PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Can Online Self-Reports Assist in Real-Time Identification of Influenza Vaccination Uptake? A Cross-Sectional Study of Influenza Vaccine-Related Tweets in the US, 2013-2017
AUTHORS	Huang, Xiaolei; Smith, Michael; Jamison, Amelia; Broniatowski, David; Dredze, Mark; Quinn, Sandra; Cai, Justin; Paul, Michael

VERSION 1 – REVIEW

REVIEWER	Alain Moren EpiConcept, Paris, France
REVIEW RETURNED	03-Jun-2018

GENERAL COMMENTS	This is a very relevant article exploring alternatives and real time data sources from social media to estimate vaccine coverage. I reviewed this article from an epidemiological point of view. Being involved in studies measuring influenza vaccine effectiveness, I found this article interesting and opening potential for new methods for measuring vaccine coverage and perhaps using it for real time effectiveness studies. However, not being a specialist of "big data analysis" I will mainly comment on the usefulness of the design. The objectives, methods and results are clearly presented and discussed both in the article and in the interesting annexes. Authors clearly discuss limitations mainly due to sample size issues. However, the optimum use of results would be to provide real time vaccine coverage according to small geographical areas corresponding to administrative areas responsible for decision making (e.g. rapid intervention to improve vaccine uptake). Authors are clearly stating the sample size limitation. Authors may want to comment on the possibility to obtain estimates by age groups, chronic conditions and any other relevant characteristics. Two major target groups for vaccination, young children and elderly people, are likely to be missed using Tweeter data. Is addition it would be interesting to check if the proposed design and analytical methods are dependent upon small population charcateristics and time of the season. Authors have chosen to compare coverage estimates with ECDC data. Rather than estimating coverage, authors may want to consider using their tweeter data source with a different design, the Lot Quality Assurance Sampling (LQAS) which, rather than

	providing an estimate of the vaccine coverage, allows to make a decision based on the probability that the expected coverage is likely to be below, above a pre-defined threshold, or in between two thresholds. This method was used by the WHO Expended Programme on Immunization (EPI) and aimed at identifying local administrative areas with potentially low vaccine coverage and then better targeted interventions. The advantage of the design is the much smaller sample size required which may allow authors to use their methods and data source to identify small local areas or, if feasible, narrower age group to target real time intervention. Beyond estimating vaccine coverage authors may also want to discuss other ways to use results. Their design may be used to compute estimates in smaller geographical areas and age group that could be linked to databases including laboratory confirmed influenza outcomes. This would allow estimating on a real time basis vaccine effectiveness using aggregated data for the case coverage (screening) method. If feasible, this would constitute a real added value allowing, very early in the influenza season, to identify clade specific VE. This would guide alternative control measures. Providing mutation specific VE very early in the influenza season would also complement the laboratory results already helping WHO to select (in February for the Northern hemisphere and September for Southern hemisphere) vaccine strains to be included in the next season vaccines. In summary, this is a very challenging study that needs to be replicated and which methods and assumptions (e.g. algorithms) need to be further reviewed by big data specialists.

REVIEWER	Adam Dunn Macquarie University, Australia
REVIEW RETURNED	07-Jul-2018

GENERAL COMMENTS	Thank you for the opportunity to review this work. The authors have processed data from repeated Twitter keyword searches to model national and state-level estimates (time-series) of influenza vaccination drawn from surveys. The major issues with the work are related to the specification of the methods and the presentation of the results. The work is interesting and will be a valuable contribution to the field. Major comments: Abstract: The abstract could be substantially improved. For example, it does not explain that the CDC data are surveys that are used to estimate influenza vaccine uptake in the United States. There is no need to explain that the methodology is "new"; instead explain what the method actually is. Data sources and their size should be explained, the number of outcome variables and methods of testing performance should be included. Correlations should be reported with 95% confidence intervals. Abstract: Be very careful when comparing the proportions/number of tweets with individual behaviours in sub-populations (i.e. the comment on women). Readers will almost certainly misinterpret this to assume that you are directly linking Twitter users to vaccination behaviour when this is a population-level study. Strengths and limitations: This should clearly specify that these are identifying correlations with population level estimates. Suggesting

the behaviours can be estimated using Twitter sounds like it is estimating individual-level behaviours.

