Elucidating user behaviours in a digital health surveillance system to correct prevalence estimates

Dennis Liu[1,2*], Lewis Mitchell[1,2], Robert C. Cope[1],
Sandra J. Carlson[3], Joshua V. Ross[1,2]

[1]School of Mathematical Sciences, The University of Adelaide,
North Terrace, Adelaide, SA 5015, AUS
[2]ARC Centre of Excellence for Mathematical and Statistical Frontiers, AUS
[3]Hunter New England Population Health, Wallsend, NSW 2287, AUS

[*]To whom correspondence should be addressed;
E-mail: dennis.liu01@adelaide.edu.au

Abstract

Estimating seasonal influenza prevalence is of undeniable public health importance, but remains challenging with traditional datasets due to cost and timeliness. Digital epidemiology has the potential to address this challenge, but can introduce sampling biases that are distinct to traditional systems. In online, participatory health surveillance systems, the voluntary nature of the data generating process must be considered to address potential biases in estimates. Here we examine user behaviours in one such platform, FluTracking, from 2011 to 2017. We build a Bayesian model to estimate probabilities of an individual reporting in each week, given their past reporting behaviour, and to infer the weekly prevalence of influenza-like-illness (ILI) in Australia. We show that a model that corrects for user behaviour can substantially effect ILI estimates. The model examined here elucidates several factors, such as prior reporting of ILI and consistency of prior reporting, that are strongly associated with the likelihood of participating in online health surveillance systems. This framework could be applied to other digital participatory health systems where participation is inconsistent and sampling bias may be of concern.

Influenza is a substantial public health concern. There are many challenges in estimating influenza activity and forecasting the spread of influenza, given that disease transmission in the general population is largely unobserved. Traditional influenza surveillance relies on public health system monitoring, such as through hospital admissions, notifications of laboratory confirmed cases, or voluntary reporting from local physicians [2, 11]. However, in these systems, estimates of disease prevalence can be limited by the time taken to collect, collate and publish through public health systems.

Online participatory health surveillance systems attempt to address these challenges by providing a convenient, simple and near real-time platform for self-reporting of symptoms. FluTracking [13] is one such system for monitoring influenza-like-illness (ILI), with a principal aim to contribute to community level ILI surveillance in Australia and New Zealand. Similar platforms exist in the US (Flu Near You [10]) and Europe (Influenzanet [22]).

These platforms often publish estimates of the incidence or prevalence of ILI in the population, usually derived as a proportion of the total number of reports received that week. These estimates have been found to correlate well with clinical surveillance by public health bodies [4, 21, 24], including across different definitions of ILI cases [22].

While systems such as FluTracking show promise in estimating the incidence of ILI in the population, there is evidence that these systems can be effected by user participation [4, 12, 23]. Very little is known about how these biases could affect disease prevalence estimates and there is a clear lack of studies that attempt to quantitatively adjust for them. Attempts to correct for these biases have all been based on the removal of data, such as only considering users who frequently participate [7, 19, 21, 22], removing the first report of a user, or by creating noise filtering algorithms that minimise sudden departures from sentinel data due to changes in participation [25]. While not unreasonable, these methods do not give insight into user behaviour, and do not examine the effect of their corrections on the estimates they examine. Some examinations of the heterogeneity of users and their behaviours have been conducted, such as inferring significant predictors for and classifying users on their participation levels [3, 5], and examining the demographic representativeness of the participating population [5, 7, 16]. However, user behaviour has not been analysed within a systematic, statistical modelling framework, in particular one which can simultaneously correct disease prevalence estimates. This work addresses this gap by using information about how individuals behave with regard to survey completion to inform our estimate of the weekly prevalence of ILI.

FluTracking Data

Participants can register in FluTracking at any stage of the year, and upon registration, they provide various demographic details about themselves, including age, gender, and postcode. Participants can optionally register and report on behalf of others in their household. Those participants who are submitting reports, on behalf of themselves and/or their household, will henceforth be referred to as 'masters', while any individual observed, master or not, will continue to be referred to generally as 'participants'. After registering, masters are sent an email on the Monday of each week during the influenza season to respond to an online questionnaire about the presence of fever or cough that they or their household members may have experienced in

the week prior, and whether they have received the influenza vaccine this year.

Surveys can be submitted for up to 5 weeks from their first reminder. Note that FluTracking often publishes estimates of ILI incidence to examine the spread of new cases of ILI [8]. In this study, we will focus on estimates of prevalence in order to determine if user behaviour is influenced by ILI status.

If a participant reports both a fever and a cough, follow up questions are revealed, such as enquiring about any health-seeking behaviour taken or if a sore throat was experienced. In this study, we define an ILI case as a survey response with a fever and a cough, which closely resembles the World Health Organization surveillance case definition [26].

Note that a report, and therefore an observation of ILI status in a household, is generated by the action of the master, and so it is the report of the master and subsequent reports on behalf of the rest of the household that is the outcome of interest. Also note that in order to build a framework for near real-time prediction, we have chosen to define a report submitted less than 7 days after the initial request as an 'on-time' report and not a 'late' report. This is the interval of time before the subsequent request for the next survey is sent. However, if symptoms are submitted in a late report, this information is not excluded, but used in the derivation of predictor variables for subsequent survey weeks, irrespective of submission date.

FluTracking operates between May and October every year in the winter season in the Southern Hemisphere. We examine all Australian reports submitted between 2011 and 2017, totalling 3,459,339 unique reports from 30,564 households and 52,773 participants. Of these reports, 352,287 (approx. 10%) are late reports.

Using the registration date of individuals, we can infer the weeks in which household reports were missed and never submitted as survey weeks without a report after their registration date for each season. This includes another 496,175 missed surveys.

The Model

We develop a framework to adjust ILI prevalence estimates by correcting for user behaviour, and construct a model to predict the probability of a user reporting in a given week, based on their prior behaviour and demography (Figure 1). An individual household i in a given week w will receive a survey request to participate, and will then proceed to either report and provide information on their symptoms, or not report. Given that they report, and their symptoms are therefore known, they will either report on-time, or report late. We will compare the following two estimates of the prevalence of ILI:

- the naïve estimate (an extension of the conventional estimate); and,
- a behaviour corrected estimate, (our new framework).

