Resubmission Memo:
Survey Data and Human Computation for Improved Flu Tracking

We thank the reviewers and editor for their thoughtful comments. We have sought to address every issue highlighted by the reviewers and editor. Consistent with the editor's note in response to our original submission, we did our best to address the underlying concerns and made changes based on the editor's instructions. We found the comments very helpful and sought to address them as fully as possible.

Responses to Reviewer 2

1. Make comparison with seeking care from a healthcare provider
    i. *We agree this is an interesting comparison to make and further motivates the core findings in the paper. We added analysis to the main document and the supporting information to address it. To incorporate the idea behind this comment, we looked at differences in search rates between those who sought information about symptoms from a healthcare provider versus those who sought information from the Internet. We found that only 6% of respondents saying they asked for information from their healthcare provider and did not look online. By contrast, 33% reported looking for information online and did not ask their provider, and 26% reported looking online* ***and*** *asking their healthcare provider. We now dedicate space to this in the survey findings section of the main paper, and address this question in the supporting information (see section "Information from Healthcare Providers vs. the Internet").*

    b. Reference relevant literature on recall - specifically that there are two day periods of recall
        i. *We have added citations to literature that tackles issues of recall. We now go into greater detail to describe the problems of recall relating to the length of time between the event and the survey in the limitations section of the main paper and in the supporting information. Specifically, we cite sources showing that memory is more effective for recent events, which will create a recency effect. Dating error increases in a linear fashion as period increases - though remains unbiased. These are located in the limitations section in the main paper and in SI: 3.4.*

    c. Check for differences between respondents that completed the survey closer to the start or peak of the influenza season.
        i. *We checked whether respondents who completed the survey closer to the start of the season were more or less likely to report flu symptoms personally or in the household. We fielded an initial survey wave from March 19th to March 27th 2015, then a second wave from April 27th to April 31. We looked at differences between these two survey waves. We found no statistically significant differences in the rates of flu symptom reporting between the early and late responders. Including whether the respondent submitted their survey in the earlier cohort did not*

substantively alter the coefficient on the household flu variable in our main model either. We have added discussions and tables discussing this to the SI (Si: 3.4).

2. Address time range issues:
    a. When was the survey fielded and did it correspond to flu activity? Timeline of when the browser data was collected, when the survey was fielded, when field period was closed, etc.
        i. *As noted above, we fielded an initial survey wave from March 19th to March 27th 2015, then a second wave from April 27th to April 31. Browser data was collected between Nov. 1, 2014 and continued until March 31, 2015. The flu season typically peaks in February, but sometimes peaks in December, January, and March. During the 2014-2015 season, the flu peaked in December (https://www.cdc.gov/flu/pastseasons/1415season.htm). We added a sentence on this to the survey methods section of the main text and go into detail in the SI (SI: 3.2).*
3. Fixes to forecasting section:
    a. What was the volume, temporal scale, and geographic coverage of the Bing search queries from 2011 to 2016?
        i. *We are not permitted to share the total numbers of searches for any given query from Bing. Instead, we generated z-scores for three core quantities across all states and years: the total number of searches, the total number of searches related to flu (again drawing on the DOC2VEC expansion technique), and the total number of A1 searches (based on human labels). This is similar to 'Google Trends' data. For the state-level estimates, we first divided the number of queries by the state's population to create a per capita per year estimate. We then averaged this estimate by year to create an annual state estimate. Finally, we subtracted the overall annual state average and divided by the standard deviation to create a z-score.*
    b. Where did the covariate demographic data (used in the MRP model) come from? At the very least, there needs to be a more clearly described section about these data in the supplement.
        i. *We have added more information about this in the main document and in the SI (SI: 6.6). The variables education, age, and the number of children per house came from the American Community Survey (2014 and 2015, 5 year estimates). They were downloaded via the American Community Survey Application Programming Interface (API) using the ACS package in R (https://cran.r-project.org/web/packages/acs/acs.pdf). Education is a binned (by quantile) measure of the proportion of individuals in each zip code who had completed a post-high school degree program, divided by the population of the zip code (these came from the 2014 5-year ACS estimates). Age is a binned (by quantile) measure of the median age in a zip code (from 2014 5-year ACS estimates). Finally, the number of*

children per house is a binned quantile measure of the average number of children in each household within a zip code (also from the 2014 5-year ACS estimates).

