Public Health England

Protecting and improving the nation's health

Daily Google Search Queries – Added Value to National Influenza Surveillance?

Katie Owens[1], Richard Pebody[1], Vasileios Lampos[2], Ingemar Cox[2], Simon de Lusignan[3,4], Alex Elliot[1]

1. Public Health England (PHE), 2. University College London (UCL), 3. Royal College of General Practitioners, 4. University of Surrey

INTRODUCTION

Surveillance systems monitoring influenza-like illness (ILI) within Public Health England (PHE) are predominantly run on data which is routinely provided through established healthcare systems. ILI surveillance systems produce regular outputs that are either daily or weekly depending on the system. The timeliness in which some of these systems report can range from a few days to two weeks.

Online user generated content, in the form of Google search queries has the potential to provide near real-time estimates and expand the community surveillance of ILI by potentially picking up cases that do not seek medical attention.

To complement existing systems by potentially improving the timeliness of ILI surveillance, the use of online user generated content, in the form of ILI related Google search queries, was assessed via daily estimates.

METHODS

A non-linear Gaussian Process model was developed by Lampos *et al.* using a combination of natural language processing and machine learning techniques[1,2]. Ground truth data, in the form of historical Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) weekly ILI GP consultations, was used to train the supervised model. Near real-time daily estimates were produced by the supervised model based on the frequency of ILI related search queries within a 10%-15% sample of all queries issued, extracted daily from Google's Health Trends Application Programming Interface (API).

Weekly and daily lags of the supervised model estimates were assessed against a number of PHE established surveillance systems for metrics such as the week number in which ILI started to increase and the date and/or week number in which maximum peak in ILI was detected.

To observe underlying trend, a three day moving average was applied to the estimates produced by the supervised model and a seven day moving average was performed for the Real-time syndromic surveillance systems GP in hours (GPIH), GP out of hours (GPOOH) and NHS 111 adjusting for weekend and bank holidays[3].

RESULTS

The underlying trend of daily ILI estimates from the supervised model for weeks 40 to 15 of the 2017/18 season was comparable to those seen within established surveillance systems. An increase in daily estimated rates was observed during week 46, before peaking at the end of week 2 (Figure 1).

The peak in estimates was observed one week earlier than that of RCGP data (week 3). Weekly averages of the supervised model estimates provided a very strong positive Pearson correlation to the weekly RCGP ILI consultation rates at *r* = 0.97 (*p* = <0.001).

RESULTS (Continued)

Comparative analysis of the supervised model, showed that the date the estimates peaked was similar to the three syndromic surveillance systems (Figures 3-5). GPIH, GPOOH and NHS 111 all peaked within a few days of each other during weeks two and three when observing the raw data (un-adjusted). The supervised model peaked at the same time as the proportion of cold and flu calls observed through NHS 111, which was one day after GPOOH, all of which were over the weekend period. A peak in GPIH ILI rates was observed the following day.

Pearson correlation of the supervised model to the syndromic surveillance systems was again positive with strong correlation to GPOOH and NHS 111, equal to *r* = 0.92 (*p* = <0.001) and 0.97 (*p* = <0.001) respectively. The supervised model had weaker correlation to GPIH, equal to *r* = 0.66 (*p* = <0.001).

When adjusting for weekends and bank holidays the peak in ILI shifted to week two for GPIH, GPOOH and NHS 111 with all systems ILI peaking with a range of one to four days before.