The conventional estimate in the literature is given by the total number of individuals that have reported with ILI, divided by the number of on-time reports in the given week. Our naïve estimate extends this by considering a Bayesian estimate, which respects the same mode value as the conventional estimate (see Materials and Methods). The behaviour corrected estimate is inferred from a framework that incorporates an observation process in the model, whereas

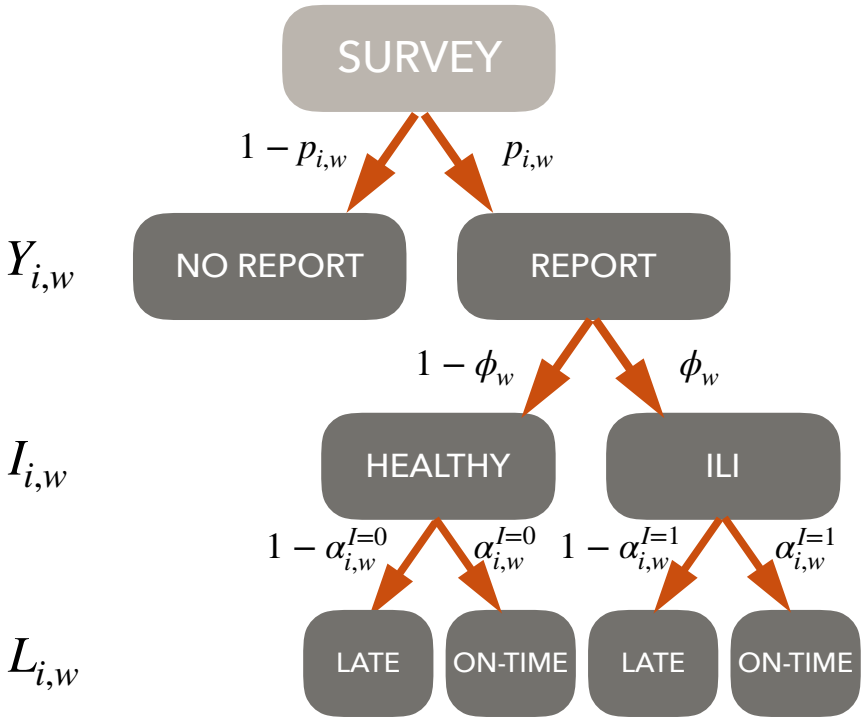

Figure 1: Visualisation of model of user outcomes for an individual i in week w, who has a probability $p_{i,w}$ of reporting and a probability π_I of having ILI. Given that a user reports, they then have a probability of $\alpha_{i,w}^{I}$ of reporting on-time. The hierarchical nature of the model is representative of how the data is thought to be generated.

the naïve model does not. Both estimates assume the number of individuals with ILI in the population is binomially distributed.

Each of the two estimates will be split into vaccinated and unvaccinated estimates of prevalence.

Materials and Methods

Naïve Model

The conventional estimate takes the total number of individuals that have reported with ILI this week X_w and divides this by the number of on-time reports this week $\mathcal{N}_w$. If we consider that each individual in the population has an equal probability $\hat{\pi}_w$ of having ILI in week w, then X_w can be modelled with a binomial likelihood with probability $\hat{\pi}_w$ and trials $\mathcal{N}_w$. Note that this interpretation does not consider any uncertainty in $\mathcal{N}_w$. Given a uniform prior Beta$(1,1)$ and a binomial likelihood, the posterior distribution of $\hat{\pi}_w$ is then Beta$(X_w+1, \mathcal{N}_w - X_w + 1)$. This results in the mode of the posterior distribution of $\hat{\pi}_w$ as $X_w/\mathcal{N}_w$, which is the conventional point

estimate often used in the literature (See [18, 21, 22, 25]). We define this posterior distribution of $\hat{\pi}_w$ as the naïve estimate, which is an estimate for the prevalence of ILI in the population, without consideration of user behaviours. In this work, we will split the cohort by vaccination status, and compare naïve estimates for the vaccinated and unvaccinated populations to the behaviour corrected estimates from the model proposed below.

Behaviour Model

We construct a model (Figure 1) that accounts for user behaviour and informs an estimate for the prevalence of ILI in the population, for vaccinated and unvaccinated individuals. For a household i in week w, let $Y_{i,w}$ be a binary indicator of an on-time report submitted by the master of the household and Π_w the probability of an individual having ILI in a given week w.

Then $Y_{i,w}$ can be modelled as a Bernoulli random variable,

$$Y_{i,w} \sim \text{Ber}\left(p_{i,w}\right), \tag{1}$$

where $p_{i,w}$ is the reporting probability.

Logistic regression can then be used to estimate a household's probability of reporting $p_{i,w}$, where for some set of predictors $\boldsymbol{x}_{i,w}$ and parameter vector $\boldsymbol{\gamma}$. The link function is defined as

$$\log\left(\frac{p_{i,w}}{1-p_{i,w}}\right) = \boldsymbol{\gamma}^T \boldsymbol{x}_{i,w}. \tag{2}$$

Let $I_{i,w}$ be a binary indicator of at least one participant of the household having ILI. Then, given that a household i has reported in week w, the household may have at least one participant reporting ILI symptoms, or the household has no participants reporting ILI.

Assume each household member has an independent probability $\Pi_w(V)$ of having ILI in week w given V, where V is a binary indicator for their vaccination status. Then let

$$\Pi_w(V) = \begin{cases} \lambda_w & \text{if } V = 1, \\ \pi_w & \text{if } V = 0. \end{cases}$$

The probability $\phi_{i,w}$ that a household of size n_i with v_w vaccinated individuals this week has at least one participant having ILI ($I_{i,w} = 1$) is then

$$\phi_{i,w} = 1 - (1-\pi_w)^{n_i - v_w}(1-\lambda_w)^{v_w}, \tag{3}$$

and the probability a household has no participants with ILI ($I_{i,w} = 0$) is $1 - \phi_{i,w}$.

Given that a household i reports in week w and has ILI status $I_{i,w}$, we can then model the event of reporting either on-time ($L^I_{i,w} = 1$) or late ($L^I_{i,w} = 0$) as another Bernoulli random variable,

$$L^I_{i,w} \sim \text{Ber}\left(\alpha^I_{i,w}\right), \tag{4}$$

where $\alpha^I_{i,w}$ is the probability that a master will report on-time, given the household ILI status and they have reported.

As in Equation 2, $\alpha^I_{i,w}$ can be modelled through logistic regression by

Table 1: Variables from the dataset used in constructing predictors.