1. Background: Social media data have not revolutionized infectious disease surveillance. There are many examples of studies using social media data to model infectious disease surveillance (including in the systematic reviews cited) but very few if any have been translated into practice to date. I would argue they have become a stream of research that has not yet been translated.

Methods: The tools for extracting information about location and demographics are fine but be very careful specifying "gender" (detected on Twitter by users self-reporting) and then conflating that with "sex" ("male" and "female") in the CDC survey results. Gender is defined differently from "sex" and these differences need to be carefully respected in public health reporting.

Methods: The methods state the use of Carmen but later in the manuscript the authors describe "geolocated" tweets. Perhaps a definition for geolocated versus inferred home locations of users would be useful to avoid confusion with geotagged tweets.

Methods: Very little detail is provided to explain how the number of tweets corresponding to "intention/receipt" are used to "predict" CDC data. It would not seem appropriate to just assume that the number of tweets were the number of receipts estimated by the surveys. If the approach was to construct a linear regression based on the signal from Twitter then it is arguably not a prediction unless it was used to predict new data from a new time period.

Results: I find it very unusual that of 1.1 million tweets referring to influenza vaccination that 367K (32.8%) were expressing that the individual user received or intended to receive an influenza vaccine. In most studies of vaccination tweets the vast majority of posts correspond to news sources and a tiny proportion correspond to intentions of individuals. Given that the search terms include "shot" this is especially surprising. For this to be true, Twitter users included in the study must have been filtered (is it that the location inference selects for human users really well?), the definition of intention/receipt must be very loose, AMT workers must have been very generous with positive labels, or the search terms must be highly specific (which they weren't). Much more detail is needed here to explain how this could have happened.

Results: Specify p-values exactly rather than classifying them by some level of significance. Regardless of the form of reporting for p-values, all correlations should be reported with 95% confidence intervals to help with interpretation (see below).

Results: Use a consistent number of decimal places or significant figures throughout the manuscript according to the journal requirements. The form of percentages and correlations are inconsistent.

Discussion: "This is one of the first studies to have utilized Twitter data to track vaccination behavior, and many of our analyses were exploratory." I think this comment is unfair and should (a) explain that it tracks population-level estimates from a survey rather than behaviour (which readers tend to read as individual-level), and (b)

	should be explained in context of other research that has used Twitter data to model the results of national surveys on vaccination. Discussion: The change in the size of data cannot be used to explain that they are the reason for the lower values found in localised correlations. If the theory is valid and Twitter is a reliable signal of vaccination in a population then it should change the confidence interval around the correlation rather than the correlation itself. This is why it is critical to report the confidence intervals around the correlations. Discussion: Parts of the discussion read like an advertisement for the approach rather than an evaluation of the approach in context of prior research. I would recommend trimming it down to a standard format of (a) what was found and were the aims met; (b) how do the results compare to the most similar types of studies; (c) what are the implications, avoiding too much speculation; (d) what were the limitations; (e) conclude. Appendix: Figures 2.1 and 2.2 seem to be equivalent to figures published in Reference 5 in the Appendix. I did not spot this reference in the main manuscript. If it covers similar work or is an extension, then that reference should be included and explained in detail in the main manuscript and referred to in the background where the authors state: "To date, efforts to track influenza vaccination through social media have been much less frequent than efforts to track disease." It should not be hidden in the appendix. In addition to the comments above, please better connect the impressive computer science work with current standards for reporting in epidemiology and public health surveillance, and seek detailed feedback on the specification of the model connecting estimates from Twitter to outcome variables from your public health expert.

REVIEWER	Soo-Yong Shin Sungkyunkwan University
REVIEW RETURNED	10-Sep-2018

GENERAL COMMENTS	This manuscript proposed a Twitter-based real-time identification method of influenza vaccination behavior. The results are quite promising and research question is also interesting. Well-written and organized. The comments are 1) Though the authors said this study was exploratory approach, the authors had better describe the results more detail. For example, why did 2013-2014 season show the lowest correlation? And why did the next season (2014-15 season) show the highest correlation? 2) Why was location not correlated compared to other factors? 3) I cannot access https://figshare.com/account/projects/31742/articles/6213878.

VERSION 1 – AUTHOR RESPONSE

Review 1

> Method. Authors may want to comment on the possibility to obtain estimates by age groups, chronic conditions and any other relevant characteristics. Two major target groups for vaccination, young children and elderly people, are likely to be missed using Tweeter data.

