Decision on manuscript NCOMMS-19-14890.

Bernadett.Gaal@nature.com

Tue 7/30/2019 6:49 AM

To: Stefan Wojcik <swojcik@pewresearch.org>;

** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors **

Dear Dr Wojcik,

Your manuscript entitled "Survey Data and Human Computation for Improved Flu Tracking" has now been seen by 2 referees, whose comments are appended below. First of all, please accept our apologies for the delay in the review process! You will see from their comments copied below that while they find your work of considerable potential interest, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version that addresses these concerns.

We hope you will find the referees' comments useful as you decide how to proceed. Should further analysis allow you to address these criticisms and to further develop your manuscript along the lines recommended by Reviewer 2, we would be happy to look at a substantially revised manuscript. However, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. If the revision process takes significantly longer than six months, we will be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Nature Communications or published elsewhere in the meantime.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if you wish to discuss the revision or if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When resubmitting your paper, please highlight all changes in the manuscript text file. We also ask that you ensure that your manuscript complies with our editorial policies. Specifically, please ensure that the following requirements are met, and any relevant checklists are completed or updated and uploaded as a Related Manuscript file type with the revised article:

IMPORTANT
To improve the quality of methods and statistics reporting in our papers, we are now asking all authors to complete an editorial policy checklist that verifies compliance with all required editorial policies. Please ensure that the checklist is completed and uploaded with your revised article. All points on the policy checklist must be addressed; if needed, please revise your manuscript in response to these points. Please note that this form is a dynamic ‘smart pdf’ and must therefore be downloaded and completed in Adobe Reader, instead of opening it in a web browser.
Editorial policy checklist:
https://www.nature.com/documents/nr-editorial-policy-checklist.pdf

Reporting summary:
https://www.nature.com/documents/nr-reporting-summary.pdf

Please also fill out the code and software submission checklist. Please note that this form is a dynamic ‘smart pdf’ and must therefore be downloaded and completed in Adobe Reader, instead of opening it in a web browser.
https://www.nature.com/documents/nr-software-policy.pdf

Data availability statements and data citations policy: All Nature Communications manuscripts must include a section titled "Data Availability" as a separate section after the Methods section but before the References. For more information on this policy, and a list of examples, please see
https://www.nature.com/documents/nr-data-availability-statements-data-citations.pdf
In particular, the Data availability statement should include:

- Accession codes for deposited data
- Other unique identifiers (such as DOIs and hyperlinks for any other datasets)
- At a minimum, a statement confirming that all relevant data are available from the authors
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.
- If a source data file is provided, please add a reference to this in the data availability statement. For example:
- "The source data underlying Figs 1a, 2a–d, 6d, h and 7c and Supplementary Figs 1a and 5d are provided as a Source Data file."

DATA SOURCES: We strongly encourage authors to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided here: http://www.nature.com/sdata/policies/repositories

If a community resource is unavailable, data can be submitted to generalist repositories such as figshare (https://figshare.com/) or Dryad Digital Repository (http://datadryad.org/). Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publicly available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be included in the reference list and cited where relevant.

Please refer to our data policies here: http://www.nature.com/authors/policies/availability.html

Springer Nature encourages all authors and reviewers to adopt an Open Researcher and Contributor Identifier (ORCID). ORCID is a community-based initiative that provides an open, non-proprietary and transparent registry of unique identifiers to help disambiguate research contributions. All authors who link their ORCID to their account in our submission system will have their ORCID published on their articles, if the article is accepted for publication. Please note that this is only possible if ORCIDs are linked prior to acceptance, that is, it is not possible to add ORCIDs at proof.

Please ensure that all co-authors are aware that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

To add an ORCID please follow these instructions:

1. From the home page of the MTS click on 'Modify my Springer Nature account' under 'General tasks'.
2. In the 'Personal profile' tab, click on 'ORCID Create/link an Open Researcher Contributor ID (ORCID)'. This will re-direct you to the ORCID website.

3a. If you already have an ORCID account, enter your ORCID email and password and click on 'Authorize' to link your ORCID with your account on the MTS.

3b. If you don't yet have an ORCID account, you can easily create one by providing the required information and then clicking on 'Authorize'. This will link your newly created ORCID with your account on the MTS.

If you experience problems in linking your ORCID, please contact Platform Support

Please use the following link to submit your revised manuscript, point-by-point response to the referees' comments (which should be in a separate document to any cover letter) and any completed checklist:
https://mts-ncomms.nature.com/cgi-bin/main.plex?el=A3S3BZOv4A5JbeE2I2A9ftdztHXqYM497FwkTA6ru6ozAZ
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further. Thank you for the opportunity to review your work.

