Using News Articles to Model Hepatitis A Outbreaks: A Case Study in California and Kentucky

Marie Charpignon
MIT
Cambridge, MA
mcharpig@mit.edu

Maria Mironova
Capital Health System
Trenton, NJ
mmironova@capitalhealth.org

Saeyoung Rho
MIT
Cambridge, MA
srho@mit.edu

Maimuna S. Majumder
MIT
Cambridge, MA
maimuna@mit.edu

Leo A. Celi
MIT
Cambridge, MA
lceli@mit.edu

Abstract

Although the number of hepatitis A cases has declined by 95% since the invention of the vaccine in 1995, the US has recently experienced a large outbreak. The near real-time incidence estimates are crucial for the government to project an effective vaccination rate, thus controlling the outbreak's spread. The traditional source of those estimates is routine surveillance data provided by the Center for Disease Control (CDC), which often arrives with a significant delay. Therefore, there is a need for an alternative approach. In this paper, based on the cases in California and Kentucky, we demonstrate that the outbreak dynamics modeled with the news articles data produced a comparable result to the one with CDC data, in terms of the target vaccination rate. In addition, we show how Natural Language Processing (NLP) techniques applied to health-related news content can extract insights to inform policymakers, thus further elevating a potential social impact of our work.

1 Introduction

Hepatitis A can be successfully prevented with an appropriate vaccination, albeit it is a highly contagious disease that can lead us to serious morbidity and occasional mortality. However, there was an upsurge in the incidence of the disease in the US—in September 2019, there were almost 25,000 cases reported from 30 states in the US, which is seventeen times higher than in 2015 [1, 2]. There is a possibility that the rapid spread of Hepatitis A Virus (HAV) infection can revert us to the pre-vaccine era.

The government needs a reliable prediction of the outbreak's spread to set target vaccination rates and effectively curb the outbreak, but real-time surveillance data from the CDC is often delayed. Accordingly, the use of alternative data sources has been proposed and already demonstrated some success in tracking infectious diseases outbreaks [3, 4, 5]. In this paper, we assess the potential of using news articles to model the recent hepatitis A outbreak.

2 Data and Methods

This study focuses on the outbreak dynamics in California (CA) and Kentucky (KY), the states that greatly represent the most at-risk populations. CA holds 24% of the total homeless population in the

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

US, while KY is among the top five states in terms of drug overdose-related deaths with 54 counties (45%) receiving special CDC funding [6, 7].

A surveillance dataset from CDC represents traditional data, and hepatitis A related news articles retrieved from HealthMap are used as a non-traditional data source. The Incidence Decay and Exponential Adjustment (IDEA) model was applied to both datasets and its performances were compared [8]. In addition, we analyzed the body of the news articles with NLP techniques and measured the degree of language similarity to provide further insights to health policymakers.

2.1 Data

CDC WONDER Dataset CDC Wide-ranging Online Data for Epidemiologic Research (WONDER) is a search engine to select datasets from the CDC. We extracted the weekly reports from 4/3/2017 to 3/31/2019 on hepatitis A incidence. These are voluntarily submitted by local health departments to the National Notifiable Diseases Surveillance System.

HealthMap News Articles HealthMap is a publicly available database that contains disease-related news articles. From 3/24/2017 to 3/31/2019, a total of 568 HealthMap news articles from statewide outlets were obtained that had a mention of the current hepatitis A outbreak. Sources with coverage at the county-level only were removed and the number of cases reported in the news was gathered for the purpose of the IDEA analysis. For each news article from CA and KY, the content was parsed for the text analysis. There were 25 articles from CA and 85 articles from KY, which aggregate to 1,841 unique terms and 2,915 unique terms, respectively.

2.2 Incidence Decay and Exponential Adjustment (IDEA) Model

The IDEA model is a single-equation model used for short-term epidemiological forecasting. It depends on two parameters: the basic reproduction number R_0 and a discounting factor d. R_0 represents the number of successful transmissions per infected person when she first enters a completely susceptible population and describes initial exponential growth of an outbreak. The formula below shows the cumulative incidence (I) over time (T) as a function of R_0 and d. T denotes the number of time points between the start and end of data collection process. In this scenario, the reporting frequency is in weeks.

$$I(t) = \sum_{t=0}^{T} \left(\frac{R_0}{(1+d)^t} \right)^t$$

Note that it is appropriate to use the parametric IDEA model when R_0 is reasonably low (less than 5), which is the case for hepatitis A virus propagation. To calibrate the model parameters, we applied a non-linear optimization procedure and fitted the theoretical representation to empirical data using Python *scipy.optimize.curve_fit* method.