Chronological	Epidemiological	Demographic
Survey week	Had fever	Participant ID
Year	Had cough	Household ID
Join date	Days of absence	Postcode
Submit date	Sought medical advice?	Age group
	Diagnosis	Health worker
	Test results	Vaccinated

$$\log\left(\frac{\alpha_{i,w}^{I}}{1-\alpha_{i,w}^{I}}\right) = \boldsymbol{\theta_I}^T \boldsymbol{\omega}_{i,w} \tag{5}$$

where $\boldsymbol{\omega_{i,w}}$ is a vector of predictors and $\boldsymbol{\theta_I}$ is a vector of parameters dependent on I.

So for every household in every week, there can be one of five outcomes $z_{i,w}$:

- $z_{i,w} = 1$: A household submits a report on-time, and no participants report having ILI ($Y_{i,w} = 1, I_{i,w} = 0, L_{i,w} = 1$),
- $z_{i,w} = 2$: A household submits a report late, and no participants report having ILI ($Y_{i,w} = 1, I_{i,w} = 0, L_{i,w} = 0$),
- $z_{i,w} = 3$: A household submits a report on-time, and at least one participant reports having ILI ($Y_{i,w} = 1, I_{i,w} = 1, L_{i,w} = 1$),
- $z_{i,w} = 4$: A household submits a report late, and at least one participant reports having ILI ($Y_{i,w} = 1, I_{i,w} = 1, L_{i,w} = 0$), or
- $z_{i,w} = 5$: A household does not submit a report ($Y_{i,w} = 0$).

Therefore, we can describe the outcome $z_{i,w}$ as following a categorical distribution:

$$z_{i,w} \sim \text{ Categorical}(\boldsymbol{p}), \tag{6}$$

where

$$\boldsymbol{p} = \begin{pmatrix} p_{i,w}(1-\phi_{i,w})\alpha_{i,w}^{I=0} \\ p_{i,w}(1-\phi_{i,w})(1-\alpha_{i,w}^{I=0}) \\ p_{i,w}\phi_{i,w}\alpha_{i,w}^{I=1} \\ p_{i,w}\phi_{i,w}(1-\alpha_{i,w}^{I=1}) \\ 1-p_{i,w} \end{pmatrix}.$$

Model Fitting

We use a Bayesian framework to estimate the posterior distribution of the parameters $\boldsymbol{\gamma}$, $\boldsymbol{\theta_I}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$, conditional on the data $\boldsymbol{X}$ and $\boldsymbol{y}$, via Bayes' rule:

$$\mathrm{P}(\boldsymbol{\gamma}, \boldsymbol{\theta_I}, \boldsymbol{\pi}, \boldsymbol{\lambda} | \boldsymbol{X}, \boldsymbol{y}) \propto \mathrm{P}\left(\boldsymbol{y} | \gamma, \boldsymbol{\theta_I}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{X}\right) \mathrm{P}\left(\boldsymbol{\gamma}, \boldsymbol{\theta_I}, \boldsymbol{\pi}, \boldsymbol{\lambda}\right). \tag{7}$$

Prior distributions for the parameters were:

$$\begin{aligned} \boldsymbol{\gamma} &\sim \mathrm{Norm}(\mathbf{0}, 0.7\mathcal{I}), \\ \boldsymbol{\theta_I} &\sim \mathrm{Norm}(\mathbf{0}, 0.7\mathcal{I}), \\ \boldsymbol{\pi}, \boldsymbol{\lambda} &\sim \mathrm{Unif}(0, 1), \end{aligned}$$

where $\mathbf{0}$ is the zero vector and $\mathcal{I}$ is the identity matrix.

The variance of the prior distributions for $\boldsymbol{\gamma}$ and $\boldsymbol{\theta_I}$ were chosen to provide a near uniform prior density for the probability of reporting and reporting on-time, given the distribution of the training set predictors. Figure S1 in the Supplementary Information shows the prior distribution transformed from the log odds scale to the probability scale, after multiplication with samples from the training set predictors. Results were not sensitive to this choice of variance.

All predictors $\boldsymbol{\omega_{i,w}}$ and $\boldsymbol{x_{i,w}}$ were scaled to be between 0 and 1, to allow for simplicity in comparing the predictive strength of each parameter. The set of predictors used for regression were derived from variables listed in Table 1, and the complete set of predictors used can be seen in Table S1.

To estimate the posterior distribution of the parameters of the model, we used Hamiltonian Monte Carlo (HMC) implemented in the software package Stan (version 2.18) [9]. HMC is particularly effective over other Markov chain Monte Carlo methods, such as Random Walk Metropolis-Hastings, in high dimensional parameter spaces such as in this analysis, and the Stan implementation uses adaptive tuning of algorithm parameters, reducing the need to tune the inference algorithm [6].

For model training and inference of results, 80% of the households in each year were used in the training set for the model, and the remainder left in a test set for cross-validation of model predictions. For each season examined, the first week of reports was used in determining the prior behaviours of users and used in calculating predictor values, but were not used to infer parameters of the model, as predictors involving prior behaviours could not be well defined in the first week.

Data Access

Reproduction of this study would require individual level data, which is precluded by ethics approval (The University of Adelaide Human Research Ethics Committee (HREC) H-2017-131).

Results

Comparison to Naïve Estimate

Without correcting for user behaviour, our results show that naïve estimates of ILI prevalence are biased and overestimate the actual prevalence. The distribution of the naïve estimates