Best regards,

Bernadett

Bernadett Gaál, DPhil
Senior Editor, Nature Communications
Nature Research

@GaalBernadett
@naturecomms
ORCID: orcid.org/0000-0003-1926-7648
www.nature.com/ncomms

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

This is the second time I have reviewed the study by Wojcik et al and I see that while some of my previous concerns have been address, several of my main points have not. I will reiterate them here and still maintain their importance. In general I think the survey results are very interesting, novel, and an advance for the field and that the ILI prediction model detracts from the manuscript.

Major comments:
- The ILI prediction model is not necessary. First, from a philosophical point of view, this reviewer thinks there are a surfeit of ILI prediction algorithms – why do we need another? Second, the presented model does not outperform previous models – again, what does this add? 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.
- The language is better, but still needs work. I count the phrase "influenza-like-illness (ILI)" 7(!) times in the text. Please be consistent.
- 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.
- 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?
- 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?
- Why is DE omitted from Figure 3?

Reviewer #2 (Remarks to the Author):

The goal of this study was to examine associations between symptoms and internet searches of influenza-like illness and to build a behavior-adjusted model to forecast ILI incidence with internet search terms. While other studies have examined population-level associations between ILI surveillance and internet search activity, this study provides novelty and interest in its validation of the relationship between disease and ILI-related search activity on an individual level and its attempt to adjust for non-representativeness in using internet search activity in a surveillance context.

While the study has collected valuable data in linking the use of digital data streams in surveillance and made progress in increasing the specificity of ILI-related searches, I think the authors could be more ambitious in their analyses and more applied in the framing and reporting of results in order to increase the value of their work in an epidemiological context. Alternatively, the authors may choose to target their work for a more statistical and machine-learning audience, in which case, the methods and comparisons of models should be strengthened and expanded upon. In the current manuscript, there seem to be critical descriptions of modeling methods that are missing, incorrect, or in need of clarification.

Primary comments:

- Considering the survey include a question on seeking care with a healthcare provider, I am surprised that the authors did not do any comparisons of this data and search behavior. Even outside of the development of forecasting models, one existing narrative about the advantage of digital epidemiology is its potential to identify disease outbreaks in “real-time” and among a greater population, not just among those individuals that are captured in traditional surveillance systems several days later. It would be useful to gain insight into whether internet search behavior for ILI might precede or supplant visits to a healthcare provider, with the goal of describing the populations that might be captured in digital versus traditional surveillance.

- I am concerned about the period of symptom recall proffered in the survey question. Particularly when the ILI symptoms are mild, three months is a long recall period for an individual, much less for an individual to recall about other household members. There is literature to suggest that accurate recall is on the order of two days for an experience of diarrheal illness and up to two months if a hospital visit was required. I think it is particularly problematic that the directionality of recall bias could tend in both directions; on one hand, individuals may forget they had the symptoms since they are relatively common; alternatively, individuals may reason that there is a high probability they have experienced these symptoms in the past three months since they are relatively common. I suggest that the authors consider the implications for recall bias more seriously in their paper and identify ways in which this limitation may affect their results.

- What was the specific time range for the survey and did it correspond with active influenza activity? (This should be included in the main text.) With regards to my earlier concern about recall bias, are there differences between respondents that completed the survey closer to the start or peak of the influenza season?

- I think the Forecasting Methods section needs to include substantially more detail.
  - What was the volume, temporal scale, and geographic coverage of the Bing search queries from 2011 to 2016?
  - 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.
  - How were the forecasting models implemented (e.g., which software) and will code be shared?

- The MRP model described on page 4 needs to be further explained:
  - 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?
  - 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?
  - 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.
  - 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’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.
  - 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?
  - Why is the prior for Income missing, and why is it described by beta when all other effects are described by alpha?
  - 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.

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

- 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. 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?

- 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? Weret the plotted predictions were made two time-steps prior? 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. As mentioned before, the authors should comment on the utility of these models in capturing peak timing and magnitude, not just with regards to model error.

- 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?

Minor comments:

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

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

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

- 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.

** See Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits

This email has been sent through the Springer Nature Tracking System NY-610A-NPG&MTS

Confidentiality Statement:

This e-mail is confidential and subject to copyright. Any unauthorised use or disclosure of its contents is prohibited. If you have received this email in error please notify our Manuscript Tracking System Helpdesk team at http://platformsupport.nature.com .

Details of the confidentiality and pre-publicity policy may be found here http://www.nature.com/authors/policies/confidentiality.html

Privacy Policy | Update Profile

DISCLAIMER: This e-mail is confidential and should not be used by anyone who is not the original intended recipient. If you have received this e-mail in error please inform the sender and delete it from your mailbox or any other storage mechanism. Springer Nature Limited does not accept liability for any statements made which are clearly the sender's own and not expressly made on behalf of Springer Nature Ltd or one of their agents.
Please note that Springer Nature Limited and their agents and affiliates do not accept any responsibility for viruses or malware that may be contained in this e-mail or its attachments and it is your responsibility to scan the e-mail and attachments (if any).