Vaccine xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Evaluation of a multinational, multilingual vaccine debate on Twitter

Benedikt F.H. Becker [a,*], Heidi J. Larson [b], Jan Bonhoeffer [c,d], Erik M. van Mulligen [a], Jan A. Kors [a], Miriam C.J.M. Sturkenboom [a]

[a] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands
[b] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK
[c] University of Basel Children's Hospital, Basel, Switzerland
[d] Brighton Collaboration Foundation, Basel, Switzerland

ARTICLE INFO

Article history:
Received 8 July 2016
Received in revised form 31 October 2016
Accepted 1 November 2016
Available online xxxx

Keywords:
Vaccination programme
Pentavalent vaccine
Social media
Vaccine debate
Multilingual analysis

ABSTRACT

Background: Public confidence in an immunization programme is a pivotal determinant of the programme's success. The mining of social media is increasingly employed to provide insight into the public's sentiment. This research further explores the value of monitoring social media to understand public sentiment about an international vaccination programme.
Objective: To gain insight into international public discussion on the paediatric pentavalent vaccine (DTP-HepB-Hib) programme by analysing Twitter messages.
Methods: Using a multilingual search, we retrospectively collected all public Twitter messages mentioning the DTP-HepB-Hib vaccine from July 2006 until May 2015. We analysed message characteristics by frequency of referencing other websites, type of websites, and geographic focus of the discussion. In addition, a sample of messages was manually annotated for positive or negative message tone.
Results: We retrieved 5771 messages. Only 3.1% of the messages were reactions to other messages, and 86.6% referred to websites, mostly news sites (70.7%), other social media (9.8%), and health-information sites (9.5%). Country mentions were identified in 70.4% of the messages, of which India (35.4%), Indonesia (18.3%), and Vietnam (13.9%) were the most prevalent. In the annotated sample, 63% of the messages showed a positive or neutral sentiment about DTP-HepB-Hib. Peaks in negative and positive messages could be related to country-specific programme events.
Conclusions: Public messages about DTP-HepB-Hib were characterized by little interaction between tweeters, and by frequent referencing of websites and other information links. Twitter messages can indirectly reflect the public's opinion about major events in the debates about the DTP-HepB-Hib vaccine.

1. Introduction

Vaccination programmes are among the most effective means for improving population health. Particularly at the time of programme introduction, they tend to be accompanied by public discussion [1,2]. This may increase public awareness of the vaccine and affect the programme beneficially [3]. However, public concern may lead to reduced uptake or even jeopardize the entire immunization programme [4,5]. Therefore, detecting changes in public sentiment early is important to understand its origin and dynamics and to inform appropriate measures to investigate concerns, guide public health decision-making, or help identify issues with the vaccine or the vaccination programme.

Public attention and sentiment about vaccines have been evaluated previously by analysing different types of social-media messages and user-generated web content. Messages from the social-media platform *MySpace* were used for monitoring public sentiment about the human papillomavirus (HPV) vaccine [6]. Public news items about the HPV vaccine were shown to influence the public's awareness and opinion about HPV infection and vaccine in the United States [7]. Sentiments about an influenza vaccine shared through Twitter messages were found to correlate highly with US vaccination rates as reported by the Centers for Disease Control and Prevention [8]. International debates about vaccines and the course and drivers of public confidence have also been studied through analysis of media sources such as news sites, blogs, and governmental reports [2,9]. Twitter and other social media have frequently been used for post-marketing surveillance of pharmaceutical safety issues [10–12]. Some studies have concluded that monitoring social media is more suitable for measuring

* Corresponding author at: Department of Medical Informatics, Erasmus University Medical Center, P.O. Box 2040, 3000 DR Rotterdam, Netherlands.
E-mail address: b.becker@erasmusmc.nl (B.F.H. Becker).

http://dx.doi.org/10.1016/j.vaccine.2016.11.007
0264-410X/

public awareness of known safety issues than for providing clues about new safety signals [13].

Since 2001, a pentavalent paediatric vaccine against diphtheria, tetanus, pertussis, hepatitis B and *Haemophilus influenzae* type b (DTP-HepB-Hib) has been introduced into more than 70 low- and middle-income countries (LMICs) [14]. In a number of countries, the introduction of the vaccine was accompanied by a critical debate following a suspected association with the death of children, none of which have been deemed as causally related to the vaccine [15]. In India, a petition and a lawsuit was filed against the vaccine [16,17]. In Sri Lanka, Bhutan, and Vietnam, the market authorization for the vaccine was even temporarily suspended [18].

In this study, we explore the value of public Twitter messages to gain insight into the multinational debate on the pentavalent vaccine.

2. Methods

2.1. Data collection

We used Twitter's advanced search web interface to collect messages retrospectively. The messages were collected on 1 May 2015. The advanced search interface provides the content and date of messages from the entire history of Twitter since 2006. We queried Twitter's web API (application programming interface) to retrieve additional data fields describing the language of the content, the identity of the author, the geographical location in his or her user-profile, and the interaction status of the message (original post, repost, or reply).