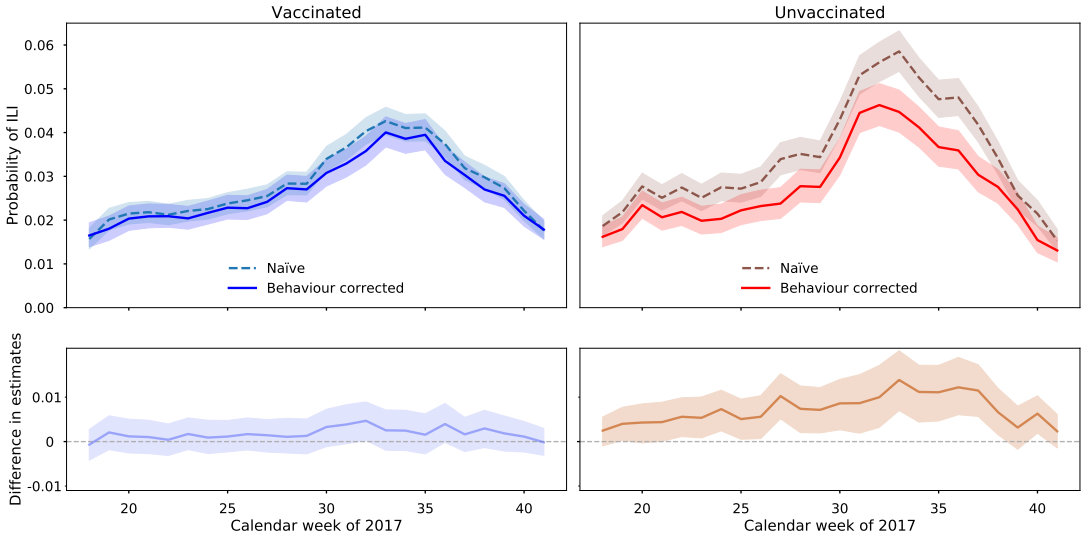

Figure 2: Comparison of behaviour corrected posterior estimate of the prevalence of ILI in the population (solid) to the naïve estimate (dashed) for the winter of 2017, and the difference between the two distributions over time for both vaccinated (blue) and unvaccinated (red) individuals. Lines represent the median of the distribution and shaded regions are 95% credible intervals of the posterior distributions.

have nearly all their probability density higher than the distribution of the behaviour corrected estimates for the ILI prevalence in vaccinated individuals λ_w and for unvaccinated individuals π_w (Figure 2). The effect is greatest for unvaccinated individuals. This bias was found to be greatest when prevalence of ILI was greatest, and implies that users may be more likely to report when they have ILI. This trend was observed for every year of analysis (see Supplementary Information). Figure 2 shows the results from training the model on the year 2017. While it can be tempting to use the difference between the vaccinated and unvaccinated estimates as a measure of influenza vaccine effectiveness, in this study we examine cases of ILI, not influenza. Any examination of vaccine effectiveness will require further modelling.

Reporting Behaviour

Predictors of reporting behaviour were determined from chronological, epidemiological and demographic data (Table 1). Given a consistent identification number for each individual and each household, predictors such as the week of the survey, whether the individual reported having a single symptom previously this year, and whether an individual is reporting on behalf of others can be determined. A full list of predictors used and their explanations can be found in Table S1.

Unsurprisingly, the greatest predictor for reporting behaviour was found to be the proportion of reports submitted on-time in the year so far for an individual, with all posterior samples of log odds ratio greater than 4 (see Figure S10). Here we will present predictors of interest, and

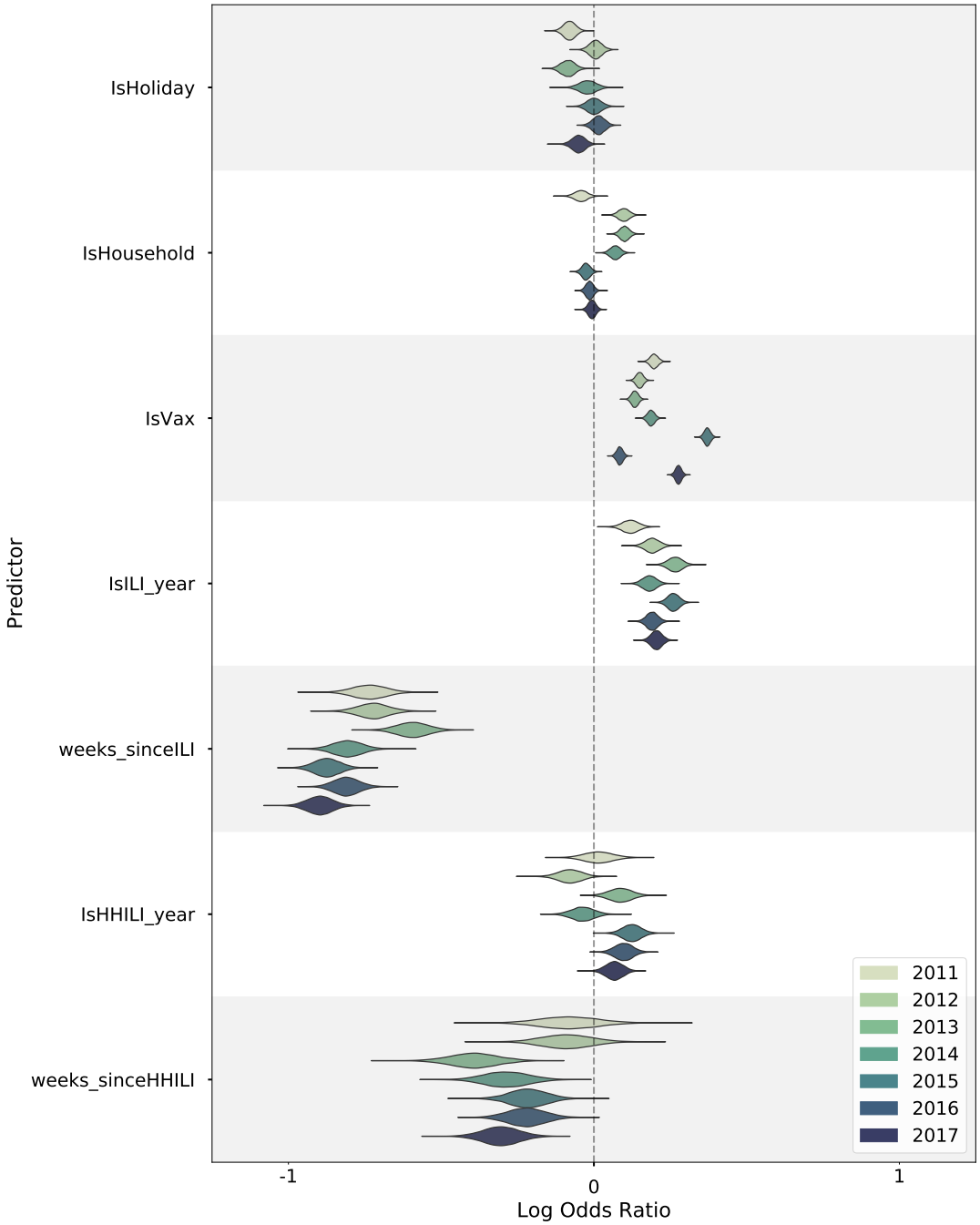

Figure 3: The marginal posterior densities of the log odds ratio of certain regression coefficients for predicting the probability of an individual reporting across all years examined. For the complete set of regression coefficients, see the Supplementary Information.

the marginal densities of all predictors are presented in the Supplementary Information.

