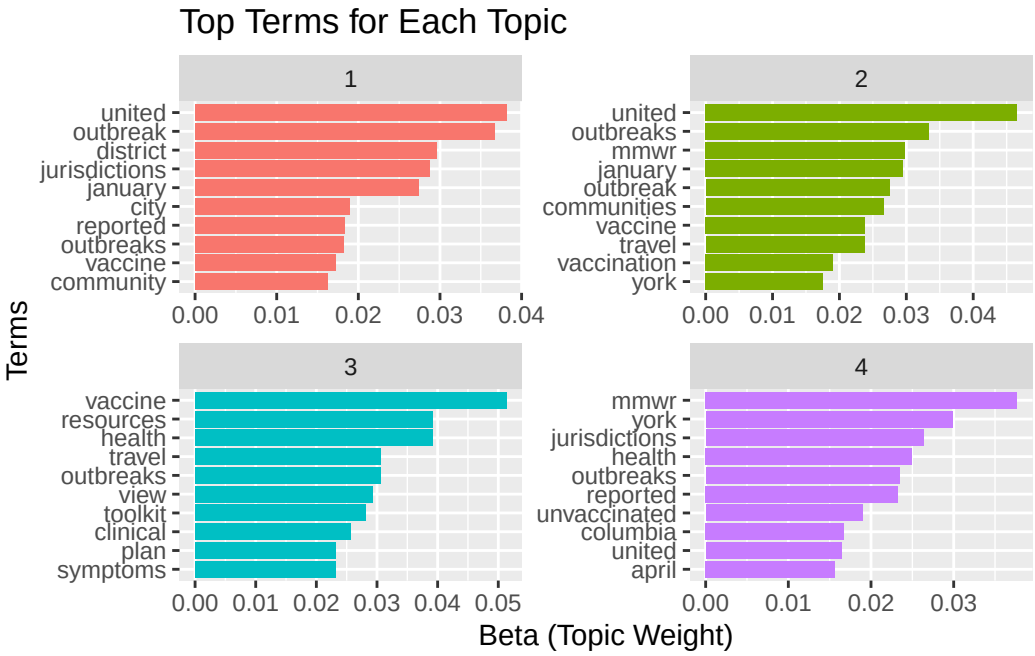
The second scatter plot examines the relationship between outbreak mentions and the percentage of persons with a bachelor's degree (education variable). For this analysis, education levels were standardized by calculating the proportion of the educated population that holds a bachelor's degree. The plot shows a nearly flat horizontal regression line, indicating a weak or negligible relationship between the two variables. This suggests that education level has little correlation with the frequency of mentions in CDC outbreak reports.

Methodology

Word Frequency Analysis and Topic Modeling

We conducted a word frequency analysis to identify the most common terms in the CDC outbreak text, which helped us understand the concepts and language frequently used in measles reporting. The graphs below shows the top ten terms for each topic.

Additionally, to examine underlying patterns, we used Latent Dirichlet Allocation (LDA) for topic modeling. We fit an LDA model with k = 4 topics, choosing this number based on observed patterns in the text and the interpretability of the resulting topics. The topic modeling identified four distinct themes in the CDC outbreak text. Each topic was interpreted using its highest weighted terms.



The themes that emerged from the analysis are shown in the table below.

topic	label	top_terms	summary
1	National Outbreak Summaries	united, outbreak, district, jurisdictions, january, city	National case reporting and jurisdiction summaries.
2	Vaccination Guidance and Public Health Resources	united, outbreaks, mmwr, january, outbreak, communities	Public health guidance, vaccination, and prevention resources.
3	CDC Surveillance and MMWR Reporting	vaccine, resources, health, travel, outbreaks, view	Surveillance updates and MMWR reporting.
4	State Level Outbreaks and Vaccination Status	mmwr, york, jurisdictions, health, outbreaks, reported	State specific outbreak reporting and vaccination details.

Results

The following themes for each topic were observed:

Topic 1: National Outbreak Reporting and Jurisdiction Summaries

This topic reflects national-level outbreak reporting, including summaries of affected jurisdictions, monthly updates, and general communication about reported cases across states and districts.

Topic 2: Surveillance Updates, Community Spread, and Travel Risk

This topic centers on CDC surveillance and MMWR reporting. It captures themes of community transmission, travel-related risk, vaccination information, and references to specific locations such as New York.

Topic 3: Public Health Guidance and Prevention Resources

This topic represents public guidance materials, including vaccination recommendations, health resources, clinical toolkit information, travel advisories, and symptom-based guidance.