   c. How were the forecasting models implemented (e.g., which software) and will code be shared?
      i. *The MRP models were implemented in R using the lme4 package (Bates et al. 2015). The forecasting models were implemented in R using the Forecast package by Rob Hyndman. Code will be made available as supplementary files.*

4. Fixes to MRP description:
   a. Can you please describe the response variable more clearly? I guess it is the proportion of search queries possessing an A1 search term in zipcode i?
      i. *That interpretation is correct for the MRP section - the proportion of search queries possessing an A1 search term in zipcode i. We edited the main text and supplementary file to make the response variable in the MRP section more clear. The window time component was not represented in the main text, following a similar representation as Wang et al. (2015), but we added it for additional clarity. Additionally, we have otherwise edited the equations for consistency in indexing and readability.*

   b. Perhaps I am missing something, but I do not understand how the first model equation is representing a smoothing and re-weighting process. The alphas appear to be regression coefficients, but is there also supposed to be some covariate data in the equation?
      i. *We have made edits to clarify this in the main document and in the supporting information. As written in the original submission, we did not show the formula for re-weighting process, but we have added this to the main text for full clarity. We have updated the model equations to make the terms clearer in addition to the equation for poststratification. The model is based on Multilevel Regression and Poststratification (so-called MRP) models described elsewhere (see e.g Wei, et al., 2014, Lax and Phillips, 2009). We emphasize in the text that the idea behind the smoothing is that zip codes with distinct combinations of demographic features will be rare, and thus suffer from variability due to small sample sizes. By creating priors for demographic variables, the multilevel model borrows some power from zip codes with similar characteristics. The poststratification is conducted by generating weights based on the true prevalence of zipcodes of type i, then weights the data accordingly.*

   c. The equation indexes (i,j,k,p,q) in the model equations need to be defined. Presumably some of them represent bins, but it is not immediately clear what those bins are
      i. *Yes, a majority of the variables are binned quantiles, with the exception of income which is a fixed coefficient (following advice from Buttice and Highton (2013), see more explanation below). We have added further*

elaboration to the model equations to more clearly depict what each index means and added explicit statement in the text about what is being binned and what is not.

d. The second-to-last paragraph describes the application of a time series model over a rolling three-day window, but the primary model equation does not appear to have any time-varying components.
    i. *We appreciate this point. We sought to follow the clearest possible representation of the model from the literature for consistency and readability. We have added subscripts (t) to indicate the time window and hopefully clarify the model for all readers. The t subscript now depicts the time window in which the model is being estimated.*

e. I'm a bit confused in general about the time-varying nature of the MRP model; the demographic factors would vary across zipcodes, they will not vary much over time, so I don't understand the benefit of having this factor be time-varying.
    i. *Correct, the demographic variables do not change over time, but the underlying flu rate within each zip code does change over time, and the MRP model smooths and reweights the average prevalence of flu-like searches to capture this change. We have tried to make this clearer in the present draft. Our paper follows a similar structure as Wang et al. (2015), in which they apply a MRP model over a moving window to survey data from a video game console. Similarly, a moving window allows us to generate a smoothed and reweighted average for the last day of the time window in every zip code or geographic area desired. We then use the smoothed signal from this method as input to our time series model.*

f. Are the zipcode level models being run independently or jointly? Was there examination of spatial dependence? How is the MRP signal being aggregated to the national level?
    i. *Models are estimated with a multilevel model (so-called 'partial pooling'), and therefore are not fully disaggregated by zipcode nor fully pooled. Each row in the query data represents one query at a period of time in a particular zipcode and state. We have varying densities of searches in different zip codes, which makes it very challenging to fully disaggregate by zip code. Instead, the multilevel model will treat zip codes with greater search densities as stronger signal than zip codes with lesser search densities, and borrow statistical power from demographically similar zip codes. Regarding spatial correlation, the state where the zipcode is located is the only spatial variable in the model. Spatial dependence will make the standard errors of the model inaccurate, but not the coefficient estimates. We do not use the MRP model for calculating statistical significance measures - the MRP technique is only used as a smoothing and re-weighting step of the flu signal. We tried to clarify these points in the paper.*

g. Why is the prior for Income missing, and why is it described by beta when all other effects are described by alpha?