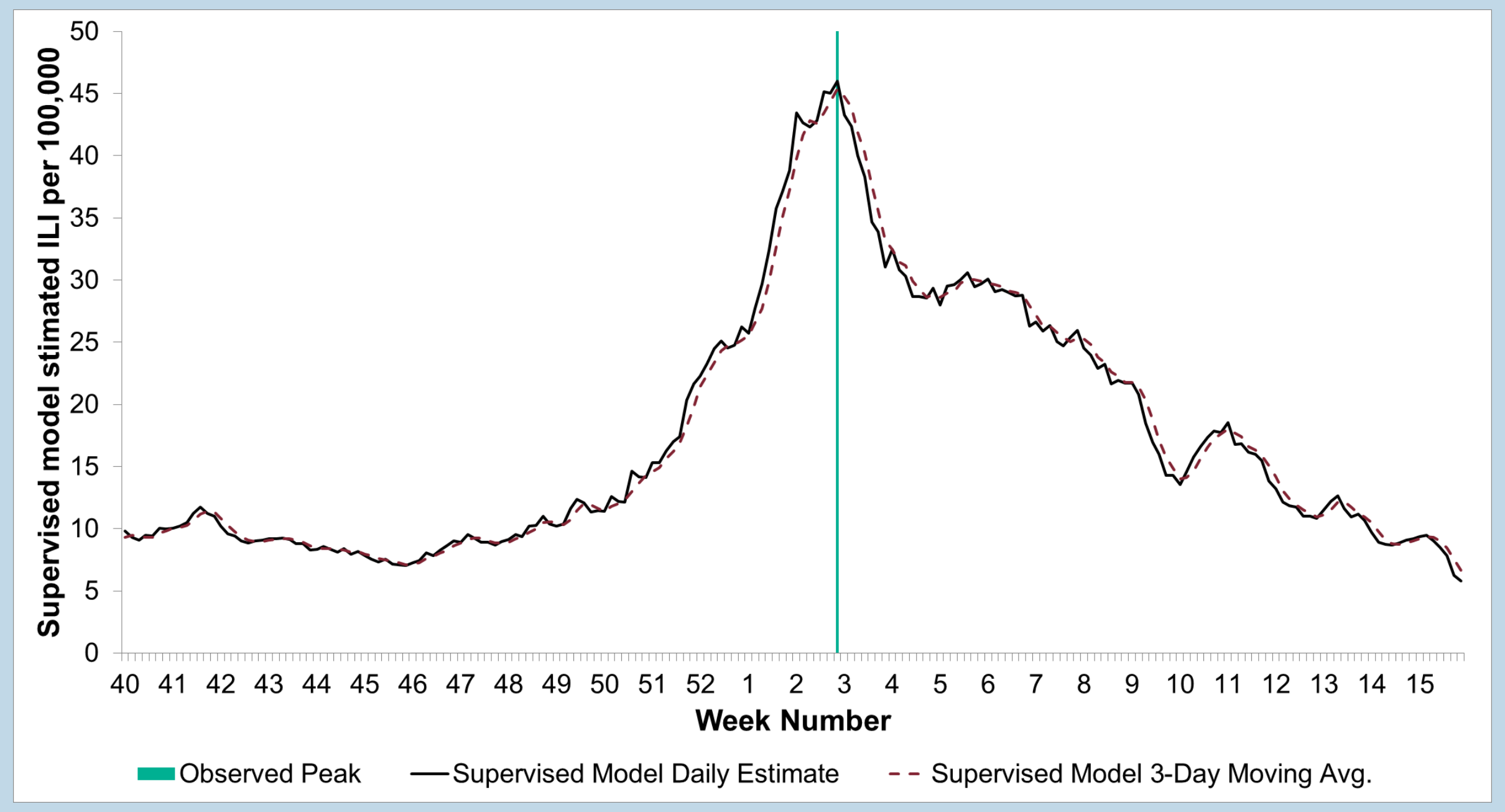

Figure 1. Supervised model daily ILI estimates per 100,000 population and 3-day moving average indicating the maximum estimate of ILI: Weeks 40 to 15, 2017-18 season.

