Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States

Adam G. Dunn a,*, Didi Surian a, Julie Leask b, Aditi Dey c, Kenneth D. Mandl d,e, Enrico Coiera a

a Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia
b School of Public Health and Population Sciences, The University of Sydney, Sydney, NSW 2006, Australia
c National Centre for Immunisation Research and Surveillance, The Children’s Hospital at Westmead, Sydney 2145, NSW, Australia
d Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA 02115, United States
e Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, United States

ARTICLE INFO

Article history:
Received 8 March 2017
Received in revised form 19 April 2017
Accepted 20 April 2017
Available online 29 April 2017

Keywords:
Human papillomavirus vaccine
Social media
Content analysis
Acceptability
Vaccine refusal
Immunization coverage

ABSTRACT

Background: Together with access, acceptance of vaccines affects human papillomavirus (HPV) vaccine coverage, yet little is known about media’s role. Our aim was to determine whether measures of information exposure derived from Twitter could be used to explain differences in coverage in the United States.

Methods: We conducted an analysis of exposure to information about HPV vaccines on Twitter, derived from 273.8 million exposures to 258,418 tweets posted between 1 October 2013 and 30 October 2015. Tweets were classified by topic using machine learning methods. Proportional exposure to each topic was used to construct multivariable models for predicting state-level HPV vaccine coverage, and compared to multivariable models constructed using socioeconomic factors: poverty, education, and insurance. Outcome measures included correlations between coverage and the individual topics and socioeconomic factors; and differences in the predictive performance of the multivariable models.

Results: Topics corresponding to media controversies were most closely correlated with coverage (both positively and negatively); education and insurance were highest among socioeconomic indicators. Measures of information exposure explained 68% of the variance in one dose 2015 HPV vaccine coverage in females (males: 63%). In comparison, models based on socioeconomic factors explained 42% of the variance in females (males: 40%).

Conclusions: Measures of information exposure derived from Twitter explained differences in coverage that were not explained by socioeconomic factors. Vaccine coverage was lower in states where safety concerns, misinformation, and conspiracies made up higher proportions of exposures, suggesting that negative representations of vaccines in the media may reflect or influence vaccine acceptance.

© 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

There is a growing recognition that news and social media channels can be exploited to shape individual views through evidence and misinformation. While the value of news and social media as a population intervention to positively influence health behaviors has been examined [1,2], the impact that socially-shaped misinformation has on decision-making and health outcomes is an emerging concern. We explored a healthcare decision that appears susceptible to influence of news and social media—the choice of administering the human papillomavirus (HPV) vaccine to an adolescent.

For vaccines generally, rates of coverage are influenced by the policies and practices that make it easier or harder to access vaccines, as well as the representation of their safety and efficacy in public discourse that may affect acceptance. Socioeconomic indicators related to education, poverty, and insurance are associated with levels of immunization across countries [3,4], and these are likely to reflect differences in both access and acceptance. In countries where access to healthcare is a less important barrier, socioeconomic factors tend to display weaker correlations with immunization levels, while attitudes and beliefs play an increasingly important role [5]. This phenomenon is described as an emerging inverse relationship [6], and suggests the need to address vaccine hesitancy in public health practice [7,8]. Until now, our understanding of the relationship between the representation of vaccines in news media and acceptance has come from hypotheses generated by surveys and manual forms of media tracking [9,10].
HPV vaccines were first available in 2006, and are used to prevent the infections that can cause cervical cancers, as well as cancers of the anus, vagina, penis, and throat [11–13]. Surveys and registries have been used to estimate HPV vaccine coverage by location [14,15], but are resource intensive [16]. Around one third of females in developed countries aged 10–20 have received a full course of the vaccine [17], but there is substantial variation between and within countries, especially in the United States (US) [18,19]. Variation in both the intention to vaccinate and receipt of HPV vaccines has been linked to socioeconomic factors including income, education, race and ethnicity, and interactions with the health system [20–23].

The relationships between the representation of HPV vaccines in public discourse, vaccine acceptance, and observed vaccine coverage are complicated. Describing the reasons why HPV vaccination has fallen short of targets, the 2012 US President’s Cancer Panel report recognized that acceptance of the HPV vaccine is affected by the knowledge, attitudes, and beliefs of parents and healthcare providers [24]. Differences in exposure to information that is critical or noncritical of HPV vaccines may influence the attitudes and beliefs of the target populations, their parents, and health professionals [25]. The representation of HPV vaccine information in news and social media varies by source [26–29], and there is limited evidence to suggest that acceptance of HPV vaccines may be influenced by critical news media [30,31].