        i. *Here, we follow the advice of Buttice and Highton (2013) and the example of other MRP papers to include one state-level covariate in our models. In their work, they show that including at least one state-level covariate increases the correlation and reduces absolute bias between the final signal and the underlying true value. We now include this point in the main text in the MRP methods section.*

    h. The indexes for the alpha equations do not match. For example, $\alpha_j ^ State$ is normally distributed for all h numbered 1 to 52? In any case, I would recommend the authors use different indexes so that there is less confusion with the p and q indexes in the SARIMA model.

        i. *We appreciate this point. There was a typo in the section of equations displaying distributional assumptions, which we have now fixed. The assumption we were trying to convey (and we think we now convey better!) was that the alphas for the states are normally distributed, which is an assumption we made for all the terms described by alphas. We now recognize that the distributional assumptions were not particularly helpful to a general audience so we moved them to the supporting information and corrected the mismatched alphas.*

    i. The authors should use consistent model names in the Forecasting Results, State Level Findings and Tables in the Results section. It would be good to introduce the model names when describing the model structures in the Methods as well.

        i. *We made edits to standardize mentions of the 'behavioral' model and the 'tracking' model throughout the main text. We also added text to the introduction of the methods section to introduce definitions for these models.*

5. I recommend that the authors provide more descriptive captions for the tables and figures, and add the long-form terminology for metrics that are abbreviated.

    a. *We have updated the captions on the tables in the main text, and edited text to avoid short-form abbreviations where it could be confusing.*

6. Forecasting model issues:

    a. While it's important to show the model performance, I think the Results should report more results in the context of ILI rates. I think this will make the paper more relevant to an epidemiological audience.

        i. *We have edited to deliver a more substantive interpretation of the results in the context of ILI in the main text. Specifically, we now discuss the raw correlation between the MRP input signal and the flu rates (~90). We also describe in terms of the overall rate of flu, the percentage away we were from a perfect prediction.*

    b. Are the models prone to under- or over-estimation at different times of the flu season? Are the models capable of capturing the peak timing and magnitude of seasonal outbreaks? The authors should comment on the utility of these models in capturing peak timing and magnitude, not just with regards to model error.

i. *For the national model, we found that models based on history were more likely to overestimate the peaks relative to the MRP-based model. We comment on this in the forecasting results section. Based on this and other comments throughout, we include more substantive interpretation of these results in the main results section.*

c. I don't quite understand what is plotted in Figure 2B since the text describes these as 2-step ahead predictions. Which date does the x-axis represent? Were the plotted predictions were made two timesteps prior?

i. *The x-axis corresponds to the current date. Correct, the prediction lines at a given date correspond to the two-step ahead prediction for that day. We have clarified.*

d. Also, is there some reason why some of the models were excluded from the prediction figure? Regardless, it's curious that the model predictions seem to lag behind the actual ILI signal.

i. *Delaware data is more sparse compared to the other states, with low counts and as such susceptible to more erratic and unreliable predictions (for all methods, not just ours), so we initially omitted it from the figure, even though the MRP model still displays lower prediction error compared to the historical models. We now provide all the absolute error plots in the Supporting Information. Regarding lagging of the prediction signal from the ILI signal, it is somewhat hard to dissect the trajectory of the predictions, in that undershooting at one time point will propagate to the future. The model also incorporates a good amount of information from prior flu levels (the underlying history, which is extremely powerful), so the MRP model prediction will partially reflect the prior values in its prediction.*

7. Limitations: Is it possible that multiple users are creating logged searches in the browser during the survey? Should the search term data represent more of a "household" measure instead of an individual measure?

a. *This is a good point, and indeed the household measure is plausible. To check for this, when we restrict the data to cases where the respondent indicates that she is the primary user of the machine where the browsing tracker is installed, the results are similar. We include a discussion of the challenges of associating searches from a particular machine with a single user.*

8. Minor comments:

a. Please add more references with greater specificity to sections in the supplement.

