Author’s response to reviews

Title: Respiratory Syncytial Virus tracking using Internet Search Engine Data

Authors:
Eyal Oren (eoren@email.arizona.edu)
Justin Frere (jfrere@email.arizona.edu)
Eran Yom-Tov (gri120@walla.co.il)
Elad Yom-Tov (eladyt@microsoft.com)

Version: 1 Date: 08 Jan 2018

Author’s response to reviews:

We thank the reviewers for their constructive comments, which helped us improve the paper significantly. We have revised our paper to incorporate all the comments we have received from the reviewers. In what follows, we explain how we addressed each and every comment we received, with references to parts of the paper that have been modified pursuant to reviewers’ suggestions.

Reviewer reports:

Nusrat Homaira (Reviewer 1):

Methods: I am unclear about how the geographical variation in different data sets used in the analyses were accounted for and then how was this interpreted, for e.g two of the data sets were only for Arizona while lab data were for five states and the NREVSS RSV data were for all 50 states and not sure how much the Google data covered.

RESPONSE: Google Trends data was available for all 50 states, as noted in Table 1. Due to the difference in geographical and temporal coverage of the datasets, each was used for different purposes: Bing and hospitalization data were used solely for finding the best correlated terms. This is because of the limitations posed by Google Trends in the number of terms that can be
queried, compared to Bing data, for which we had full access, but only for a limited time span. The terms identified in Bing were then extracted from Google Trends. Lab data was the only long-term RSV incidence data to which we had access. Therefore, we used these data to train our models. Finally, the models were tested on NREVSS from all 50 states over a one-year period. We added a paragraph at the beginning of the Methods to clarify this, stating that:

“Each of the datasets differ in geographical coverage, temporal availability, volume, and clinical accuracy. Thus, each was utilized for different purposes. Bing query data and Arizona hospitalization data were used for finding the search terms best correlated with RSV incidence. Bing data could be used for prospectively seeking these search terms, as full access to all queries on Bing was available to the investigators, in contrast to data from Google Trends, which is limited in the number of terms that can be queried. The terms identified in Bing were then extracted from Google Trends over a longer time span and for more states than that available in Bing data. We then constructed models for predicting lab-reported incidence from the frequency of searches on Google Trends. Finally, the models were tested on CDC NREVSS from all 50 states over a one-year period.”

It's surprising to me that the for a list of keywords the authors did not include ARI or ALRI (two of the most commonly used terms)

RESPONSE: We apologize for this omission. Following the reviewer's comment, we extracted these data from Google Trends. The average correlation over the 50 states between the term ‘RSV’ and the term ‘ARI’ over a period of 5 years was -0.03, and for ‘ALRI’ it was 0.07. Thus, these terms would appear to be only weakly correlated with RSV incidence and could be omitted from the model.

In estimation of the RSV incidence the authors make the assumption that similar diseases burden will result in a similar volume of use of particular keywords, did the author consider the population demographics, States with more younger children (<2 years) are likely to have larger burden of RSV diseases or is it that population demographics is pretty homogeneous.
RESPONSE: We agree that demographics probably play an important part in the estimation of RSV incidence. However, we assume that the differences in keywords use caused by such differences will be small compared to the overall use of these keywords. We addressed this issue in two ways. First, we added to the Methods that:

“We make one basic assumption in using domain adaptation, which is, that a similar disease incidence in a population will result in a similar volume of use of particular keywords. Since internet use is not uniform across demographics this assumption is only approximate. However, we assume that the differences in keyword search volume caused by different demographics will be minor relative to the overall use of the keywords.”

Second, we added to the Discussion the following text:

“Our data was analyzed without information on the demographics of those searching for information on RSV. As noted in the Methods, we make a basic assumption that similar disease incidence will result in similar use of search keywords. It is, however, likely that some of the discrepancies between actual and predicted RSV incidence are due to the differences in use of search engines by different demographics. Future work will investigate how to incorporate demographic information into a finer-grained model.”

Results: I am surprised to see in the initial filtering of keywords section key words with highest correlations there is no mention of wheeze (under symptoms group)

RESPONSE: The correlation of the term ‘wheeze’ with the term ‘rsv’ was only 0.07. We assume that this low correlation is due to the fact that wheezing could be associated with other respiratory infections, and thus has too low a specificity for the model to use.
Ryan Malosh (Reviewer 2):

Page 3, Line 19-23: I respectfully disagree that there is low urgency or recognition of the burden of RSV disease. This statement may be accurate for certain aspects of the population or specific geographic regions, but certainly not all. In particular the burden of disease in children aged < 1 year and adults > 65 years is recognized. Evidence for this recognition and urgency on the part of public health researchers is evidenced by the over 40 vaccine products in various stages of clinical testing (see the PATH RSV vaccine snapshot).

RESPONSE: We agree with the reviewer noting the urgency and recognition of the condition, particularly in recent years, and have rephrased our introduction to note that rather than low urgency, there has generally been a lack of historically available data and evidence to guide decision-making activities, particularly given that RSV is not (currently) a notifiable disease. We modified the relevant sentence to “However, while RSV disease accounts for very significant health care and social costs, there is a low level of urgency or recognition of the burden of disease lack of epidemiologically available data regarding burden of disease or its seasonality.”