To better understand how the representation of HPV vaccines in media may influence or reflect vaccine behaviors, we asked whether populations that were disproportionately exposed to evidence, advocacy, misinformation, safety concerns, or conspiracies had higher or lower rates of HPV vaccine coverage. We chose to use Twitter to measure population-level differences in information exposure because it is the largest accessible source of information about the news and social media people read or see, and has been previously used to explain differences in health outcomes. Our aim was to determine whether state level differences in exposure to information on Twitter about HPV vaccines were associated with state level differences in HPV vaccine coverage in the US.

2. Material and methods

2.1. Study data

State level coverages of HPV vaccines were extracted from the National Immunization Surveys undertaken by the Centers for Disease Control and Prevention, which included 10,084 females and 10,743 males aged 13–17 in 2014 [18], and 10,508 females and 11,367 males aged 13–17 in 2015 [19]. Measures of poverty status, racial and ethnic composition, insurance coverage, and education levels were used as proxies for state level differences in health and socioeconomic status. These data were extracted from 2014 census produced by the US Census Bureau [32], for 50 states and the District of Columbia.

Twitter is a social media platform in which users post short messages (tweets) that may include images, videos, links to other tweets, or links to other webpages. Tweets are visible to users who follow them. We collected tweets in the period between 1 October 2013 and 30 October 2015, using the Twitter application programming interface (API) to repeatedly search for tweets that included keywords related to HPV vaccines (“Gardasil”, “Cervarix”, “hpv + vaccin/”, “cervical + vaccin/”). Shortly after the first time a new user posted a tweet meeting the search criteria, lists of followers were collected, as well as the profile information for each of those followers. This information included self-described location or geographical coordinates. Further details are provided in the Supplementary Material (Section 1.1).

2.2. Location inference

The locations of Twitter users were inferred using profile information. In Twitter, accurate location information can be found in only a small proportion of tweets that have coordinates stored in

Fig. 1. The distribution of the 219.7 million potential exposures to HPV vaccine tweets by county. Colors are given by each county's percentile among total exposure counts, ranging from 0 to 19.7 million (New York County, NY). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
The metadata of the tweet—known as geo-tags—approximately one percent of tweets [33]. As an alternative, around half of users describe a home location in a free text section of their user profile designed for that purpose. Nominatim, a gazetteer used to resolve locations from text, was used to extract the likely location of users from their user profiles (Supplementary Material, Section 1.3).

2.3. Topic modeling

Topic modeling is a common method for grouping documents that are semantically similar based on word level differences [34], and the approach has been used for applications in social media and public health [35]. The approach is a clustering method that takes the number of topics as a parameter, and produces a model that can be used to assign each tweet to a single topic. More details of the method and previous implementations of topic modeling using social media data for public health applications are provided in the Supplementary Material (Section 1.2).

Each tweet was assigned to one of 31 topics. Topic exposures for each state were estimated by counting the proportion of potential exposures in that state that came from the tweets from that topic, producing a value between 0 and 1 for each of the 51 by 31 state-topic combinations.

2.4. Multivariable model construction and evaluation

Following a similar approach to that of Eichstaedt et al. [36] multivariable models of vaccine coverage were constructed using combinations of predictors from the set of topic exposures and the census-derived socioeconomic factors. The socioeconomic factors from the US Census Bureau data included the proportions of people in each state who were African American; Hispanic of any...
race; living below poverty levels; had high school or bachelor-degree education; or had health insurance.

Elastic net regression was used to select and weight variables in the models in an attempt to produce robust and parsimonious models [37]. Fivefold cross-validation was used to measure performance in out-of-sample data to avoid bias due to over-fitting and the final models were constructed by averaging the coefficients across the five rounds (Supplementary Material Section 1.4). The outcome measures for the performance of the models were the r-squared values and the correlation coefficients (Pearson’s r) produced by comparing the logit-transformed values for HPV vaccine coverage from the National Immunization Surveys with the values predicted by the models. In total there were 20 models constructed across 4 series (males and females; in 2014 and 2015) of 5 combinations of predictors (insurance/poverty predictors, education predictors, all socioeconomic and health predictors, only Twitter-based predictors, all predictors), each using 50 or 51 units of analysis (one observation was unavailable in one survey). Performance was reported by the r-squared values, the correlation coefficients and their 95% confidence intervals, and the numbers of predictors used in the models.