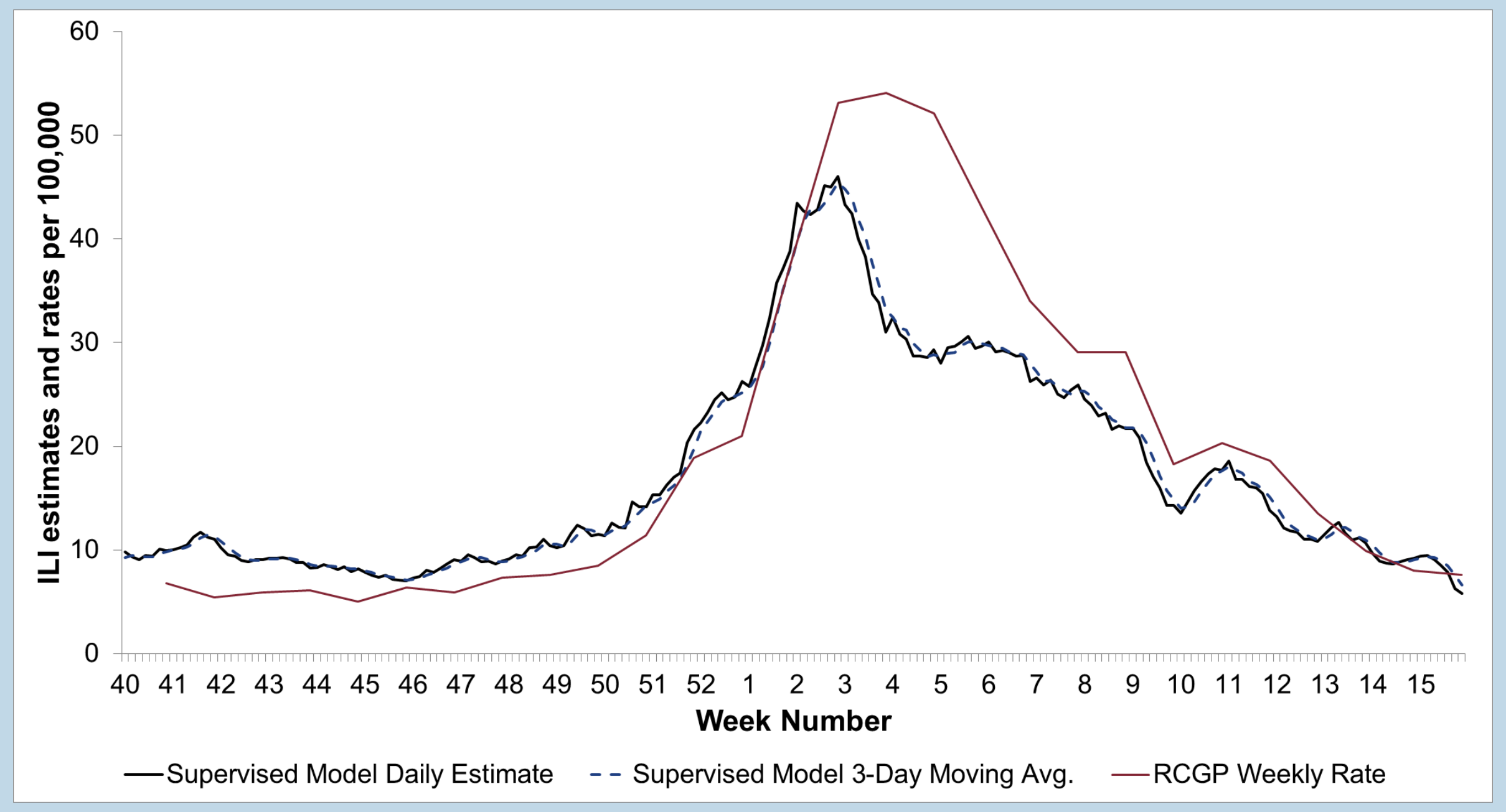

Figure 2. Supervised model daily ILI estimates (and 3-day moving average) and RCGP ILI consultation rate per 100,000 population: Weeks 40 to 15, 2017-18 season.

