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

Jamila Sani & Kehinde Adeniyi

SUMMARY OF PROPOSAL

This project attempts to analyze how measles outbreaks are described across official CDC reports and explore how these outbreak patterns relate to demographic characteristics of US states. Instead of examining public sentiment, the study focuses on public health reporting patterns, extracting state level outbreak mentions from CDC text and comparing their geographic patterns with demographic variables from the American Community Survey (ACS). The goal is to understand which states appear most frequently in outbreak narratives, what themes emerge from CDC reporting, and whether demographic factors help explain the distribution of outbreaks across states.

The analysis uses two independent data sources. The first is CDC measles outbreak text scraped from relevant CDC webpages, which provides detailed descriptions of confirmed cases, affected states, risk factors, and temporal trends. The second is ACS demographic data, which includes population size, education levels, income, and race distribution for each state. Together, these sources allow for a structured comparison between outbreak reporting and demographic context.

Text analysis techniques, including keyword frequency and topic modeling, will be applied to identify predominant outbreak themes. State mentions will be extracted and quantified to determine which states appear most often in CDC communications. These state level counts will then be merged with ACS demographic data to explore relationships between population characteristics and outbreak representation. Visualizations produced with ggplot2 will highlight geographic and thematic patterns across states.

The questions of interest may include:

1. Which US states are most frequently mentioned in CDC measles outbreak reports?

2. What major themes appear in CDC outbreak narratives based on text analysis?
3. How do state level demographic characteristics from ACS data relate to outbreak representation in CDC reports?