A comparison of marginal posterior densities of regression coefficients for predicting the probability of reporting ($p_{i,w}$) can be seen in Figure 3. Those reporting on behalf of others were not found to be much more likely to report, inconsistent with analysis in other studies [3, 5]. These studies did not consider the past reporting behaviour of users in their analysis, and this may have confounded this predictor in previous work. Individuals who have reported being vaccinated for this season were also found to be more likely to report in any given week.

Individuals who had reported having ILI in previous weeks were much more likely to submit a report. However, the increasing number of weeks since the ILI occurred was a strong predictor for not submitting a report. Given no other differences, the model then predicts an increase in probability for an individual to report when they have reported having ILI, with the increase decaying over time as the number of weeks since reporting ILI increases. This relationship was also seen in later years (2015 to 2017), when there is a report of a participant, excluding the master, reporting having ILI. However, this effect was small. The posterior densities are remarkably consistent across the years, with the direction of the effect of the predictor consistent in all years, with the exception of those with small effects.

The current week was also used as a predictor for the probability a user will report, and their marginal probabilities of reporting declined sharply in the first third of the season, before a more shallow decline as the season continues (Figure 4). This could be explained by user fatigue, where dedication to participating in a voluntary system wanes through the year. This trend was also seen in every year studied (see Supplementary Information).

Reporting On-time

A comparison of densities of the log odds ratio of the regression coefficients for predicting the probability of an on-time report given a report was submitted ($\alpha^I_{i,w}$) is presented in Figure 5. During school holidays, individuals were less likely to report on-time, with the effect stronger for those who did not have ILI at the time. This may be due to disrupted schedules during the school holidays where emails may not be checked (and thus the reminder missed), before observing the emails once returning to work. The effect may be stronger for those who have ILI as they are reminded to report due to their symptoms. Reporting on behalf of others was a small predictor of reporting on-time in the earlier years (2011 to 2013), especially those reporting ILI. In later years (2014 to 2017), the effect was mixed based on ILI status. Households reporting with ILI were more likely to report on-time, while those who did not have ILI were more likely to report late.

Those who had reported having ILI previously in the year did not appear to be more or less likely to report on-time, despite being more likely to report in general, as discussed earlier. Individuals who identified as having an occupation that works with patients were slightly less likely to report on-time.

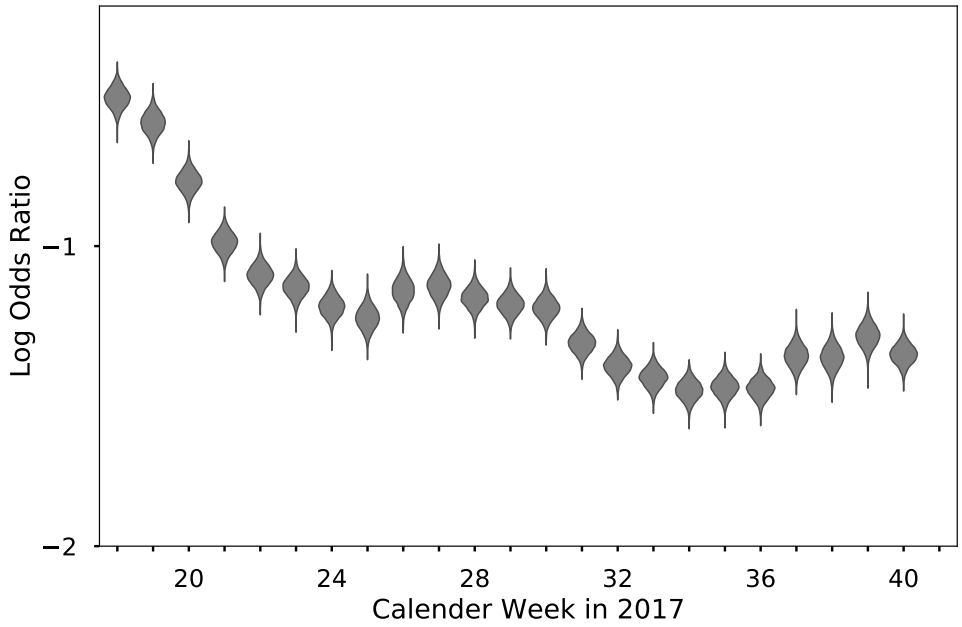

Figure 4: The marginal posterior densities of the log odds ratio of chronological regression coefficients for predicting the probability of an individual reporting in a given week of 2017. For all other years, please see Supplementary Information.

Model Validation

To validate the model, 20% of the households in each year were excluded from training the model and kept as a test set. Model predictions were compared to observations in the test set by taking 1000 samples of the parameters from the posterior distribution, and simulating outcomes from the model for each sample and each household in the test set. The simulated data classified each household in each week into one of the five outcomes (Figure 1).

The proportion of reporting households with at least one participant with ILI, and the participation rate of households of the simulated data, were compared to observations from the test set. The proportion of reporting households with at least one participant with ILI was calculated by dividing the number of households reporting on-time with ILI by the total number of households reporting on-time. The participation rate of households is determined by dividing the number of households reporting by the total number of households registered who have submitted at least one report in that year. Both summary statistics from the data were similar to simulations from the model on the test set, and these comparisons can be seen in the Supplementary Information Figures S20 to S26. The simulated data from the model matches well to the actual observations from the test set and appears to retain the autocorrelation or time dependence of the actual test outcomes, without time dependence being explicitly included in the model. As the model predictions correspond well to outcomes observed in the test set in both summary statistics, the model shows no indication of bias of observable outcomes at a

population level.

Discussion

Online participatory health surveillance systems strive to provide near real-time estimates of disease prevalence, and yield complementary insights to traditional public health systems. However, the voluntary nature of these systems, and the reliance on user participation, need to be considered.

In this work we find that history of ILI in the year, as well as other demographic factors, impacts the probability a user will submit a report. At a population level in FluTracking, this results in overestimation of the prevalence of ILI in the population when using the naïve estimate. This difference is greatest near the peak prevalence of ILI. This may be due to users being triggered to report by their symptoms, and being more likely to report in voluntary health surveillance systems when directly affected by the illness.