*The RCGP weekly ILI rate is plotted on the Sunday of each week

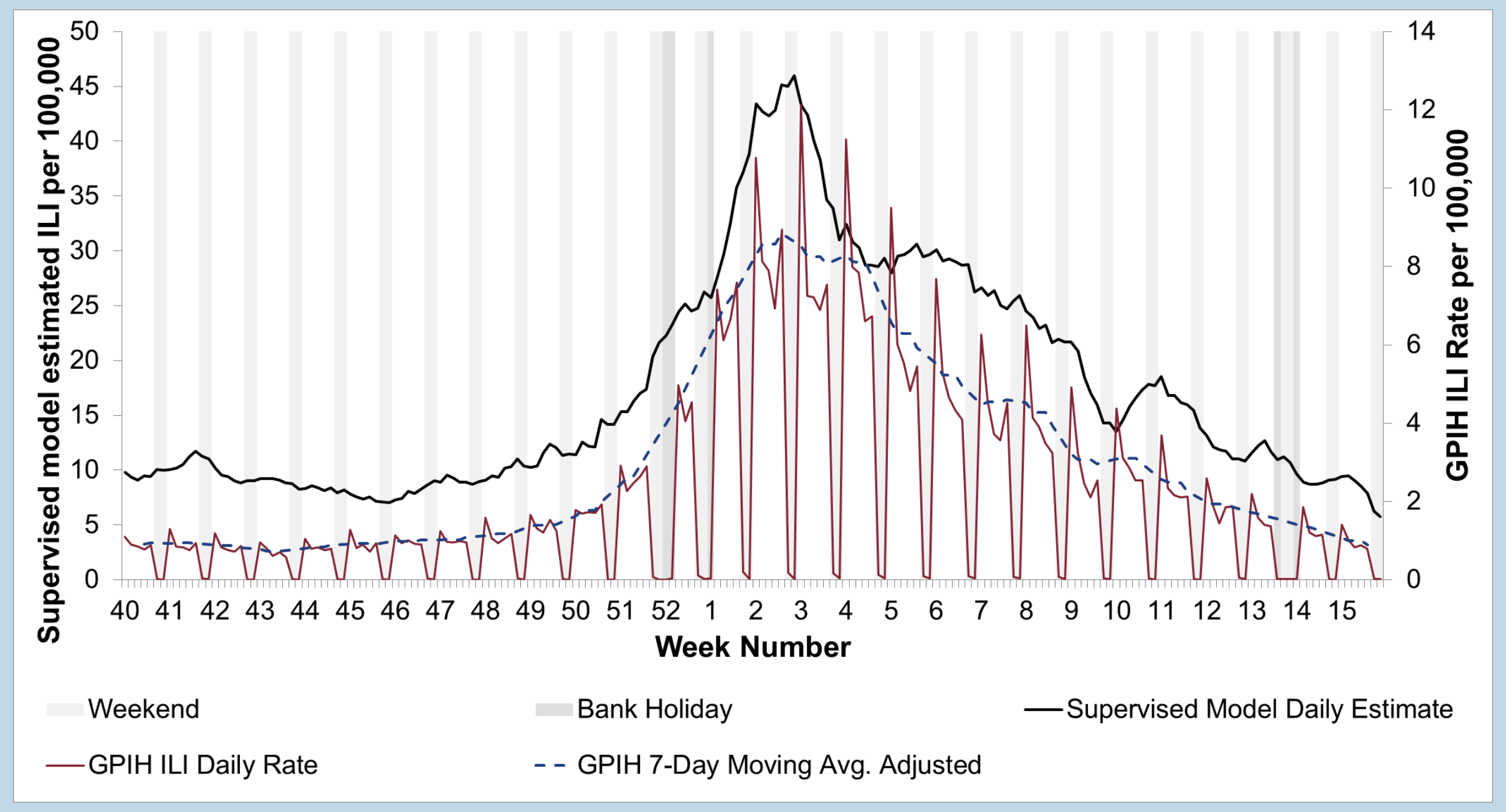

Figure 3. Supervised model daily ILI estimates and GP in hour (IH) daily ILI consultation rate (and 7-day adjusted moving average) per 100,000 population: Weeks 40 to 15, 2017-18 season.