Scraping data from CDC

```
library(rvest)
library(httr)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter()         masks stats::filter()
x readr::guess_encoding() masks rvest::guess_encoding()
x dplyr::lag()            masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(xml2)

fetch_page <- function(url) {
  tryCatch({
    res <- GET(
      url,
      add_headers(
        "User-Agent" = "Mozilla/5.0 (Windows NT 10.0; Win64; x64)",
        "Accept-Language" = "en-US,en;q=0.9"
      ),
      timeout(20)
    )
    if (status_code(res) != 200) {
      message("Status ", status_code(res), " for: ", url)
      return(NULL)
    }
    read_html(content(res, as = "text", encoding = "UTF-8"))
```

```
  }, error = function(e) {
    message("Error loading: ", url)
    return(NULL)
  })
}

extract_text <- function(page) {
  if (is.null(page)) return(character(0))
  xml_find_all(page, ".//script") %>% xml_remove()
  xml_find_all(page, ".//style") %>% xml_remove()
  raw <- page %>% html_nodes(xpath = "//*") %>% html_text(trim = TRUE)
  raw %>%
    str_squish() %>%
    discard(~ .x == "") %>%
    unique()
}

scrape_cdc_page <- function(url) {
  page <- fetch_page(url)
  if (is.null(page)) return(tibble())
  text <- extract_text(page)
  tibble(url = url, text = text)
}

cdc_urls <- c(
  "https://www.cdc.gov/measles/data-research/index.html",
  "https://www.cdc.gov/measles/index.html"
)

cdc_data <- map_df(cdc_urls, scrape_cdc_page)
cdc_clean <- cdc_data %>%
  mutate(text = str_squish(text)) %>%
  filter(str_length(text) > 30)

write_csv(cdc_clean, "cdc_measles_text.csv")
print(head(cdc_clean, 20))
```

```
# A tibble: 20 x 2
   url                                                  text
   <chr>                                                <chr>
 1 https://www.cdc.gov/measles/data-research/index.html Measles Cases and Outbr~
 2 https://www.cdc.gov/measles/data-research/index.html Measles Cases and Outbr~
```

```
 3 https://www.cdc.gov/measles/data-research/index.html Skip directly to site c~
 4 https://www.cdc.gov/measles/data-research/index.html Skip directly to site c~
 5 https://www.cdc.gov/measles/data-research/index.html An official website of ~
 6 https://www.cdc.gov/measles/data-research/index.html An official website of ~
 7 https://www.cdc.gov/measles/data-research/index.html An official website of ~
 8 https://www.cdc.gov/measles/data-research/index.html Official websites use .~
 9 https://www.cdc.gov/measles/data-research/index.html Official websites use .~
10 https://www.cdc.gov/measles/data-research/index.html A .gov website belongs ~
11 https://www.cdc.gov/measles/data-research/index.html Secure .gov websites us~
12 https://www.cdc.gov/measles/data-research/index.html A lock ( ) or https:// ~
13 https://www.cdc.gov/measles/data-research/index.html Measles (Rubeola) Explo~
14 https://www.cdc.gov/measles/data-research/index.html Measles (Rubeola) Explo~
15 https://www.cdc.gov/measles/data-research/index.html Explore This Topic Sear~
16 https://www.cdc.gov/measles/data-research/index.html For Everyone About Symp~
17 https://www.cdc.gov/measles/data-research/index.html For Everyone About Symp~
18 https://www.cdc.gov/measles/data-research/index.html For Everyone About Symp~
19 https://www.cdc.gov/measles/data-research/index.html About Symptoms and Comp~
20 https://www.cdc.gov/measles/data-research/index.html Health Care Providers C~
```

```
view(cdc_clean)
```

```
library(tidyverse)

# Load the CSV
cdc <- read_csv("cdc_measles_text.csv")
```

```
Rows: 367 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): url, text

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Basic structure
glimpse(cdc)
```

```
Rows: 367
Columns: 2
$ url  <chr> "https://www.cdc.gov/measles/data-research/index.html", "https://~
$ text <chr> "Measles Cases and Outbreaks | Measles (Rubeola) | CDC Skip direc~
```

```r
# View a sample of rows
head(cdc, 20)
```

```
# A tibble: 20 x 2
   url                                                text
   <chr>                                              <chr>
 1 https://www.cdc.gov/measles/data-research/index.html Measles Cases and Outbr~
 2 https://www.cdc.gov/measles/data-research/index.html Measles Cases and Outbr~
 3 https://www.cdc.gov/measles/data-research/index.html Skip directly to site c~
 4 https://www.cdc.gov/measles/data-research/index.html Skip directly to site c~
 5 https://www.cdc.gov/measles/data-research/index.html An official website of ~
 6 https://www.cdc.gov/measles/data-research/index.html An official website of ~
 7 https://www.cdc.gov/measles/data-research/index.html An official website of ~
 8 https://www.cdc.gov/measles/data-research/index.html Official websites use .~
 9 https://www.cdc.gov/measles/data-research/index.html Official websites use .~
10 https://www.cdc.gov/measles/data-research/index.html A .gov website belongs ~
11 https://www.cdc.gov/measles/data-research/index.html Secure .gov websites us~
12 https://www.cdc.gov/measles/data-research/index.html A lock ( ) or https:// ~
13 https://www.cdc.gov/measles/data-research/index.html Measles (Rubeola) Explo~
14 https://www.cdc.gov/measles/data-research/index.html Measles (Rubeola) Explo~
15 https://www.cdc.gov/measles/data-research/index.html Explore This Topic Sear~
16 https://www.cdc.gov/measles/data-research/index.html For Everyone About Symp~
17 https://www.cdc.gov/measles/data-research/index.html For Everyone About Symp~
18 https://www.cdc.gov/measles/data-research/index.html For Everyone About Symp~
19 https://www.cdc.gov/measles/data-research/index.html About Symptoms and Comp~
20 https://www.cdc.gov/measles/data-research/index.html Health Care Providers C~
```

```r
# Check how many rows per URL - 367 rows total (281 and 86)
cdc %>% count(url)
```

```
# A tibble: 2 x 2
  url                                                  n
  <chr>                                            <int>
1 https://www.cdc.gov/measles/data-research/index.html   281
2 https://www.cdc.gov/measles/index.html                  86
```

Sort and arrange text entries from the two urls

```r
cdc %>%
  mutate(nchar = nchar(text)) %>%
  arrange(desc(nchar)) %>%
  select(text) %>%
  head(20)
```

```
# A tibble: 20 x 1
   text
   <chr>
 1 Measles Cases and Outbreaks | Measles (Rubeola) | CDC Skip directly to site ~
 2 Skip directly to site content Skip directly to search An official website of~
 3 Measles Cases and Outbreaks Dec. 3, 2025 Español What to know Updated on Dec~
 4 What to know Updated on December 3, 2025. The data on this page reflect conf~
 5 What to know Updated on December 3, 2025. The data on this page reflect conf~
 6 Measles cases in 2025 Resources for Communities with a Measles Outbreak: Sam~
 7 Previous years 2024 As of December 31, 2024, a total of 285 measles cases we~
 8 2024 As of December 31, 2024, a total of 285 measles cases were reported in ~
 9 What to know about measles cases & outbreak data How does CDC collect and re~
10 How does CDC collect and report data on measles cases and outbreaks? State, ~
11 Measles (Rubeola) | Measles (Rubeola) | CDC Skip directly to site content Sk~
12 Skip directly to site content Skip directly to search An official website of~
13 Measles cases in 2025 Resources for Communities with a Measles Outbreak: Sam~
14 2024 As of December 31, 2024, a total of 285 measles cases were reported in ~
15 As of December 31, 2024, a total of 285 measles cases were reported in the U~
16 Español Measles (Rubeola) About Measles is a highly contagious virus. Two do~
17 About Measles is a highly contagious virus. Two doses of the MMR vaccine pro~
18 How do measles outbreaks happen in the United States? Measles is very contag~
19 Measles is very contagious. It spreads through the air when an infected pers~
20 2019 From January 1 to December 31, 2019, 1,274 individual cases of measles ~
```