Participation in FluTracking, derived as the number of on-time reports divided by the number of registered users reporting at least one in a given year, generally ranges between 70% and 90% for any given year and any given week. This is much higher than rates observed in Influenzanet and Flu Near You, where previous studies have shown that 70% and 35% of users respectively submit more than 3 reports in a season [3, 5]. This is despite FluTracking having proportionally much higher numbers of participants than the European and US systems, relative to the respective populations of these regions. While the variance of participation over time has not been examined in the other systems, correcting for user behaviour would potentially be even more critical than observed here.

The prevalence of ILI was higher for those who were unvaccinated compared to those who were vaccinated in the peak weeks of prevalence. However, this should not be seen as a measure of influenza vaccine effectiveness. The vaccine is not targeted to ILI, and the difference in prevalence over time does not consider the population size, or the number of cases of influenza reduced in a season.

However, there exists the potential to construct a measure of vaccine effectiveness by comparing the two groups. The conventional metric of test-negative cases [17] does not extend simply to the behaviour corrected estimates presented here. However, with the potential to correct for potential bias in the data and reporting behaviour across different years, further studies may help inform vaccine effectiveness estimates in near real time.

Whilst past behaviours and illness status were used to predict user participation within the season, this study did not attempt to train the model across seasons. Training the model across every season would allow for users' behaviour in past years to inform behaviour in later seasons. However, the computational difficulty in substantially increasing the size of the training data, and the number of parameters involved, remains a challenge left for further work.

Predictors for user behaviour have only been taken from within the FluTracking data set. Social media and news coverage [20] and public awareness [14] are some examples where external factors can influence behaviour during an epidemic. Inclusion of these factors in the model may further improve estimates of disease prevalence where user behaviour may be significant.

This analysis assumes each individual has an equal probability of having ILI in a given week. Incorporating a spatial mechanistic model, particularly given the postcode and demographic information in the data set may enable further insights into the mechanisms that drive the transmission of influenza and ILI in Australia. Recent analysis has shown the onset of influenza epidemics is largely synchronised across regional areas [15].

With this model we have shown that user behaviour can have a significant effect on disease prevalence estimates drawn from the data. The framework developed herein, which elucidates drivers of user reporting in voluntary health surveillance systems and improves estimates of disease prevalence, should prove useful as digital data streams burgeon in epidemiology.

Acknowledgements

We would like to acknowledge the FluTracking team for their support, and all the FluTracking users for participating. This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide. Funding and support was provided by the Data To Decisions Collaborative Research Centre (D2D CRC), the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), the National Health and Medical Research Council (NHMRC) Centre of Research Excellence in Policy Relevant Infectious diseases Simulation and Mathematical Modelling (PRISM^2), and an Australian Government Research Training Program (RTP) Scholarship.

References

[1] Australian Government. *Special dates and events: School term dates*. 2018. URL: `https://www.australia.gov.au/about-australia/special-dates-and-events/school-term-dates` (visited on 11/01/2018).

[2] Australian Government Department of Health. "2018 Influenza Season in Australia: A summary from the National Influenza Surveillance Committee". November (2018), pp. 1–5. URL: `https://www.health.gov.au/internet/main/publishing.nsf/Content/cda-surveil-ozflu-flucurr.htm`.

[3] Paolo Bajardi, Alessandro Vespignani, Sebastian Funk, Ken T.D. Eames, W. John Edmunds, Clément Turbelin, Marion Debin, Vittoria Colizza, Ronald Smallenburg, Carl E. Koppeschaar, Ana O. Franco, Vitor Faustino, Annasara Carnahan, Moa Rehn, and Daniela Paolotti. "Determinants of follow-up participation in the internet-based european influenza surveillance platform influenzanet". *Journal of Medical Internet Research* 16.3 (2014), pp. 1–15. ISSN: 14388871. DOI: `10.2196/jmir.3010`.

[4] Kristin Baltrusaitis, John S Brownstein, Samuel V Scarpino, Eric Bakota, Adam W Crawley, Giuseppe Conidi, Julia Gunn, Josh Gray, Anna Zink, and Mauricio Santillana. "Comparison of crowd-sourced, electronic health records based, and traditional health-care based influenza-tracking systems at multiple spatial resolutions in the United States of America". *BMC infectious diseases* 18.1 (2018), pp. 1–8. ISSN: 1471-2334. DOI: `10.1186/s12879-018-3322-3`.

[5] Kristin Baltrusaitis, Mauricio Santillana, Adam W Crawley, Rumi Chunara, Mark Smolinski, and John S Brownstein. "Determinants of Participants' Follow-Up and Characterization of Representativeness in Flu Near You, A Participatory Disease Surveillance System". *JMIR Public Health and Surveillance* 3.2 (2017), e18. ISSN: 2369-2960. DOI: 10.2196/publichealth.7304. URL: http://publichealth.jmir.org/2017/2/e18/.

[6] Michael Betancourt, Simon Byrne, Sam Livingstone, and Mark Girolami. "The geometric foundations of Hamiltonian Monte Carlo". *Bernoulli* 23.4A (2017), pp. 2257–2298. ISSN: 1350-7265. DOI: 10.3150/16-bej810.

[7] Pietro Cantarelli et al. "The representativeness of a European multi-center network for influenza-like-illness participatory surveillance". *BMC Public Health* 14.1 (2014), pp. 1–17. ISSN: 14712458. DOI: 10.1186/1471-2458-14-984.

[8] Sandra J Carlson, Daniel Cassano, Michelle T Butler, David N Durrheim, and Craig Dalton. "Flutracking weekly online community survey of influenza-like illness annual report, 2016". *Communicable Diseases Intelligence* 43 (2019). DOI: 10.33321/cdi.2019.43.15.

[9] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. "Stan: A Probabilistic Programming Language". *Journal of Statistical Software* 76.1 (2017). ISSN: 1548-7660. DOI: 10.18637/jss.v076.i01. URL: http://www.jstatsoft.org/v76/i01/.

[10] Rumi Chunara, Susan Aman, Mark Smolinski, and John S. Brownstein. "Flu Near You: An Online Self-reported Influenza Surveillance System in the USA". *Online Journal of Public Health Informatics* 5.1 (2013), p. 2579. ISSN: 1947-2579. DOI: 10.5210/ojphi.v5i1.4456. URL: http://journals.uic.edu/ojs/index.php/ojphi/article/view/4456.