The search query "*pentavalent OR pentavac OR quinvaxem*" was used to retrieve messages about the pentavalent vaccine. The query terms were selected to retrieve messages from multiple national discussions about the vaccine, but not from all national or language-specific discussions (which would have required, among others, the inclusion of country-specific brand names and slang terms). The terms "*pentavac*" and "*quinvaxem*" are brand names of the pentavalent vaccine and specific to the vaccine as such. The term "*pentavalent*" is also used in various other contexts (e.g., "pentavalent" is used in chemistry and as user name on Twitter). To remove unrelated messages, a message retrieved by the term "*pentavalent*" was only retained if it also contained the term "child" or "vaccine" (in the language of the message). The terms for *child* and *vaccine* in different languages were retrieved from OmegaWiki (http://www.omegawiki.org), a community-driven, multilingual dictionary. *OmegaWiki* provided 94 terms for *child* and 45 terms for *vaccine*. The terms came from 67 different languages.

2.2. Message analysis

A random sample of 10% of the messages was selected for manual analysis. The message tone was manually analysed to gain insights into the sentiment about the pentavalent vaccine as reflected on Twitter. The two categories of message tone – *positive*/*neutral* and *negative* – and the criteria to assign the categories were the same as previously described [9]. A message was coded *negative* if it contained any indication of concern about the pentavalent vaccine or vaccination programme, e.g., information about an adverse event that occurred after immunization, vaccine suspension, or any other factor that might have a negative effect on the vaccine programme. A message was coded *positive*/*neutral* if it contained no indication of public concern about the vaccine or vaccination programme. Non-English messages were translated with Google Translate (https://translate.google.com) while annotating. Google Translate covered the languages of all messages in the sample, and the tone was apparent from the translations for all messages.

All authors of the messages in the random sample and the 50 authors creating most messages overall were characterized as *private person*, *news site*, *health information*, *health organization*, *government*, vaccine-critical, *manufacturer*, or *non-governmental organization* (*NGO*) based on their public Twitter profile.

To characterize the use of references in the collected messages, the most commonly referred (top-level) web domains were categorised as *news site*, *social media, health information, health organization,* and *other*. Additionally, all messages from the random sample that contained references, were manually assessed if the author added own content (i.e., if the message contained more than a link to or the heading of the referred website).

We defined the geographical focus of a message by identifying the countries mentioned in the message or referred web pages. A dictionary of terms for geographical entities of countries (including cities and regions) was compiled from the *GeoNames* database (http://geonames.org) to identify mentions of countries automatically. To disambiguate terms that referred to entities in different countries, the country with the entity that had the largest population was selected. For example, *Bali* is the name of a city in India and an island in Indonesia in the *GeoNames* database. Because the population of the Indonesian island is larger than that of the Indian city, mentions of Bali were assigned to Indonesia. Messages that contributed to peaks in the message distribution over time were manually reviewed to identify the events that triggered the peaks.

The messages were analysed for occurrences of the standard format for reposts ("*RT @user*") to complement the information provided by the Twitter API. However, when evaluating public awareness and sentiment we did not distinguish between original posts and reposts, assuming that users primarily repost messages that reflect their own stance.

3. Results

We retrieved 7657 messages about the pentavalent vaccine from Twitter, of which 5771 (75.3%) from 2945 users remained after disambiguation. The number of messages grew over the years from 10 messages in 2008 to 2619 messages in 2013 (32 in 2009, 110 in 2010, 446 in 2011, and 1033 in 2012). The numbers of messages should be seen against the background of a strong growth of Twitter messages until 2012, as well as the expanded introduction of the pentavalent vaccine and incidents of public resistance in some countries. After 2013 the number of messages declined (1091 in 2014 and 430 until May 2015). A histogram of all messages per month from 2012 until May 2015 is shown in Fig. 1a.

In the manually annotated sample of 585 messages, 9 messages (1.5%) were false positives of the message retrieval and filtering, and unrelated to the pentavalent vaccine. Among the 576 messages referring to the pentavalent vaccine, 37% had a *negative* tone and 63% of the messages had *positive*/*neutral* tone. The percentage of negative messages in the random sample reached its maximum in 2014 (2010: 9%, 2011: 24%, 2012: 19%, 2013: 43%, 2014: 51%, 2015: 36%) when reports of alleged cases of severe adverse events in Vietnam dominated the overall debate. The striped bars in Fig. 1a show the estimated number of messages with negative tone. No personal experience reports with the vaccine were found in the manually annotated messages.

Fig. 2 shows the distribution of users from the random sample and of the top-50 tweeters over the user categories. In both sets most users are private persons or represent news sites. Health information sites, health organizations (including Global Alliance