View pulled data

```r
#View(cdc)
#head(cdc)
```

Cleaning

Clean Out Boilerplate & Navigation Junk

```
library(dplyr)
library(stringr)

boilerplate_patterns <- c(
  "Skip directly to site content",
  "Skip directly to search",
  "An official website",
  "About CDC",
  "Contact Us",
  "Privacy",
  "Policies",
  "Languages",
  "Accessibility",
  "HHS.gov",
  "USA.gov",
  "CDC Archive",
  "Sign up for Email Updates",
  "Share",
  "Facebook",
  "Twitter",
  "LinkedIn",
  "Syndicate"
)

clean_cdc <- cdc %>%
  filter(!str_detect(text, str_c(boilerplate_patterns, collapse = "|"))) %>%
  filter(nchar(text) > 30)
```

Extract Outbreak Relevant Blocks - Basis for analysis

```
outbreak_cdc <- clean_cdc %>%
  filter(str_detect(text, regex("case|outbreak|measles|MMR|vaccin", ignore_case = TRUE)))

view(outbreak_cdc)
```

Extract State Names

Which states appear most often in CDC outbreak text?

```
library(stringr)

# List of states for matching
states <- state.name

# Create state regex pattern (case-insensitive)
state_regex <- paste(states, collapse = "|")

state_mentions <- outbreak_cdc %>%
  mutate(state = str_extract_all(text, regex(state_regex, ignore_case = TRUE))) %>%
  unnest(state) %>%
  mutate(state = str_to_title(state)) %>%
  filter(str_detect(text, regex("new york", ignore_case = TRUE)))
```

State mentions count - 113 mentions for New York (highest mentions)

```
state_counts <- state_mentions %>%
  count(state, sort = TRUE)

view(state_counts)
```

Pull ACS Demographic Data

Load packages

```
library(httr)
library(jsonlite)
```

```
Attaching package: 'jsonlite'

The following object is masked from 'package:purrr':

    flatten
```

```
library(dplyr)
```

API URL for Population + Income

```
url <- "https://api.census.gov/data/2022/acs/acs5?get=NAME,B01003_001E,B19013_001E&for=state

res <- GET(url)
dat <- fromJSON(rawToChar(res$content))

acs_data <- dat[-1,] %>%
  as.data.frame() %>%
  rename(
    state = V1,
    population = V2,
    median_income = V3,
    state_code = V4
  )
```

Pull Education Variables - using proportion with bachelors as standard

```
edu_url <- "https://api.census.gov/data/2022/acs/acs5?get=NAME,B15003_001E,B15003_022E&for=st

edu_res <- GET(edu_url)
edu_dat <- fromJSON(rawToChar(edu_res$content))

edu_data <- edu_dat[-1,] %>%
  as.data.frame() %>%
  rename(
    state = V1,
    edu_total = V2,
    bachelors = V3,
    state_code = V4
  ) %>%
  mutate(
    pct_bachelors = as.numeric(bachelors) / as.numeric(edu_total) * 100
  ) %>%
  select(state, pct_bachelors)
```

Merge both ACS demographics data

```r
acs_data_full <- acs_data %>%
  select(state, population, median_income) %>%
  left_join(edu_data, by = "state")

view(acs_data_full)
view(acs_data)
```

Merge CDC and ACS Data

```r
merged_data <- state_counts %>%
  left_join(acs_data_full, by = "state")
head(merged_data)
```

```
# A tibble: 6 x 5
  state          n population median_income pct_bachelors
  <chr>      <int> <chr>      <chr>                  <dbl>
1 New York     113 19994379   81386                   21.6
2 Washington    31 7688549    90325                   23.3
3 California    30 39356104   91905                   22.1
4 New Jersey    27 9249063    97126                   25.5
5 Ohio          27 11774683   66990                   18.7
6 Florida       24 21634529   67917                   20.2
```

```r
#view(merged_data)
```

```r
state_counts <- state_mentions %>%
  count(state, sort = TRUE)

merged_data <- state_counts %>%
  left_join(acs_data_full, by = "state") %>%
  mutate(
    mentions_per_million = n / (as.numeric(population) / 1e6)
  )

view(state_counts)
```