3. Results

3.1. Distribution of topic exposures within the US

A total of 258,418 tweets about HPV vaccines were identified by repeated searches using the Twitter API. The total number of public Twitter users (globally) that may have been exposed to one or more of these tweets was 109.2 million. From this cohort, 31.4% (34.3 million) were identified within the US and 30.0% (32.8 million) could be localized to the state level (Supplementary Material, Section 2.1).

From a global total of 1.33 billion exposures to HPV vaccine tweets, 291.8 million potential exposures were associated with the 34.3 million users located in the US, of which 93.8% (273.8 million) could be localized to the state level, and 75.3% (219.7 million) were associated with users localized to the county level (Fig. 1). Total exposure counts ranged from 389,982 in Wyoming to 41.3 million in California. The ratio of exposures to population (also from the US Census Bureau) was highest for the District of Columbia, New York, Rhode Island, Massachusetts, Nevada, and Colorado (Supplementary Material Section 2.2).

The topics varied between generally positive topics describing new evidence or advocating for the use of HPV vaccines; mixed topics debating mainstream news media stories; and negative topics describing safety concerns, conspiracies, and politics (Fig. 2). A mainstream news media story related to a television program had the strongest overall negative correlations with coverage in females and males, and its highest proportional exposures were found in states with lower coverage (Fig. 3). Generally positive topics (reporting evidence and advocacy) tended to reach much larger audiences than topics related to safety concerns or conspiracies (Fig. 4). A newspaper story that was retracted and replaced with evidence-based responses had the strongest overall positive correlations with coverage in males, and its highest proportional exposures were found in states with higher coverage (Fig. 5).

3.2. Associations with HPV vaccine coverage

The percentage of the population without a Bachelor level education, high school education, and health insurance were most closely correlated with the proportions of females and males aged 13–17 in each state who had received at least one dose of HPV vaccine. No other socioeconomic factors were found to be significantly correlated with coverage. Among the topic exposure factors, the topics that exhibited consistently high correlations with HPV vaccine coverage across both males and females were related to mainstream news media events and issues, safety concerns, and conspiracies. Positive topics (reporting evidence and advocacy) had weaker correlations with coverage.
Models constructed using only socioeconomic factors explained 40% and 42% of the variance in initiation (at least one dose) for females and males, respectively, while models constructed using only topic exposure factors explained 63% and 68% of the variance in initiation (Table 1).

The differences between models that used socioeconomic factors and models that used topic exposure factors was more pronounced for initiation compared to completion (Fig. 6). When modeling initiation, the topic exposure models were able to explain an additional 25% of the variance in coverage compared to the models based on socioeconomic factors, for both females and males (Table 1). The increases for models of completion was lower; 17% and 12% in females and males, respectively.

When the socioeconomic factors and the topic exposure factors were combined, they marginally improved models in comparison to those that used topic exposure alone (Fig. 6). Models that combined socioeconomic factors and topic exposure factors explained ±1% of the variance in coverage compared to the models that only used topic exposure factors (Table 1).

4. Discussion

The results showed a strong correlation between HPV vaccine coverage and exposure to certain topics on Twitter, often negative topics that expressed concerns about safety or related to conspiracies. While socioeconomic and health factors are expected to capture differences in access to healthcare, proxies for information diets derived from social media data account for more of the variance in initiation (Table 1).

Other studies that have measured differences in the rates of coverage across populations have found that coverage varies by race and ethnicity, income, education, and interactions with the health system [20–22]. Our results were consistent with these studies in relation to insurance, education, and poverty—correlations were
identified between HPV vaccine coverage reported in the National Immunization Survey in 2014 and 2015, and education, income, and insurance data from US Census in 2014. The results were not aligned with these studies in relation to race or ethnicity—we found no significant correlations between HPV vaccine coverage reported in the National Immunization Survey in 2014 and 2015 and the proportions of Black and African American people or the proportion of Hispanic people of any race from the US Census. However, because our models were state level aggregates and previous studies were conducted within individual states and considered patient-level receipts of the HPV vaccine, the results are not directly comparable.