i. *Done, thank you for noting it.*

b. Methods, Survey Data, third paragraph: The values reported do not sum to 654 survey respondents.

i. *The values in this section report omit 10 respondents who did not have any search volumes. We added a note to the document to clarify.*

c. Many sections of the supplement appear to be duplicates of the main text. This should be cleaned up and made less redundant.

i. *Thank you, we have removed these redundancies.*

d. I don't think the supplement needs to include a section for the Discussion. All of the discussion points should be included in the main text.
    i. *Noted, we have removed the discussion from the supplement and placed all critical points in the main text.*

Responses to Reviewer 1

REVIEWER 1 COMMENTS TO ADDRESS

9. The ILI prediction model is not necessary.
    a. First, from a philosophical point of view, this reviewer thinks there are a surfeit of ILI prediction algorithms – why do we need another?
        i. *Good point. We agree a new ILI model is not necessary, and we use an off the shelf model, and just add the signal derived. So, the scientific point here is not to produce the best possible ILI predictions-- that's not our value added-- it's to show if our signal helps in a standard model. We have adjusted our language to make this point clear. The issues with previous models have to do exactly with what we emphasize here - a different way of integrating search query data. We saw the need to establish the overall validity of using search queries as a proxy for flu-like experience. This is important for understanding how powerful search queries can be for predicting flu in the population. However, the counterargument from any potential reader would be say that our method may be valid in labeling queries that have a higher prior likelihood of being associated with the flu, but it may not be able to actually predict the flu. So, we sought to test whether this method would actually work to predict the flu.*
    b. Second, the presented model does not outperform previous models – again, what does this add?
        i. *On one hand, this paper is good news for papers that have used search queries to predict the flu - searches are associated with self-reported flu-like symptoms, thus proving the validity of the first assumption of existing methods. On the other hand, we think we have demonstrated a proof of concept of a survey-based method for forecasting the flu - one that does not require comparison against every search query in a particular engine - a method that prior work has been shown to be challenging for a variety of reasons. In general, we think the method of fitting such a large number of queries to a handful of data points needs revision. We think that such techniques are more likely than ours to fail without warning.*
    c. Third, the claims that "The overshooting becomes more apparent for the history signal. MRP serves as to check this." Is not correct – it seems random when the history is above MRP and vise-versa. I think including this weakens the paper.
        i. Thanks. We removed the sentence on overshooting. We know that incorporating the MRP signal in the SARIMA-X model reduces the

variance in the prediction error, so we emphasize that in the paper. The inclusion of the MRP as an exogenous signal has a considerable beneficial impact on the prediction quality.

10. I count the phrase “influenza-like-illness (ILI)” 7(!) times in the text. Please be consistent.
    a. *Many thanks for noting this issue - we have corrected it throughout.*
11. The main text deals with A1 & A2 searches, while the supplement has A1, A2, B1, B2, C1, and D. Why are these other results not in the main text? These are the most interesting aspects of the paper.
    a. *We agree that the degree to which users search for illness information on the web is very interesting. We added a section to the main paper and another to the Supporting Information describing how often users reported searching for information online versus seeking information in-person from a healthcare provider (see SI:4). Based on input from the editor, we felt that adding each of the B1, B2, etc., category results to the main paper may distract from the main results of the paper. We kept them in the SI, however.*
12. There needs to be more information on the survey in the ms. Where/who/how were the 20,000 selected? Who collected those data? How nationally representative is the study group? There isn’t a citation for the data in the paper. Additionally: will these data be available to other researchers?
    a. *We report in section 3.1 of SI how the original panel is conducted. While we didn’t have access to report the demographics of the full 20,000 panel, we do report the demographic characteristics of those we invited and those who responded to our survey (see SI 3.1 for these results).*
13. The paper needs confidence intervals throughout. What’s the magnitude of uncertainty for a relative risk comparing flu in house v. not of 1.57?
    a. *Many thanks for suggesting this. We have added these for our case-control model throughout. We used the ‘zelig’ package in R to generate simulation-based 95% confidence intervals for all the main quantities we report.*
14. Why is DE omitted from Figure 3?
    a. *Fixed. Based on the feedback here, we moved figure 3 to the Supporting Information.*