Create Normalized Metrics - mentions per 1,000,000 persons and mentions per 100,000 persons

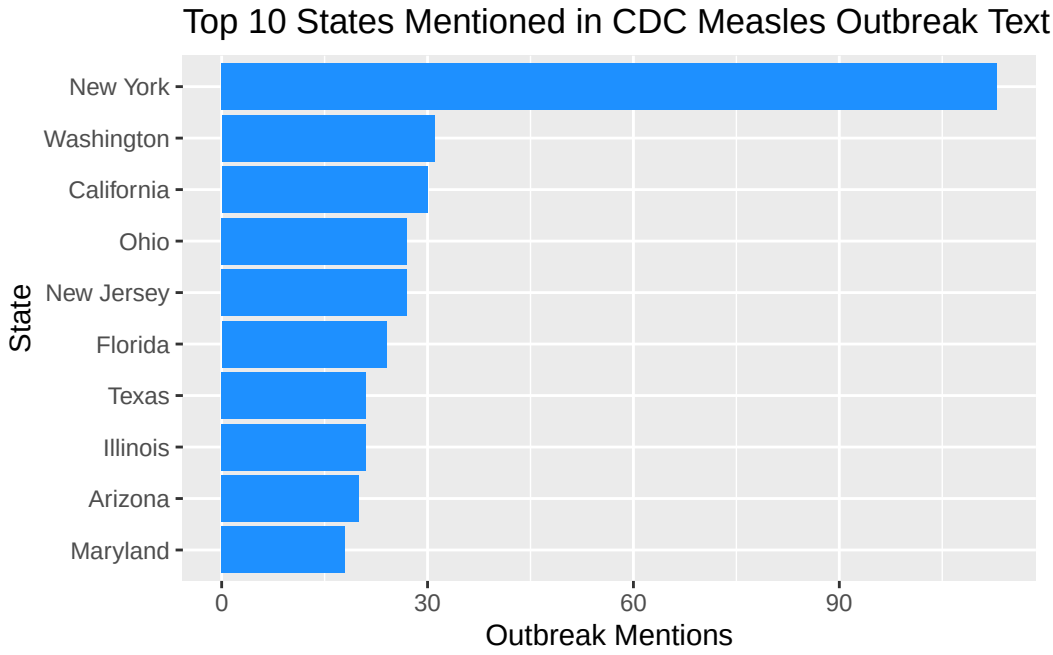
```
merged_data <- merged_data %>%
  mutate(
    mentions_per_million = n / (as.numeric(population) / 1e6),
    mentions_per_100k = n / (as.numeric(population) / 1e5)
  )
head(merged_data)
```

```
# A tibble: 6 x 7
  state          n population median_income pct_bachelors mentions_per_million
  <chr>      <int> <chr>      <chr>                  <dbl>                 <dbl>
1 New York     113 19994379   81386                   21.6                 5.65
2 Washington    31 7688549    90325                   23.3                 4.03
3 California    30 39356104   91905                   22.1                 0.762
4 New Jersey    27 9249063    97126                   25.5                 2.92
5 Ohio          27 11774683   66990                   18.7                 2.29
6 Florida       24 21634529   67917                   20.2                 1.11
# i 1 more variable: mentions_per_100k <dbl>
```