2.3 Text analysis

All news articles were aggregated in a bag-of-words model to make them ready for the text analysis. Important preprocessing steps include sentence tokenization, stopping word removal and word lemmatization. These steps were accomplished in Python using the *word_tokenize* function from *nltk.tokenize*, the list of English stopwords as provided by *nltk.corpus*, and the *WordNetLemmatizer* module from *nltk.stem*, respectively. Note that punctuation marks were removed, as well as tokens that were not alphabetic. We further measured the intersection of the two bag-of-words model for CA and KY using the Spearman rank correlation to estimate their similarity based on the words' relative frequency.

3 Results

3.1 IDEA Model

Using CDC Data: Robust to the choice of serial interval The IDEA model with CDC data produced 22% and 15% of Mean Absolute Percentage Error (MAPE) for CA and KY, respectively,

using a 3-week serial interval. When the serial interval length increased from 3 to 4 weeks, the MAPE increased to 24% and 18%, for CA and KY, respectively. These goodness-of-fit checks show that the performance of the model does not strongly depend on the choice of serial interval. In Figure 1, the upper two plots represent the epidemic curve calibration using CDC data for CA (left) and KY (right).

Using News Articles: Handling missing data Unlike the number of cases reported to the CDC, the cumulative incidence extracted from news articles data suffers from missing data. Carry-forward and linear smoothing are two of the techniques that can handle it. Both strategies were applied and similar outcomes were obtained—e.g., the CA model produced a MAPE of 20% for carry-forward, and 21% for linear smoothing. In Figure 1, the lower two plots represent the epidemic curve calibration using news articles for CA (left) and KY (right), with a 3-week serial interval and carry-forward.

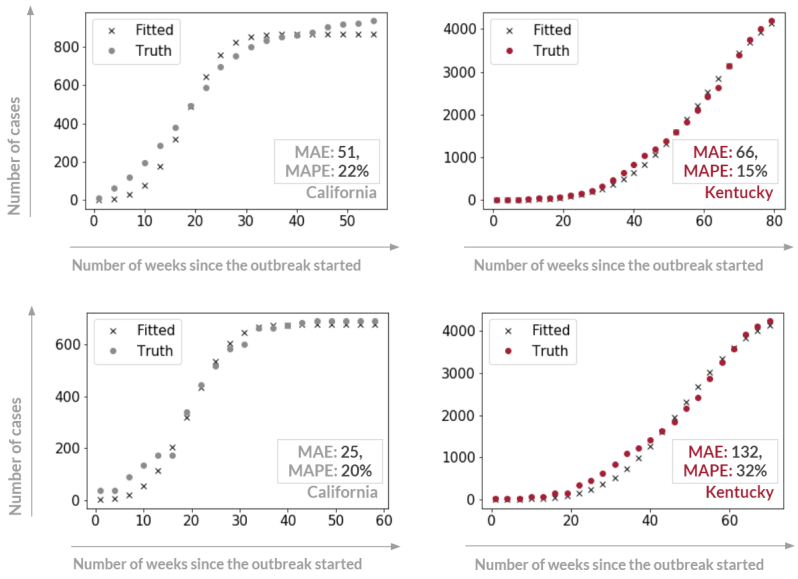

Figure 1: The epidemic curve fit using CDC WONDER (upper) and news articles (lower) with a 3-week serial interval. The carry-forward method was applied to the model using the news articles.

Parameter calibration Interestingly, the basic reproduction number was 38% higher in California than in Kentucky ($R_0^{CA} = 1.65$, $R_0^{KY} = 1.19$), signifying a much faster spread of the virus there. However, California was able to quickly curb the outbreak, with a discount factor almost ten times larger than in Kentucky ($d^{CA} = 0.0131$, $d^{KY} = 0.00142$). The reported point estimates for R_0 and d correspond to a 3-week serial interval using CDC data.

Vaccination assessment Using a 100% vaccine efficacy, the current (i.e. as of the data collection end time on 3/31/2019) vaccination rates were estimated and compared to the target thresholds to curb the outbreak. The estimates for vaccination percentages are largely different by the use of the dataset due to a higher sensitivity of the model based on news articles to missing data. However, the delta estimation (how much more vaccination is needed at this stage), which is the most important quantity in practice to inform policymakers, is quite stable (Table 1). This is because the difference between estimated threshold and estimated actual vaccination percentage is relative. As such, this difference estimator is not affected by under- or over-reporting tendencies that news articles may suffer from.

3.2 Text analysis

The intersection of the two bag-of-words model contained 1,157 terms in total (i.e. 40% vs. 63% of the language elements used in KY vs. California, respectively), which is surprisingly low considering that they are covering the same topic. Furthermore, each term of the shared vocabulary is assigned a

Table 1: Vaccination assessment in CA and KY, assuming 100% vaccine efficacy and as of the data collection end time on 3/31/2019. Columns indicate serial interval length in weeks, estimated vaccination percentage threshold, estimated actual vaccination percentage, and delta (the difference between the second and the third columns).