Page 3, Line 51-53: My understanding is that most influenza forecasting models have limited utility in long term prediction (the best results tend to be a couple of weeks in advance and prediction seems to be worse near the peak). One of the benefits of improved forecasting of RSV circulation that the authors fail to mention is the implementation of prevention strategies (timing of vaccine administration, once licensed or administration of palivizumab). It would be of great interest to hear if/how the authors believe these findings could impact those decisions.

RESPONSE: We have added in the introduction the benefits of prevention strategy implementation through improved forecasting and also in our discussion that “in addition, transmission pattern prediction would allow for more precise implementation of disease prevention strategies, including timing administration of a new vaccine, given the numerous candidates under development, or public health messaging.”
Page 5 Line 20: Did the authors consider collecting data for RT-PCR confirmation of RSV infection? The use of clinical multiplex PCR to test for a panel of respiratory viruses (and bacteria) is expanding rapidly. As PCR is highly sensitive this could be relevant as a sensitivity analysis, especially in more recent years.

RESPONSE: While RT-PCR has become more routine, it is not available from most states (or the states where antigen data was available) as part of the NREVSS system. We considered using local data where it was available, but this would not allow us to model broader spread across states.

Page 6, Line 16: Why were lungs not included in the affected body parts? This seems particularly relevant for parents of young children who observe cough and wheezing and for older adults with underlying chronic respiratory conditions.

RESPONSE: We apologize for this oversight. Following the reviewer’s suggestion, we obtained the Google Trends data for “lungs” and evaluated the correlation between these data and RSV incidence. The average correlation is 0.22, which is relatively low. We therefore did not add it to the model.

Page 7, Line 22: A reference is needed for this bootstrap method

RESPONSE: We added the requested reference.

Page 16, Lines 22-27: I think the limitations related to RSV and Influenza co-circulation warrants further discussion. While I agree that at the weekly level this may be less of a concern the amount of co-circulation will be highly variable by year. Further, influenza and RSV are difficult to distinguish clinically and influenza is included in the search terms. Perhaps a sensitivity analysis looking at years with substantial overlap in lab confirmed RSV and influenza and examining the model with and without influenza in the search terms could shed more light on this issue.
RESPONSE: We adopted the reviewer’s suggestion and did the following: We built a model for predicting RSV from the Google Trends data for the 2010 season for the 5 states analyzed in the paper. We then applied the model to the 4 following seasons and measured the correlation between the predictions of the model and NREVSS RSV incidence and NREVSS influenza incidence. These analyses were added to the Methods (where we describe the data used in the manuscript), including Table 1; to the Results, where a new section (and Table were added); and to the Discussion. Our results indicate that the predicted RSV rate is more strongly correlated with the actual RSV rate than with the influenza rates. Indeed, the correlation between the predicted RSV rate and the influenza rate is on par with the correlation between the actual RSV rate and the influenza rate. The new section in the Results reads as follows:

“Influenza and RSV frequently co-occur in some geographies. Therefore, we tested whether the models can distinguish between searches for RSV and those for influenza. We applied the models based on the term “RSV” and the models based on the 50 terms to the data from the five states for which both RSV and influenza rates were available over several years. Models were trained using data from 2010 to 2011 and tested on data from 2012 to 2015.

The results of this analysis are shown in Table 3, together with the correlation between RSV AG rates and influenza rates. As is presented in that table, the predicted influenza rate is only slightly greater than the correlation between RSV and influenza in CDC data. However, the correlation between the predicted RSV rate and the actual RSV rate is higher than both.

These results indicate that the symptoms people search for when suffering from RSV are indeed also used by people suffering from influenza, but that the prediction model more accurately weighted the terms for predicting RSV incidence.”

Figure 1: Do the authors have hypothesized explanations for the model not dropping to zero in years 2007-2011 in OH and PA? Or for the large underestimates in CA and MI in 2010?

RESPONSE: The non-zero minima in 2007-2011 are due to a rounding performed by Google Trends on their data. The underestimates for CA and MI in 2010 may be due to local news activity in these states during that year. We note that such underestimates occur in other years.
and states, for example, CA 2014, MI 2014, OH 2015 and PA 2012. We enhanced a paragraph in the Discussion, noting these underestimates and the need to further investigate them, writing that:

“We also suggest that while a model based on a single term reaches high correlation, the full search term model is preferable because it is less likely to be skewed by effects such as media attention, which, as the case of Google Flu Trends has shown (24), can easily cause prediction models to err. Such changes in attention may be the reason for the underestimates observed in Figure 1, though future work will be required to ascertain the reason for these underestimates.”

The titles for Figure 2 and 3 seem to be flipped. Additionally, the panels in figure 3 are not labeled.

RESPONSE: Fixed, and the panels were labeled.

Editorial Policies

RESPONSE: All sections were added.