Visualize State Level Patterns

Top States by Raw Outbreak Mentions

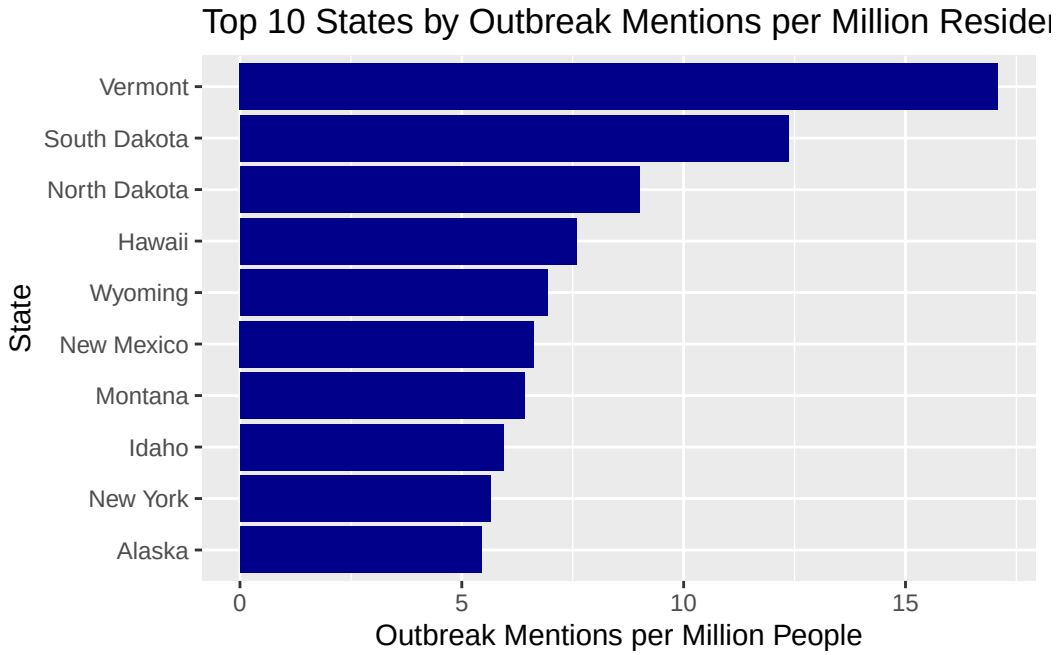
```
library(ggplot2)

merged_data %>%
  arrange(desc(n)) %>%
  slice_head(n = 10) %>%
  ggplot(aes(x = reorder(state, n), y = n)) +
  geom_col(fill = "dodgerblue") +
  coord_flip() +
  labs(
    title = "Top 10 States Mentioned in CDC Measles Outbreak Text",
    x = "State",
    y = "Outbreak Mentions"
  )
```



This plot shows the states most frequently referenced in CDC measles reporting. 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.

Mentions per Million People - VT, SD, ND, HI, WY, NM, MT, ID, NY, AK

```
merged_data %>%
  arrange(desc(mentions_per_million)) %>%
  slice_head(n = 10) %>%
  ggplot(aes(x = reorder(state, mentions_per_million), y = mentions_per_million)) +
  geom_col(fill = "darkblue") +
  coord_flip() +
  labs(
    title = "Top 10 States by Outbreak Mentions per Million Residents",
    x = "State",
    y = "Outbreak Mentions per Million People"
  )
```

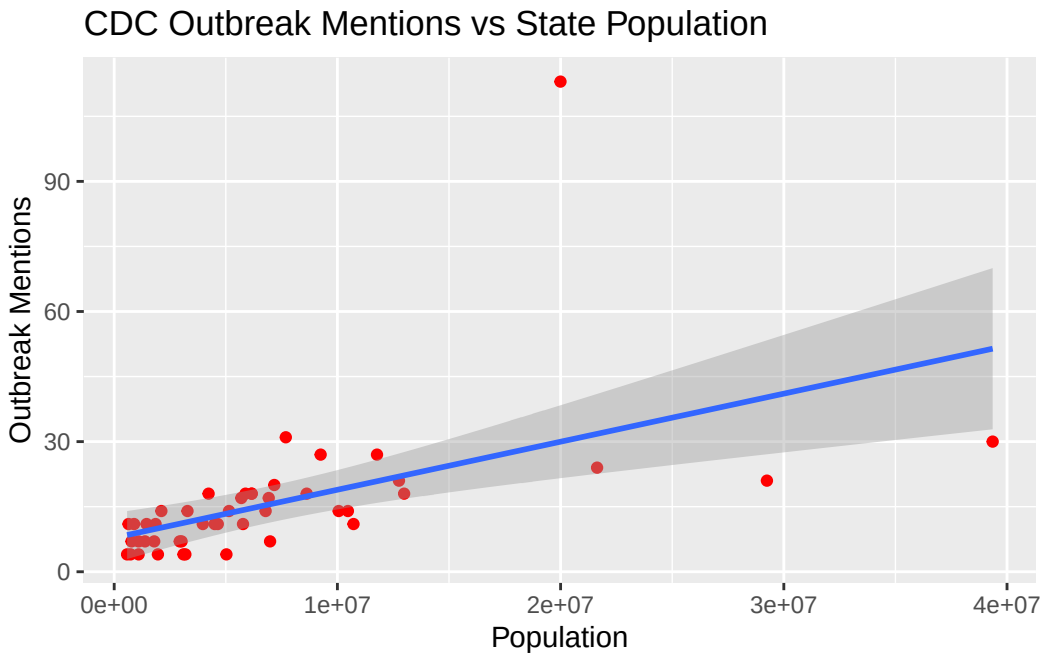


After adjusting for population size, smaller states such as Vermont and South Dakota appear most frequently per capita. This shows that outbreak impact is proportionally larger in certain low population states even when total mentions are lower.

