Modeling the hepatitis A outbreak in the United States
Marie Charpignon, MSE; Maria Mironova, MD; Saeyoung Rho, MSE;
Maimuna S. Majumder, MS, PhD; Leo A. Celi, MD, MS.
1MIT, Cambridge, MA; 2Capital Health, Trenton, NJ

Introduction

Hepatitis A virus (HAV) infection, a common cause of acute hepatitis, can lead to morbidity and occasional mortality. A vaccine introduced in 1995 led to a steady decline, with annual infection incidence decreasing by 86.9% from 2001 (10,615 cases) to 2015 (1,390 cases). However, since 2016 the US has experienced a huge hepatitis A outbreak, with more than 28,000 cases reported in 30 states. Homeless persons and drug users are at most risk. Virus transmission is facilitated in dense shelter populations and by sharing drug paraphernalia [1, 2]. The rapid spread has nearly returned us to the pre-vaccine era. Lack of real-time surveillance data limits ability to predict disease dynamics and plan state level vaccination campaigns. Such concern prompted us to use non-traditional data sources, an emerging healthcare trend that has augmented traditional approaches in tracking disease outbreaks when surveillance data are missing [3, 4, 5]. Our main objective was modeling the outbreak to quantify percentage of the susceptible population needing vaccination to curb propagation. Our second goal was evaluating whether news media data are an appropriate proxy for traditional surveillance sources to accurately model real-time disease propagation. We also discovered insights revealed by news media.

Methods

Our study focused on outbreak dynamics in two states, California (CA) and Kentucky (KY), which have major high-risk homeless and drug-use populations. CA has 24% of the US homeless population. KY is among the top five states in drug overdose-related deaths, with 54 counties (45%) receiving special CDC funding [7, 8]. We applied a combination of provisional weekly reports from CDC WONDER as a structured source, or traditional data, and news articles from HealthMap as an unstructured source, or non-traditional data. CDC WONDER weekly reports and 568 HealthMap news articles, with title and content, were obtained for the period from 3/24/2017 to 3/31/2019. We used the Incidence Decay and Exponential Adjustment (IDEA) model and a non-linear optimization procedure. The IDEA model relies on the basic reproduction number \(R_0\) and a discounting factor \(d\). To calibrate their values, we fitted the theoretical representation to empirical data using Python scipy.optimize.curve_fit method. All news articles were aggregated in a bag-of-words model to make them ready for the text analysis. Important preprocessing steps included sentence tokenization, stopping word removal, and word lemmatization. These steps were accomplished in Python using, respectively, the word_tokenize function from nltk.tokenize, the list of English stopwords as provided by nltk.corpus, and the WordNetLemmatizer module from nltk.stem. We further measured the intersection of the models for CA and KY using the Spearman rank correlation to estimate their similarity based on the words’ relative frequencies.

Results

Choice of serial interval. Our procedure performs well for both states when using traditional data, as measured by the Mean Absolute Error or MAE (Figures 1, 2). We assessed the sensitivity of our calibration to the choice of serial interval value, using in turn three and four weeks. The MAE does not significantly differ from one scenario to another.

Handling missingness in non-traditional time-series data. Unlike the CDC-reported number of cases, the cumulative incidence extracted from news articles suffers from missing data. Both carry-forward (Figures 1, 2) and linear smoothing can handle this problem. In this work, we investigated both strategies and obtained similar outcomes.

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

Vaccination assessment for CA and KY. Using a 100% vaccine efficacy, the current (as of 3/31/2019) vaccination rates were estimated and compared to the target thresholds to curb the outbreak. The estimates for vaccination percentages are different 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 important for policymakers, is quite stable (Table 1). This is because the difference between estimated threshold and estimated actual vaccination percentage is relative and not affected by the under- or over-reporting tendencies that news articles may suffer from.

Text analysis. The intersection of the two bag-of-words models contained 1,157 terms in total (i.e. 40% vs. 63% of
the language elements used in KY vs. CA, respectively), which is surprisingly low considering they cover the same topic. Furthermore, the Spearman rank correlation of 54% (p-value < 0.001) points out that the intensity of even the same words differs significantly. For example, the word "homeless" ranks 7th in CA while ranking 52nd in KY, confirming the association between homelessness and hepatitis A in CA. Also, the word "child" ranks 309th in CA but 21st in KY, also aligning with the fact that KY is more insistent on children’s vaccination than is CA, making it a state priority during the outbreak.

Figures 1, 2. The epidemic curve fit using CDC WONDER (upper) and news articles (lower) with a three-week serial interval. The carry-forward method was applied to the model using the news articles.

<table>
<thead>
<tr>
<th>California (CA), CDC WONDER data</th>
<th>Kentucky (KY), CDC WONDER data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SI</strong></td>
<td><strong>Target Vax</strong></td>
</tr>
<tr>
<td>3</td>
<td>41%</td>
</tr>
<tr>
<td>4</td>
<td>42%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>California (CA), News Articles data</th>
<th>Kentucky (KY), News Articles data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SI</strong></td>
<td><strong>Target Vax</strong></td>
</tr>
<tr>
<td>3</td>
<td>37%</td>
</tr>
<tr>
<td>4</td>
<td>37%</td>
</tr>
</tbody>
</table>

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).

Discussion

Through the comparison of the use of traditional and non-traditional data for modelling the hepatitis A outbreak, we found that the latter can be an appropriate source of information – leading to similar estimates of the basic reproduction number describing the speed at which the epidemic propagates and of the discount factor signaling how fast the state is able to react and curb it. Taking a qualitative approach with the same HealthMap news dataset, we could extract more insights to be further utilized in the health policy making process. Through both sentiment analysis and word cloud visualization, we showed the power of auxiliary information obtained from non-traditional data sources. Media features the hepatitis A outbreak differently from one state to another, e.g., strikingly focusing on its connection with the homelessness issue in CA. Finally, we determined that a larger size of the news dataset does not always commensurate to an actual outbreak and leads to the increasing predictive power of a model.

References