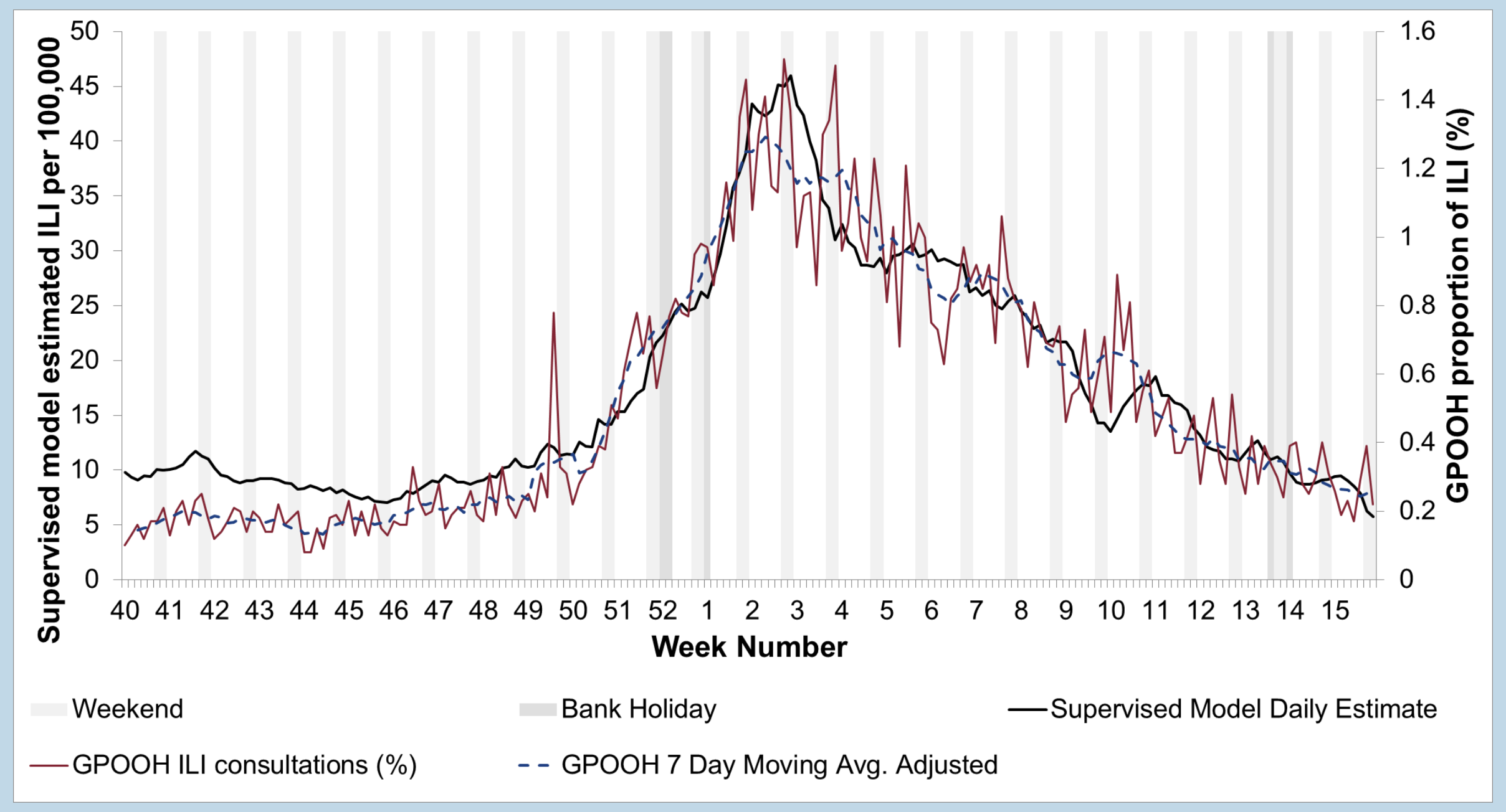

Figure 4. Supervised model daily ILI estimates per 100,000 population and proportion of GP out of hours (OOH) 111 ILI calls (and 7-day adjusted moving average): Weeks 40 to 15, 2017-18 season.