Mentions vs Populations

```
ggplot(merged_data, aes(x = as.numeric(population), y = n)) +
  geom_point(col ="red") +
  geom_smooth(method = "lm") +
  labs(
    title = "CDC Outbreak Mentions vs State Population",
    x = "Population",
    y = "Outbreak Mentions"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```

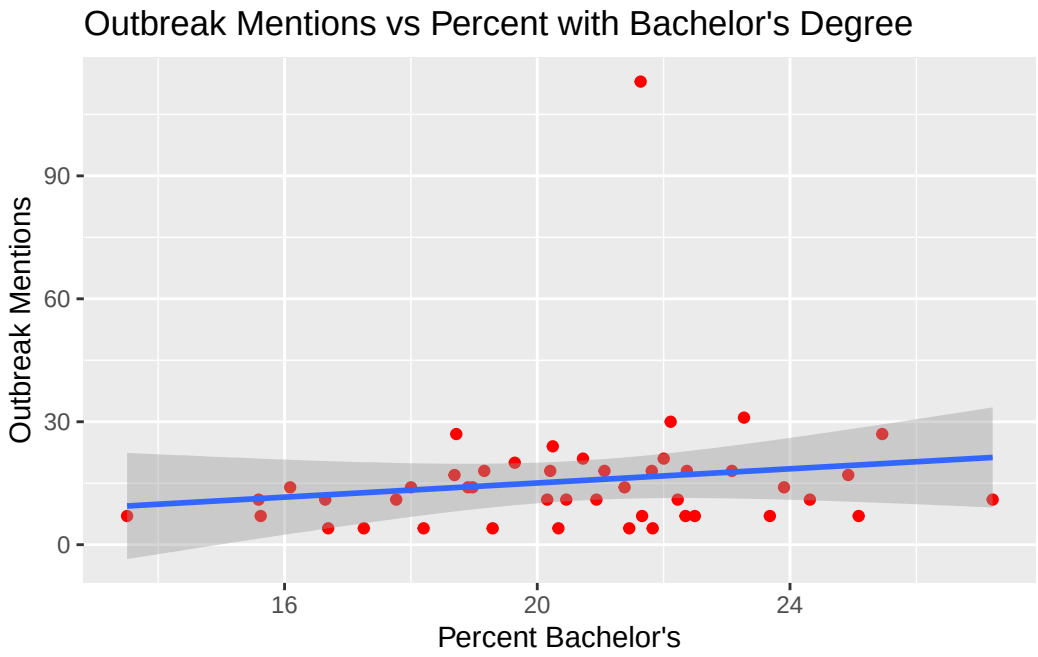


Larger states tend to appear more often in CDC outbreak reporting, which shows a positive relationship between population size and the number of reported measles events.

Mentions vs Education Level

```
ggplot(merged_data, aes(x = pct_bachelors, y = n)) +
  geom_point(col = "red") +
  geom_smooth(method = "lm") +
  labs(
    title = "Outbreak Mentions vs Percent with Bachelor's Degree",
    x = "Percent Bachelor's",
    y = "Outbreak Mentions"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Education levels only show a very weak relationship with outbreak mentions, suggesting that education is minimally associated with outbreak mentions.

Prepare Text for Topic Modelling - LDA

Topic modelling tokenization.

```
library(tidytext)

# Step 1: Basic tokenization
tokens <- outbreak_cdc %>%
  unnest_tokens(word, text)

# Step 2: Add custom stop words
custom_stop <- tibble(
  word = c("measles")
)

# Combine with default stop words
all_stops <- stop_words %>%
  bind_rows(custom_stop)
```

Keeping only meaningful words:

```
# Step 3: Remove stop words (use ALL_STOPS)
tokens <- tokens %>%
  anti_join(all_stops, by = "word") %>%
  filter(str_detect(word, "[a-z]")) %>%
  filter(nchar(word) > 3)
```

Prepare Data fo Topic Modelling (LDA)

```
# install.packages("topicmodels")
library(topicmodels)

dtm <- tokens %>%
  count(url, word) %>%
  cast_dtm(url, word, n)
```