[11] Hazel J Clothier, James E Fielding, and Heath A Kelly. "An evaluation of the Australian Sentinel Practice Research Network (ASPREN) surveillance for influenza-like illness." *Communicable Diseases Intelligence* 29.3 (2005), pp. 231–247. ISSN: 1447-4514.

[12] Robert C. Cope, Joshua V. Ross, Monique Chilver, Nigel P. Stocks, and Lewis Mitchell. "Characterising seasonal influenza epidemiology using primary care surveillance data". *PLoS Computational Biology* 14.8 (2018), pp. 1–21. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006377.

[13] Craig Dalton, David Durrheim, John Fejsa, Lynn Francis, Sandra Carlson, Edouard Tursan D'Espaignet, and Frank Tuyl. "Flutracking: A weekly Australian community online survey of influenza-like illness in 2006, 2007 and 2008". *Communicable Diseases Intelligence Quarterly Report* 33.IlI (2009), pp. 316–322.

[14] S. Funk, E. Gilad, C. Watkins, and V. A. A. Jansen. "The spread of awareness and its impact on epidemic outbreaks". *Proceedings of the National Academy of Sciences* 106.16 (2009), pp. 6872–6877. ISSN: 0027-8424. DOI: 10.1073/pnas.0810762106. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0810762106.

[15] Jemma L Geoghegan, Aldo F Saavedra, Sebastian Duchene, Sheena Sullivan, Ian Barr, and Edward C Holmes. "Continental synchronicity of human influenza virus epidemics despite climactic variation". *PLoS Pathogens* 14.1 (2018), pp. 1–16. DOI: 10.1371/journal.ppat.1006780.

[16] Caroline Guerrisi et al. "Participatory syndromic surveillance of influenza in Europe". *Journal of Infectious Diseases* 214.May (2016), S386–S392. ISSN: 15376613. DOI: 10.1093/infdis/jiw280.

[17] Michael L. Jackson and Jennifer C. Nelson. "The test-negative design for estimating influenza vaccine effectiveness". *Vaccine* 31.17 (2013), pp. 2165–2168. ISSN: 0264410X. DOI: 10.1016/j.vaccine.2013.02.053. URL: http://dx.doi.org/10.1016/j.vaccine.2013.02.053.

[18] Kyriaki Kalimeri, Matteo Delfino, Ciro Cattuto, Daniela Perrotta, Vittoria Colizza, Caroline Guerrisi, Clement Turbelin, Jim Duggan, John Edmunds, Chinelo Obi, Richard Pebody, Ricardo Mexia, Ana Franco, Yamir Moreno, Sandro Meloni, Carl Koppeschaar, Charlotte Kjelsø, Daniela Paolotti, and Ricardo Jorge. "Unsupervised Extraction of Epidemic Syndromes from Participatory Influenza Surveillance Self-reported Symptoms" (2018), pp. 1–24. DOI: 10.1101/314591. URL: http://dx.doi.org/10.1101/314591.

[19] Carl E Koppeschaar, Vittoria Colizza, Caroline Guerrisi, Clément Turbelin, Jim Duggan, W John Edmunds, Charlotte Kjelsø, Ricardo Mexia, Yamir Moreno, Sandro Meloni, Daniela Paolotti, Daniela Perrotta, Edward van Straten, and Ana O Franco. "Influenzanet: Citizens Among 10 Countries Collaborating to Monitor Influenza in Europe". *JMIR Public Health and Surveillance* 3.3 (2017), e66. ISSN: 2369-2960. DOI: 10.2196/publichealth.7429. URL: http://publichealth.jmir.org/2017/3/e66/.

[20] Lewis Mitchell and Joshua V Ross. "A data-driven model for influenza transmission incorporating media effects". *The Royal Society Open Science* 3.10 (2016), pp. 1–10. DOI: https://doi.org/10.1098/rsos.160481.

[21] Sander P. van Noort, Claudia T. Codeco, Carl E. Koppeschaar, Marc van Ranst, Daniela Paolotti, and M. Gabriela M Gomes. "Ten-year performance of Influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour". *Epidemics* 13 (2015), pp. 28–36. ISSN: 18780067. DOI: 10.1016/j.epidem.2015.05.001.

[22] D. Paolotti, A. Carnahan, V. Colizza, K. Eames, J. Edmunds, G. Gomes, C. Koppeschaar, M. Rehn, R. Smallenburg, C. Turbelin, S. Van Noort, and A. Vespignani. "Web-based participatory surveillance of infectious diseases: The Influenzanet participatory surveillance experience". *Clinical Microbiology and Infection* 20.1 (2014), pp. 17–21. ISSN: 14690691. DOI: 10.1111/1469-0691.12477. URL: http://dx.doi.org/10.1111/1469-0691.12477.

[23] Carrie Reed et al. "Estimating Influenza Disease Burden from Population-Based Surveillance Data in the United States". *PLoS ONE* 10.3 (2015), e0118369. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0118369. URL: http://dx.plos.org/10.1371/journal.pone.0118369.

[24] Moa Rehn, Anna Sara Carnahan, Hanna Merk, Sharon Kühlmann-Berenzon, Ilias Galanis, Annika Linde, and Olof Nyrén. "Evaluation of an internet-based monitoring system for influenza-like illness in Sweden". *PLoS ONE* 9.5 (2014), pp. 1–10. ISSN: 19326203. DOI: 10.1371/journal.pone.0096740.

[25] Mark S. Smolinski, Adam W. Crawley, Kristin Baltrusaitis, Rumi Chunara, Jennifer M. Olsen, Oktawia Wójcik, Mauricio Santillana, Andre Nguyen, and John S. Brownstein. "Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons". *American Journal of Public Health* 105.10 (2015), pp. 2124–2130. ISSN: 15410048. DOI: 10.2105/AJPH.2015.302696.

[26] World Health Organization. *Global Epidemiological Surveillance Standards for Influenza*. 2014. DOI: 10.1007/s13398-014-0173-7.2. arXiv: arXiv:1011.1669v3. URL: https://www.who.int/influenza/surveillance{_}monitoring/ili{_}sari{_}surveillance{_}case{_}definition/en/ (visited on 01/30/2019).

Supplementary Information

Methods

Table S.1 presents all predictors used in the analysis. Figure S.1 depicts the choice of prior distribution for the regression coefficients.