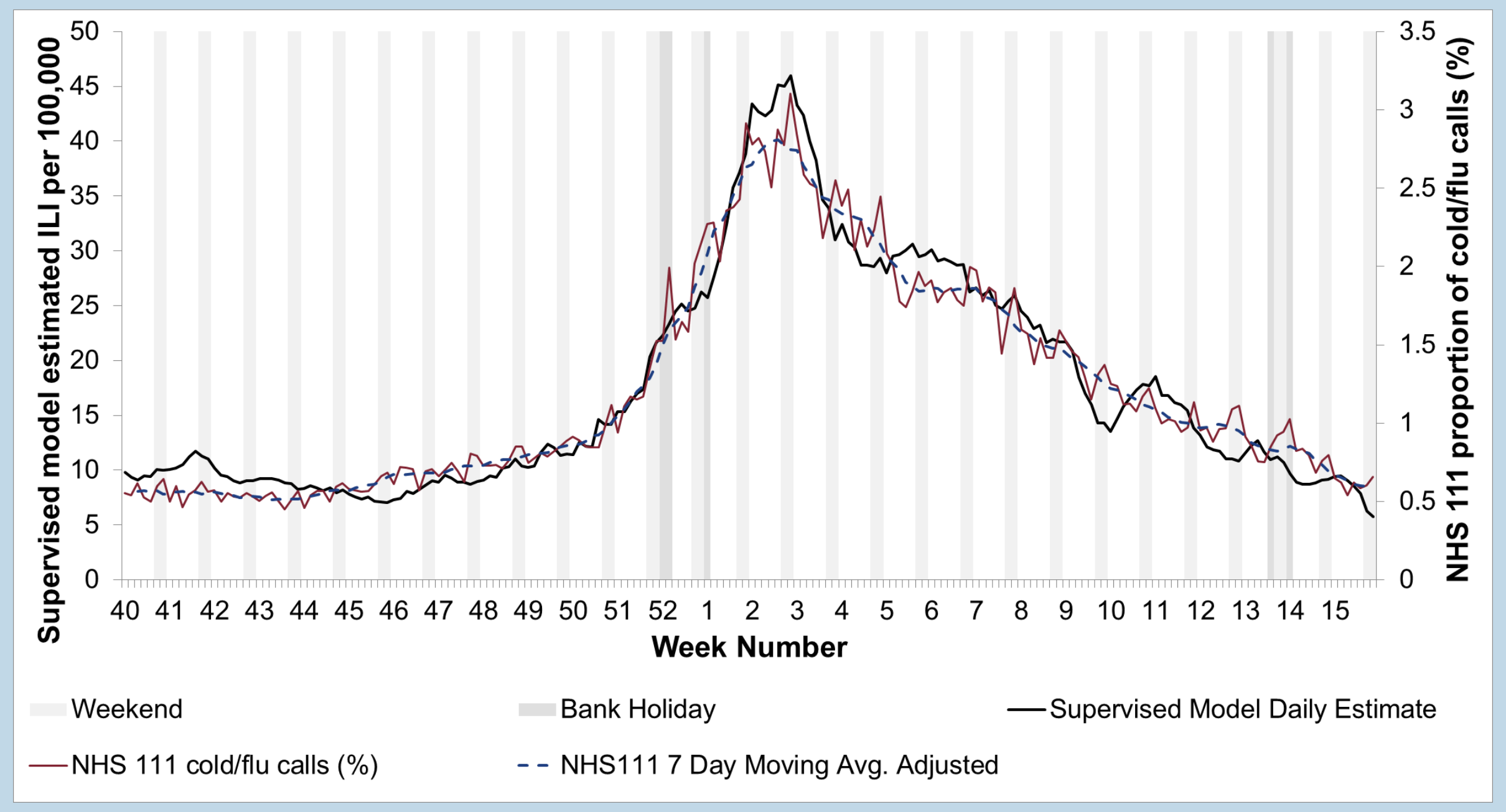

Figure 5. Supervised model daily ILI estimates per 100,000 population and proportion of NHS 111 cold/flu calls (and 7-day adjusted moving average): Weeks 40 to 15, 2017-18 season.

DISCUSSION

Daily estimates produced by the supervised model followed a similar trend to not only the surveillance systems performed in this comparative analysis, but also to that of some of PHE's other ILI surveillance systems[4]. Pearson correlation on the whole was strong between the supervised model and each of the surveillance systems used within the analysis, the exception of which was the GPIH system, however this weaker correlation could be due to the drops in rates resulting from weekend closures, when observing the raw data.

Throughout the season, daily estimates produced by the supervised model provided more timely updates at a national level compared to traditional weekly sentinel surveillance such as RCGP in which it peaked one week earlier. Similar timeliness of the estimates peaking was observed when comparing to the daily ILI syndromic surveillance data.

Unlike GPIH, GPOOH and NHS 111 the supervised model is not affected by weekday and public holiday effects[3]. The frequency of the ILI related search queries remain more stable day to day and are not influenced by what day of the week it is. Moving averages to adjust for the impact of weekends and bank holidays are therefore not a requirement to analyse the trend of ILI.

CONCLUSIONS

- User online generated content has the potential to expand the catchment of ILI within the community by including those that do not seek health care but that go online to receive advice.
- Using the supervised model as a surveillance method of ILI would be complementary to the systems already established by providing an additional estimation of ILI cases in the general population in near real-time.

ACKNOWLEDGEMENTS

This research was funded by EPSRC IRC in Early-Warning Sensing Systems for Infectious Diseases (i-sense) EP/K031953/1.

We acknowledge support from: RCGP RSC, PHE's Real-Time Syndromic Surveillance Team (ReSST) and Google, NHS 111 and NHS Digital; OOH providers and Advanced; TPP and participating SystmOne practices and University of Nottingham, ClinRisk, EMIS and EMIS practices submitting data to the QSurveillance database.

i-sense
EPSRC IRC in Early Warning Sensing Systems for Infectious Diseases

REFERENCES

1. Lampos V et al. Enhancing feature selection using word embeddings: the case of flu surveillance. Proceedings of the 26th International Conference on the World Wide Web, 2017. doi: 10.1145/3038912.3052622
2. Lampos V et al. Advances in nowcasting influenza-like illness rates using search query logs. Scientific Reports. 2015 3;5. doi:10.1038/srep12760
3. Buckingham-Jeffery E et.al. Correcting for day of the week and public holiday effects: improving a national daily syndromic surveillance service for detecting public health threats. BMC public health. 2017; 17:477. doi:10.1186/s12889-017-4372-y
4. Public Health England. Surveillance of influenza and other respiratory viruses in the UK: Winter 2017 to 2018.