This is a good suggestion, and we have added a comment about limitations of estimating age (more challenging than gender), and noted as a limitation that these particular groups are underrepresented in Twitter.

> Methods. In addition it would be interesting to check if the proposed design and analytical methods are dependent upon small population characteristics and time of the season.

Thank you for the suggestions. The supplementary document contains some of these details, due to word count limitations. For example, we explore state-level vaccine in each flu season, which shows how the methods are affected by smaller populations.

> Results. Authors have chosen to compare coverage estimates with CDC data. Rather than estimating coverage, authors may want to consider using their tweeter data source with a different design, the Lot Quality Assurance Sampling (LQAS) which, rather than providing an estimate of the vaccine coverage, allows to make a decision based on the probability that the expected coverage is likely to be below, above a pre-defined threshold, or in between two thresholds.

Thank you for the suggestion. We have noted this as a possible future direction in the Discussion.

> Results. Beyond estimating vaccine coverage authors may also want to discuss other ways to use results. Their design may be used to compute estimates in smaller geographical areas and age group that could be linked to databases including laboratory confirmed influenza outcomes. This would allow estimating on a real time basis vaccine effectiveness using aggregated data for the case coverage (screening) method. If feasible, this would constitute a real added value allowing, very early in the influenza season, to identify clade specific VE. This would guide alternative control measures. Providing mutation specific VE very early in the influenza season would also complement the laboratory results already helping WHO to select (in February for the Northern hemisphere and September for Southern hemisphere) vaccine strains to be included in the next season vaccines.

We have added more discussion about how these results may be used.

Review 2

> Abstract: The abstract could be substantially improved. For example, it does not explain that the CDC data are surveys that are used to estimate influenza vaccine uptake in the United States. There is no need to explain that the methodology is "new"; instead explain what the method actually is. Data sources and their size should be explained, the number of outcome variables and methods of testing performance should be included. Correlations should be reported with 95% confidence intervals.

We have added details on CDC methods and removed references to “new methodologies”. We have also clarified the study design (cross-sectional).

> Be very careful when comparing the proportions/number of tweets with individual behaviours in sub-populations (i.e. the comment on women). Readers will almost certainly misinterpret this to assume

that you are directly linking Twitter users to vaccination behaviour when this is a population-level study.

We have clarified this throughout.

> Social media data have not revolutionized infectious disease surveillance. There are many examples of studies using social media data to model infectious disease surveillance (including in the systematic reviews cited) but very few if any have been translated into practice to date. I would argue they have become a stream of research that has not yet been translated.

We have made this clear and dampened the language.

> Methods: The tools for extracting information about location and demographics are fine but be very careful specifying "gender" (detected on Twitter by users self-reporting) and then conflating that with "sex" ("male" and "female") in the CDC survey results. Gender is defined differently from "sex" and these differences need to be carefully respected in public health reporting.

While CDC reports this variable as “sex”, the surveys behind the CDC data actually ask for “gender”, which is comparable to the variable from Twitter. We have added this explanation.

> Methods: The methods state the use of Carmen but later in the manuscript the authors describe "geolocated" tweets. Perhaps a definition for geolocated versus inferred home locations of users would be useful to avoid confusion with geotagged tweets.

We have added this clarification.

> Methods: Very little detail is provided to explain how the number of tweets corresponding to "intention/receipt" are used to "predict" CDC data. It would not seem appropriate to just assume that the number of tweets were the number of receipts estimated by the surveys. If the approach was to construct a linear regression based on the signal from Twitter then it is arguably not a prediction unless it was used to predict new data from a new time period.

We clarify that “prediction” here simply means inferring an unknown value, using standard terminology in our discipline. To avoid confusion, we have replaced “predictions” with “estimates” or “inferences”. We would also note that while this study does not do forecasting, we do show that the variables extracted from Twitter can be incorporated into ARIMA and linear regression models.

> Results: I find it very unusual that of 1.1 million tweets referring to influenza vaccination that 367K (32.8%) were expressing that the individual user received or intended to receive an influenza vaccine. In most studies of vaccination tweets the vast majority of posts correspond to news sources and a tiny proportion correspond to intentions of individuals. Given that the search terms include "shot" this is especially surprising. For this to be true, Twitter users included in the study must have been filtered (is it that the location inference selects for human users really well?), the definition of intention/receipt must be very loose, AMT workers must have been very generous with positive labels, or the search terms must be highly specific (which they weren't). Much more detail is needed here to explain how this could have happened.