A previous study used Twitter data to predict county-level heart disease mortality by sampling and counting tweets without searching for terms related to the application domain [36]. Rather than counting the number of tweets in a topic, we examined exposure by collecting data about Twitter followers, and only used tweets that were directly relevant to HPV vaccines. While the results are not directly comparable, both studies identified correlations that

---

**Fig. 5.** The state level proportional topic exposure for Topic 19 (the Toronto Star controversy). Topic 19 exposures made up between 1.74% (Rhode Island) and 5.12% (District of Columbia) of exposures in each state and exhibited a strong positive correlation with HPV vaccine coverage.

**Table 1**
Correlations between model predictions and HPV vaccine coverage (2015 National Immunization Survey).

<table>
<thead>
<tr>
<th>Models</th>
<th>Number of predictors</th>
<th>Predictors used</th>
<th>Pearson’s R [95% CI]</th>
<th>p-value</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female aged 13–17, one dose</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance and poverty</td>
<td>2</td>
<td>1</td>
<td>0.48 [0.24–0.67]</td>
<td>3.75 × 10⁻⁴</td>
<td>0.23</td>
</tr>
<tr>
<td>Education predictors only</td>
<td>2</td>
<td>2</td>
<td>0.51 [0.27–0.69]</td>
<td>1.43 × 10⁻⁴</td>
<td>0.26</td>
</tr>
<tr>
<td>All socioeconomic/health</td>
<td>6</td>
<td>4</td>
<td>0.63 [0.43–0.77]</td>
<td>7.40 × 10⁻⁷</td>
<td>0.40</td>
</tr>
<tr>
<td>Twitter only</td>
<td>31</td>
<td>22</td>
<td>0.79 [0.66–0.88]</td>
<td>4.07 × 10⁻⁷</td>
<td>0.63</td>
</tr>
<tr>
<td>All predictors</td>
<td>37</td>
<td>22</td>
<td>0.82 [0.71–0.90]</td>
<td>1.07 × 10⁻⁷</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>Male aged 13–17, one dose</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance and poverty</td>
<td>2</td>
<td>2</td>
<td>0.53 [0.30–0.70]</td>
<td>7.01 × 10⁻⁵</td>
<td>0.28</td>
</tr>
<tr>
<td>Education predictors only</td>
<td>2</td>
<td>2</td>
<td>0.58 [0.37–0.74]</td>
<td>6.53 × 10⁻⁶</td>
<td>0.34</td>
</tr>
<tr>
<td>All socioeconomic/health</td>
<td>6</td>
<td>5</td>
<td>0.64 [0.45–0.78]</td>
<td>3.44 × 10⁻⁷</td>
<td>0.42</td>
</tr>
<tr>
<td>Twitter only</td>
<td>31</td>
<td>10</td>
<td>0.82 [0.71–0.90]</td>
<td>1.33 × 10⁻³</td>
<td>0.68</td>
</tr>
<tr>
<td>All predictors</td>
<td>37</td>
<td>18</td>
<td>0.82 [0.70–0.89]</td>
<td>2.07 × 10⁻³</td>
<td>0.67</td>
</tr>
<tr>
<td><strong>Female aged 13–17, three doses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance and poverty</td>
<td>2</td>
<td>2</td>
<td>0.62 [0.42–0.77]</td>
<td>1.15 × 10⁻⁶</td>
<td>0.39</td>
</tr>
<tr>
<td>Education predictors only</td>
<td>2</td>
<td>2</td>
<td>0.63 [0.43–0.77]</td>
<td>6.07 × 10⁻⁷</td>
<td>0.40</td>
</tr>
<tr>
<td>All socioeconomic/health</td>
<td>6</td>
<td>4</td>
<td>0.75 [0.59–0.85]</td>
<td>2.99 × 10⁻⁴</td>
<td>0.56</td>
</tr>
<tr>
<td>Twitter only</td>
<td>31</td>
<td>11</td>
<td>0.78 [0.64–0.87]</td>
<td>1.45 × 10⁻¹</td>
<td>0.61</td>
</tr>
<tr>
<td>All predictors</td>
<td>37</td>
<td>20</td>
<td>0.82 [0.70–0.89]</td>
<td>2.49 × 10⁻³</td>
<td>0.67</td>
</tr>
<tr>
<td><strong>Male aged 13–17, three doses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance and poverty</td>
<td>2</td>
<td>2</td>
<td>0.44 [0.18–0.64]</td>
<td>1.30 × 10⁻³</td>
<td>0.19</td>
</tr>
<tr>
<td>Education predictors only</td>
<td>2</td>
<td>2</td>
<td>0.47 [0.22–0.66]</td>
<td>5.51 × 10⁻⁴</td>
<td>0.22</td>
</tr>
<tr>
<td>All socioeconomic/health</td>
<td>6</td>
<td>5</td>
<td>0.53 [0.30–0.70]</td>
<td>6.88 × 10⁻⁵</td>
<td>0.28</td>
</tr>
<tr>
<td>Twitter only</td>
<td>31</td>
<td>10</td>
<td>0.71 [0.54–0.82]</td>
<td>5.07 × 10⁻⁵</td>
<td>0.51</td>
</tr>
<tr>
<td>All predictors</td>
<td>37</td>
<td>14</td>
<td>0.79 [0.53–0.82]</td>
<td>8.01 × 10⁻⁵</td>
<td>0.50</td>
</tr>
</tbody>
</table>
matched expectations and prior evidence—topics related to negative emotions like anger and stress were positively associated with heart disease mortality; and here topics related to harms, side effects, and conspiracies were negatively associated with HPV vaccine coverage.