Topic 4: State-Level Outbreak Details and Vaccination Status

This topic discusses outbreaks and vaccination status specific to individual states. It includes detailed references to specific jurisdictions like New York and Columbia, case reporting, and time markers such as April.

Discussion

This study highlights significant disparities in how measles outbreaks are depicted in CDC narratives, revealing that public health reporting is heavily influenced by state population size. The analysis confirmed that populous states, particularly New York, dominate the total number of CDC outbreak mentions, likely due to the sheer number of cases and the state's history of repeated outbreaks. However, when adjusting for population, a different pattern emerged. Smaller states like Vermont, South Dakota, and North Dakota had the highest mentions per capita. This indicates that while large states produce more overall reports, outbreaks in less populated areas might have a proportionally higher impact on their communities or prompt different reporting protocols relative to their size.

The text analysis further revealed four distinct topics, ranging from national surveillance summaries to specific vaccination guidance, which shows that the agency's focus is on monitoring transmission and distributing public health resources. Additionally, the comparison with ACS demographic data indicated a strong positive linear relationship between population size and outbreak mentions, reinforcing the link between high-density areas and reporting frequency. Interestingly, the analysis found no significant correlation between a state's education level (percentage of residents with a bachelor's degree) and the frequency of outbreak mentions. This suggests that a state's prominence in CDC reports is driven more by population size than by its educational characteristics.

Limitations

Several limitations should be considered when interpreting these findings.

First, the data were restricted to text scraped from specific CDC webpages, which might not include all measles reporting or outbreak communications shared through other channels, such as social media, internal state reports, or non-digital press releases. Consequently, the "mentions" quantified here represent the CDC's web-based outbreak text rather than a purely epidemiological count of cases hence not generalizable to the public.

Additionally, a specific challenge emerged in delineating the geography of New York narratives. The text extraction process relied on keyword matching, which may not have effectively distinguished between references to "New York State," "New York City," or the general term "New York." Consequently, the large number of mentions linked to New York State probably combines city-specific outbreak reports with broader state-level data. This ambiguity makes it difficult to determine whether the main narratives are driven by the unique urban density and international travel hubs of New York City or if they reflect broader state-wide trends, potentially misleading the interpretation of demographic patterns for this region.

Also, the demographic analysis was conducted at the state level, and by aggregating education and population data across entire states, the study may have masked localized demographic patterns specific to the communities where outbreaks occurred, such as counties or neighborhoods with lower vaccination rates.

Additionally, the observed correlation between population size and mentions cannot be used to establish causality for why certain states are featured more prominently in outbreak narratives beyond demographic scaling.

Finally, we decided to remove "measles" as a stop word after our initial analysis where it dominated every topic. While this step was necessary to reveal distinct themes, it may have inadvertently distorted meaningful information depending on specific context.

Conclusion

The merged CDC and ACS data provided insights to raw mention counts and demographics, which showed that larger and more populous states tended to appear more often, consistent with their population sizes, higher incidence/prevalence rates, and higher likelihood of reporting measles cases. However, a normalized metric (mentions per million persons) which was calculated to allow for a robust comparison across states of different sizes, showed that population size alone is not a complete indicator and may obscure other nuances. In addition to raw counts, more comprehensive assessment examine demographic structure, socioeconomic characteristics, geographic distribution and behavioral patterns maybe necessary.

When CDC outbreak mentions were compared with ACS demographic variables, a clear positive relationship was observed between state population and the number of outbreak mentions. Populous states generally had more outbreak references. However, the relationship between outbreak mentions and education level was weak, indicating that educational attainment (using Bachelor's degree as standard) has minimal to no influence on how often a state appeared in CDC outbreak reporting.

Topic modeling revealed four distinct themes interpreted using its highest weighted terms in the CDC outbreak texts. The observed themes seem consistent with the CDC's priorities of providing health guidance and preventive measures to the American public.

References

Centers for Disease Control and Prevention. (2025, December 3). *Measles cases and outbreaks*. https://www.cdc.gov/measles/data-research/index.html

Centers for Disease Control and Prevention. (n.d.). *Measles (Rubeola)*.https://www.cdc.gov/measles/index.html

Harvard Health Publishing. (2025, November 21). *Measles is making a comeback: Can we stop it? Seven things to know about the recent measles outbreaks*.https://www.health.harvard.edu/blog/measles-is-making-a-comeback-can-we-stop-it-202503063091