A possible explanation for this discrepancy is that we do remove retweets (stated in the manuscript), which could account for much of the news sharing. We did not do other filtering, but we will point out as a sanity check that you can do a Twitter search for these terms and see that a large proportion of tweets are relevant, e.g., https://twitter.com/search?f=tweets&vertical=default&q=flu%20shot&src=typd . We also note that tweets must contain flu-related terms in addition to vaccine-related terms, so this dataset does not include “shot” in other contexts.

> Results: Specify p-values exactly rather than classifying them by some level of significance. Regardless of the form of reporting for p-values, all correlations should be reported with 95% confidence intervals to help with interpretation (see below).

We have updated our results to show 95% confidence intervals.

> Results: Use a consistent number of decimal places or significant figures throughout the manuscript according to the journal requirements. The form of percentages and correlations are inconsistent.

We have fixed this.

> Discussion: "This is one of the first studies to have utilized Twitter data to track vaccination behavior, and many of our analyses were exploratory." I think this comment is unfair and should (a) explain that it tracks population-level estimates from a survey rather than behaviour (which readers tend to read as individual-level), and (b) should be explained in context of other research that has used Twitter data to model the results of national surveys on vaccination.

We have clarified that we mean this is the first to utilize self-reports on Twitter to track vaccine uptake. We had overlooked the possibility of surveys utilizing items on self-reported Twitter use. We have also added a paragraph discussing how our research relates to other machine learning studies of vaccine behavior.

> Discussion: The change in the size of data cannot be used to explain that they are the reason for the lower values found in localised correlations. If the theory is valid and Twitter is a reliable signal of vaccination in a population then it should change the confidence interval around the correlation rather than the correlation itself. This is why it is critical to report the confidence intervals around the correlations.

We have clarified an important distinction when referring to the size of the data. Normally when constructing confidence intervals for correlations, the data size refers to the number of points being correlated, and decreasing that would increase variability of the estimate without decreasing the expected correlation, as you’ve noted. In our case, the number of points is constant, while the amount of data used for calculating the values of the points has decreased, so the points themselves have higher variability. This increases variability of the correlation confidence interval, but it also systematically decreases the expected correlation because the points themselves are noisier. This should be more clear with confidence intervals that we have added to our results.

> Discussion: Parts of the discussion read like an advertisement for the approach rather than an evaluation of the approach in context of prior research. I would recommend trimming it down to a standard format of (a) what was found and were the aims met; (b) how do the results compare to the most similar types of studies; (c) what are the implications, avoiding too much speculation; (d) what were the limitations; (e) conclude.

We have substantially rewritten the discussion section and reorganized the contents. We attempted to adhere to the structure suggested in the BMJ author’s guide.

> Appendix: Figures 2.1 and 2.2 seem to be equivalent to figures published in Reference 5 in the Appendix. I did not spot this reference in the main manuscript. If it covers similar work or is an extension, then that reference should be included and explained in detail in the main manuscript and

referred to in the background where the authors state: "To date, efforts to track influenza vaccination through social media have been much less frequent than efforts to track disease." It should not be hidden in the appendix.

We have made this reference more clear in the main manuscript.

Review 3

> Methods. Though the authors said this study was exploratory approach, the authors had better describe the results more detail. For example, why did 2013-2014 season show the lowest correlation? And why did the next season (2014-15 season) show the highest correlation?

We have added a paragraph that partially explains the poor performance in 2013-14.

> Methods. Why was location not correlated compared to other factors?

The location is an attribute of the tweet. Each tweet also has other independent attributes such as user gender, time, etc. The correlation scores of location are low because when we dive into 50 states over 4 flu seasons, the data becomes much more sparse than without considering the location. Comparing to the state level, we can observe that when we focus on the region level, which contains multiple states and more tweet data, the correlation scores go up.

> I cannot access https://figshare.com/account/projects/31742/articles/6213878.

The link will be public upon publication, but in the meantime, we have created a private link that you can access for review: https://figshare.com/s/231f70e7424401497d07

VERSION 2 – REVIEW

REVIEWER	Adam Dunn Macquarie University, Australia
REVIEW RETURNED	15-Nov-2018

GENERAL COMMENTS	Thank you for the opportunity to have a look at the revised manuscript and the response, which was clear, concise, and well-argued. An impressive and thoughtful piece of work.