California (CA), CDC WONDER data			
SI	Target Vax	Current Vax	Delta
3	41%	24%	17pts
4	42%	19%	23pts

California (CA), News Articles data			
SI	Target Vax	Current Vax	Delta
3	37%	20%	17pts
4	37%	15%	22pts

Kentucky (KY), CDC WONDER data			
SI	Target Vax	Current Vax	Delta
3	17%	4.0%	13pts
4	18%	3.3%	15pts

Kentucky (KY), News Articles data			
SI	Target Vax	Current Vax	Delta
3	21%	6.3%	15pts
4	22%	5.1%	17pts

rank for both CA and KY, based on their relative frequency in each state, so as to obtain the Spearman rank correlation. Evaluated at 54% (p-value < 0.001), it points out that even when the same words are used, their intensity differs significantly. For example, the word "homeless" ranks 7^{th} in CA while ranking 52^{nd} in KY, confirming the association between homelessness and hepatitis A in CA. Also, the word "child" ranks 309^{th} in CA, whereas it ranks 21^{st} in KY. This also aligns with the fact that KY insists on children's vaccination more than CA, making it a state's priority during the outbreak.

4 Discussion

The results above demonstrate the consistency in evaluating the difference between the target vaccination percentage and the current vaccination percentage (Delta in Table 1). Since this relative statistic is a key for the government to plan for vaccination-related policy, there is a potential for news media data to be utilized as an alternative data source when an epidemic is ongoing and surveillance data is not yet available.

On the other hand, we note that the quality of the news articles may impact the results. The IDEA model for KY produced much higher MAPE when using news articles compared to when using CDC data, while the MAPE did not change significantly for CA. The difference between CA and KY could be explained by the different attitude of the local news media: the first news report in CA went out the week the outbreak was declared, whereas in KY, a 11-week delay was observed. In addition, the model outputs lower vaccination rate estimates when using news articles than when using CDC WONDER data for CA, while the situation was reversed for KY. This may suggest a scenario of over-reporting the number of cases in CA news and under-reporting in KY.

Despite those factors, the model demonstrates robustness for its major purpose of assessing the gap between an estimated vaccination percentage threshold and estimated actual vaccination percentage. At the same time, an increase in the quality of health-related news coverage will elevate the value of news media as an alternative data source.

5 Limitations and Risks

Despite its potential to estimate the relative statistic, the difference between the target and current vaccination percentage, the absolute estimates using news articles for each of those varied from the ones using CDC data. Therefore, we would like to note one caveat: this method should not be recommended to policymakers if their primary goal is to obtain the individual statistics. In addition, the availability and quality of news reporting an epidemic is another impediment for further adoption of the introduced method.

6 Conclusion and Future Work

We demonstrated that news articles can be effectively used for hepatitis A outbreak modeling to obtain the incremental vaccination that is needed to curb its progress. Using news media can bring additional benefits when it comes to understanding an epidemics. Performing text analysis and comparing the elements of language used in CA and KY revealed that news media is featuring the current US hepatitis A outbreak differently from one state to another, e.g. strikingly focusing on its connection with homelessness in CA while insisting more on immunization programs in KY. The next step is to apply this method to other US states, which requires a robust way to overcome the lack of news articles in certain geographical areas (e.g., utilizing Google Trends relative search indices to smooth the time series data with missing values) and a careful adjustment for over/under-reporting tendency.

References

[1]CDC (2017) *Viral hepatitis surveillance, United States 2016*. Atlanta, GA: US Department of Health and Human Services, CDC.

[2]CDC (2019) Widespread person-to-person outbreaks of hepatitis A across the United States. https://www.cdc.gov/hepatitis/outbreaks/2017March-HepatitisA.html. Accessed on September 5, 2019.

[3]Ghosh, S., Chakraborty, P., Nsoesie E.O. & al. (2017) Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. *Sci Rep* **7**:40841.

[4]Majumder, M.S., Kluberg, S., Santillana M. & al. (2015) 2014 Ebola Outbreak: Media Events Track Changes in Observed Reproductive Number. *PLoS Curr* **28**:7.

[5]Majumder, M.S., Santillana, M., Mekaru S.M. & al. (2016) Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015-2016 Colombian Zika virus disease outbreak. *JMIR Public Health Surveill* **2**(1):e30.

[6]Kaiser Family Foundation (2019) Estimates of Homelessness. https://www.kff.org/other/state-indicator/estimates-of-homelessness. Accessed on September 5, 2019.

[7]CDC (2018) Vulnerable Counties and Jurisdictions Experiencing or At-Risk of Outbreaks. https://www.cdc.gov/pwid/vulnerable-counties-data.html.

[8]Fisman, D.N., Hauck, T.S., Tuite A.R. & Greer, A.L. (2013) An IDEA for Short Term Outbreak Projection: Nearcasting Using the Basic Reproduction Number. *PLOS One* **8**(12):e83622.