Our methods may be used to generate further hypotheses about why people who choose not to vaccinate tend to be geographically clustered [10,38–40]. While opinions are known to cluster within communities through homophily and contagion, there has been little research previously to examine how the effect of external influences such as news and social media may be localized by community and location. The results also help reveal the limits and biases implicit in sampling information exposure via social media and suggest that even without accounting for sampling biases, proxies for information diets derived from Twitter data may be used as a signal of health behaviors that are otherwise difficult or costly to measure.

Routine systems that monitor exposure to relevant topics could be constructed to identify locations where misinformation or low-quality evidence is over-represented in the news media, and guide public health interventions to amplify high-quality evidence and guide social media interventions [2]. More precise identification of the safety concerns, misinformation, and conspiracies that are important to a segment of the public could provide a basis from which to construct targeted and cost-effective news and social media based interventions.

4.1. Limitations

Twitter users may not be representative of the general population and some Twitter users represent entities other than individuals, including organizations and non-human users. However, the relationship is measured between the representation of evidence and misinformation in the public domain and vaccination coverage, rather than an individual-level prediction of vaccine decision-making by population. Second, the method also relied on self-reported locations from Twitter users, which captures around half of the users that may have been exposed to evidence and misinformation. This is a standard approach but it could be improved through location and demographic inference methods (see Supplementary Material). Third, the study did not consider temporal patterns in exposure to topics corresponding with changes in coverage between 2014 and 2015, which limits the ability to conclude about whether changes in information exposure were lagging or leading indicators of decision-making in the population. Fourth, recommendations for two-dose models were introduced in 2016 and we considered survey data for one dose and three doses in 2014 and 2015, relating those to initiation and completion. Future epidemiological analyses would need to consider these changes. Finally, socioeconomic factors were selected based on a literature review of the evidence about observed differences in recipients of vaccines but there may be other measurable socioeconomic and health factors correlated with HPV vaccine coverage that were not included in the analysis.

Fig. 6. Correlations between the HPV vaccine coverages for initiation (at least one dose) and completion (three doses). Coverage data are from the 2015 National Immunization Survey, and coverages estimated by each of five models including combinations of socioeconomic factors and topic exposures; with 95% confidence intervals.
5. Conclusions

Measures of exposure to HPV related tweets explained more of the variance in state level HPV vaccine coverage than was explained by socioeconomic factors. Our study suggests that in states where negative opinions about HPV vaccines are popularized by mainstream media, the coverage is often lower than would be expected by socioeconomic differences alone. These differences suggest that news media may reflect, amplify, or influence vaccine acceptance, and that measures of information exposure derived from Twitter may be a surrogate indicators for localized differences in acceptance.

Conflicts of interest

No conflicts of interest.

Funding

This work was supported by the National Health and Medical Research Council (Grant No. APP1128968). The NHMRC played no role in the study design, data collection, analysis and interpretation, the writing of the report, or the decision to submit the article for publication.

Contributions

AGD, DS, JL, AD, KM, EC designed the study; AGD collected the data; AGD, DS undertook the analysis; AGD, JL, KM, AD interpreted the results; AGD drafted the article; AGD, DS, JL, AD, KM, EC critically revised the article for important intellectual content. All authors have approved the final article.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.vaccine.2017.04.080. These data include Google maps of the most important areas described in this article.

References