Topic Modeling

```
lda_model <- LDA(dtm, k = 4, control = list(seed = 123))
topics <- tidy(lda_model, matrix = "beta")
head(topics)
```

```
# A tibble: 6 x 3
  topic term        beta
  <int> <chr>      <dbl>
1     1 abroad  1.71e- 3
2     2 abroad  6.20e- 4
3     3 abroad  2.32e-24
4     4 abroad  2.66e- 3
5     1 absence 1.10e- 4
6     2 absence 1.96e- 3
```

View topic keywords

```
top_terms <- topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 10)
```

```
view(top_terms)
```

Topic Modeling Results

The topic model identified four distinct themes in the CDC outbreak text. Each topic was interpreted using its highest weighted terms.

Topic 1. National Outbreak Reporting and Jurisdiction Summaries

Top terms: united, outbreak, district, jurisdictions, january, city, reported, outbreaks, vaccine, community
Summary:
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

Top terms: united, outbreaks, mmwr, january, outbreak, communities, vaccine, travel, vaccination, york
Summary:
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

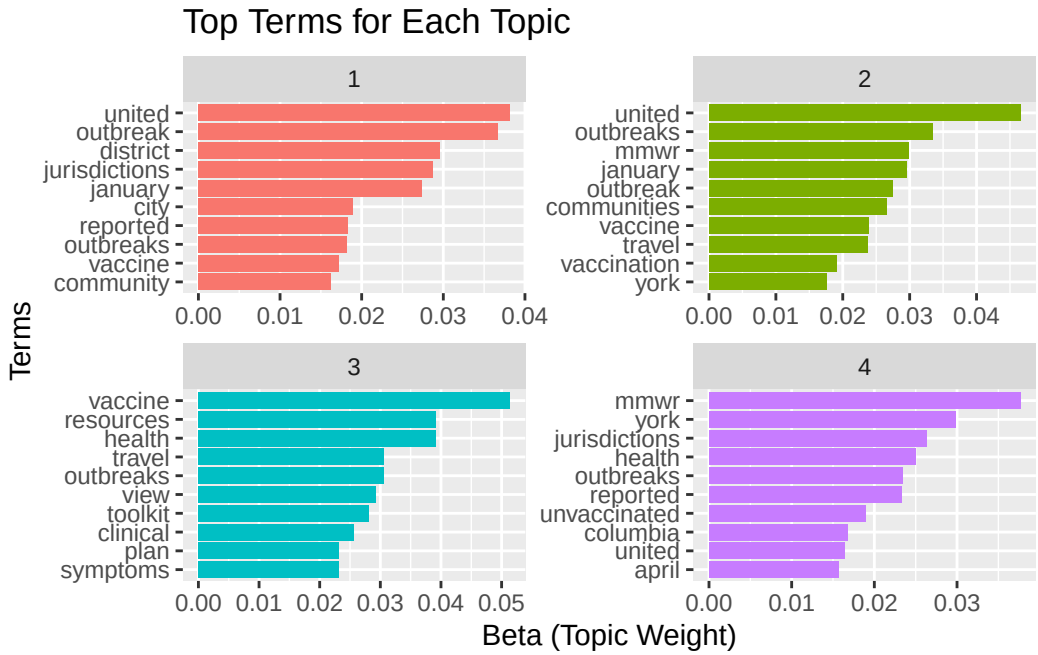
Top terms: vaccine, resources, health, travel, outbreaks, view, toolkit, clinical, plan, symptoms
Summary:
This topic represents public guidance material, including vaccination recommendations, health resources, clinical toolkit information, travel advisories, and symptom-based guidance.

Topic 4. State Level Outbreak Details and Vaccination Status

Top terms: mmwr, york, jurisdictions, health, outbreaks, reported, unvaccinated, columbia, united, april
Summary:
This topic reflects state specific outbreak reporting and vaccination status. It includes detailed references to particular jurisdictions such as New York and Columbia, case reporting, and time markers like April.

Topic Visualization: Top Words per Topic

```
library(ggplot2)

top_terms %>%
  group_by(topic) %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered() +
  labs(
    title = "Top Terms for Each Topic",
    x = "Terms",
    y = "Beta (Topic Weight)"
  )
```



Final Tables

Table 1: CDC + ACS Merged Summary Table

```
final_table <- merged_data %>%
  select(
    state,
    n,
    mentions_per_million,
    population,
    median_income,
    pct_bachelors
  ) %>%
  arrange(desc(n))

final_table
```

```
# A tibble: 46 x 6
   state          n mentions_per_million population median_income pct_bachelors
   <chr>      <int>                <dbl> <chr>      <chr>                  <dbl>
 1 New York     113                 5.65  19994379   81386                   21.6
```

```
 2 Washington    31              4.03  7688549    90325              23.3
 3 California    30              0.762 39356104   91905              22.1
 4 New Jersey    27              2.92  9249063    97126              25.5
 5 Ohio          27              2.29  11774683   66990              18.7
 6 Florida       24              1.11  21634529   67917              20.2
 7 Illinois      21              1.65  12757634   78433              22.0
 8 Texas         21              0.718 29243342   73035              20.7
 9 Arizona       20              2.79  7172282    72581              19.6
10 Maryland      18              2.92  6161707    98461              22.4
# i 36 more rows
```

Table 1 presents the combined CDC and ACS dataset, including raw outbreak mentions, normalized mentions per million residents, and key demographic variables for each state.

Table 2: Topic Modeling Summary Table

```
topic_table <- tibble(
  topic = c(1, 2, 3, 4),
  label = c(
    "National Outbreak Summaries",
    "Vaccination Guidance and Public Health Resources",
    "CDC Surveillance and MMWR Reporting",
    "State Level Outbreaks and Vaccination Status"
  ),
  top_terms = c(
    "measles, united, jurisdictions, outbreaks, January, total",
    "measles, vaccine, resources, health, travel, outbreaks",
    "measles, mmwr, outbreak, united, jurisdictions, reported",
    "measles, york, outbreaks, reported, vaccination, unvaccinated"
  ),
  summary = c(
    "National case reporting and jurisdiction summaries.",
    "Public health guidance, vaccination, and prevention resources.",
    "Surveillance updates and MMWR reporting.",
    "State specific outbreak reporting and vaccination details."
  )
)

topic_table
```

```
# A tibble: 4 x 4
  topic label                                                top_terms       summary
  <dbl> <chr>                                                <chr>           <chr>
1     1 National Outbreak Summaries                          measles, unite~ Nation~
2     2 Vaccination Guidance and Public Health Resources     measles, vacci~ Public~
3     3 CDC Surveillance and MMWR Reporting                  measles, mmwr,~ Survei~
4     4 State Level Outbreaks and Vaccination Status         measles, york,~ State ~
```

Table 2 provides a summary of the four topics identified in the CDC text. Each topic is labeled, with its key terms and a brief interpretation.

Results

State Mentions in CDC Text

CDC outbreak text included references to multiple states, with some states appearing more frequently than others. Raw mention counts showed that larger and populous states tended to appear more often, consistent with their larger population sizes and higher likelihood of reporting measles cases. After merging with ACS data, a normalized metric (mentions per million residents) was calculated to allow for a robust comparison across states of different sizes.

Demographic Comparison

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 of CDC Text

Topic modeling identified four main themes in the CDC outbreak text.
Topic 1 captured national outbreak summaries, including total case counts and the number of jurisdictions reporting cases.
Topic 2 focused on public health guidance, vaccination recommendations, and prevention resources.
Topic 3 centered on surveillance reporting and MMWR updates.

Topic 4 contained state specific outbreak information, including reported cases, vaccination status, and time based details.
These topics indicate that CDC communications encompasses national level reporting, surveillance updates, public health guidance, and specific outbreak details.

Normalized Outbreak Patterns

After adjusting for population, some smaller states (such as Vermont and the Dakotas) showed higher outbreak mentions per million residents. This indicates that population size alone does not fully explain outbreak reporting patterns. The normalized measure offers a clearer view of outbreak intensity relative to population and adds context beyond raw mention counts.

Overall, the results show that CDC outbreak reporting varies across states, with population size strongly associated with mention frequency. Topic modeling also reveals consistent themes that reflect the structure of CDC communication, including reporting, guidance, and surveillance.

Research Questions Answered

1. Which states were most frequently mentioned in CDC outbreak reporting?
Larger states such as Texas, California, and New York appeared most often, and population size showed a positive relationship with mention frequency.

2. What themes appear in CDC measles outbreak narratives?
Topic modeling identified four themes: national outbreak summaries, vaccination and public health guidance, surveillance reporting, and state specific outbreak details.

3. Do demographic factors help explain outbreak mention patterns?
Population showed a strong relationship with CDC mention frequency, while education level showed a weak or negligible relationship. Normalized values indicated that some smaller states had relatively higher outbreak intensity.