Results

All figures are presented in reverse chronological order, as the more recent years have the most directly comparable populations of participants to 2017, the year that was primarily discussed in the main paper. Figures S.2 to S.7 compare the naïve estimate to model posterior distributions of ILI prevalence in vaccinated and unvaccinated individuals across all years studied. 2017 results are presented in the main article. Figures S.8 to S.11 show the marginal posterior densities of the log odds ratio of all regression coefficients for predicting the probability of an individual reporting across all years examined. Figures S.12 to S.14 show the marginal posterior densities of the log odds ratio of all regression coefficients for predicting the probability of an individual reporting on-time given they have reported and their ILI status, across all years examined. Figures S.15 to S.21 compare the model predictions against summary statistics of the test sets across all years examined.

Table S.1: List of predictor variables used and their definitions. Asterisks indicate variables in which the value is divided by the number of weeks in the season to scale the predictor between 0 and 1. School holidays breaks are sourced from Australian government sources [1].

Predictor	Description
HH has reported having ILI previously	Has the household (HH) reported having ILI previously this year, exclusive of the master.
HH has reported having symptoms	Has the HH reported having either fever or cough, but not both, previously this year, exclusive of the master.
Master works with patients	Is the master in a patient facing occupation
Report is during a State/Territory school holiday	Does the current survey week cover days that are in the master's state school holiday period.
Master is reporting on behalf of others	Is the master reporting on behalf of others.
Master reported having ILI previously	Has the master reported having ILI previously this year.
Master reported having a symptom previously	Has the master reported having either fever or cough, but not both, previously this year.
Master is vaccinated	Has the master reported being vaccinated for influenza this year.
Proportion of reports so far	Proportion of reports submitted on-time so far this year.
Number of weeks since HH reported having ILI*	The number of weeks since the HH, but not the master, was reported to have ILI.
Number of weeks since HH reported having symptoms*	The number of weeks since the HH, but not the master, was reported to have a fever or a cough, but not both.
Number of weeks since reporting ILI*	The number of weeks since the master reported having ILI.
Number of weeks since reporting symptom*	The number of weeks since the master reported having a fever or a cough, but not both.
Week	Categorical predictor for the week of the survey. Not used in predicting whether a report will be on-time or late given a report has been submitted.
Intercept	Intercept variable to capture base level reporting rate. Not used in predicting whether a report will be on-time or late given a report has been submitted.

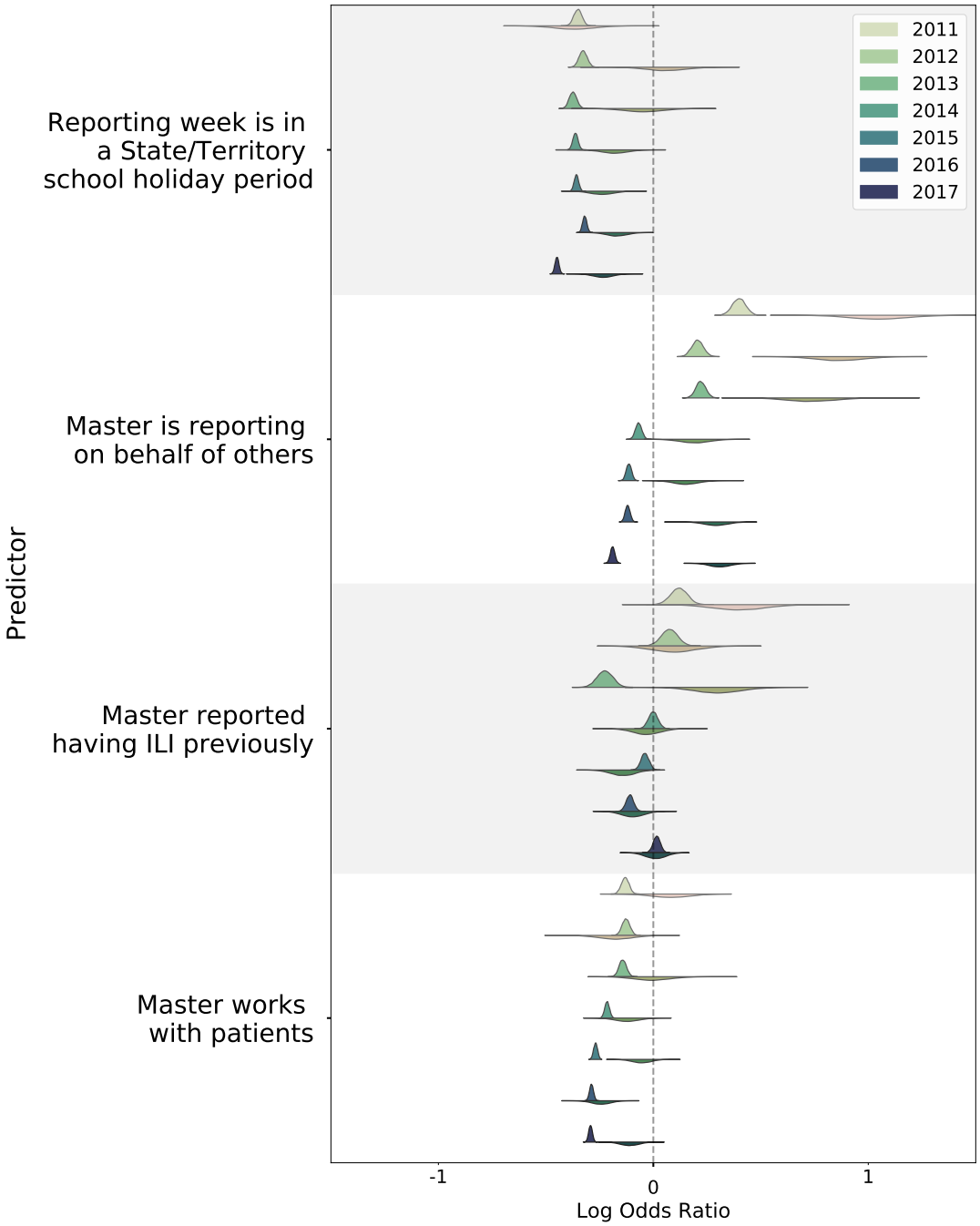

Figure 5: The marginal posterior densities of the log odds ratio of certain regression coefficients for predicting the probability of an individual reporting on-time given they have reported and their ILI status, across all years examined. The upper density is the density of the regression coefficient given that the household has no participant with ILI, and the lower density is the density given the household has at least one participant with ILI. For the complete set of regression coefficients, see the Supplementary Information.

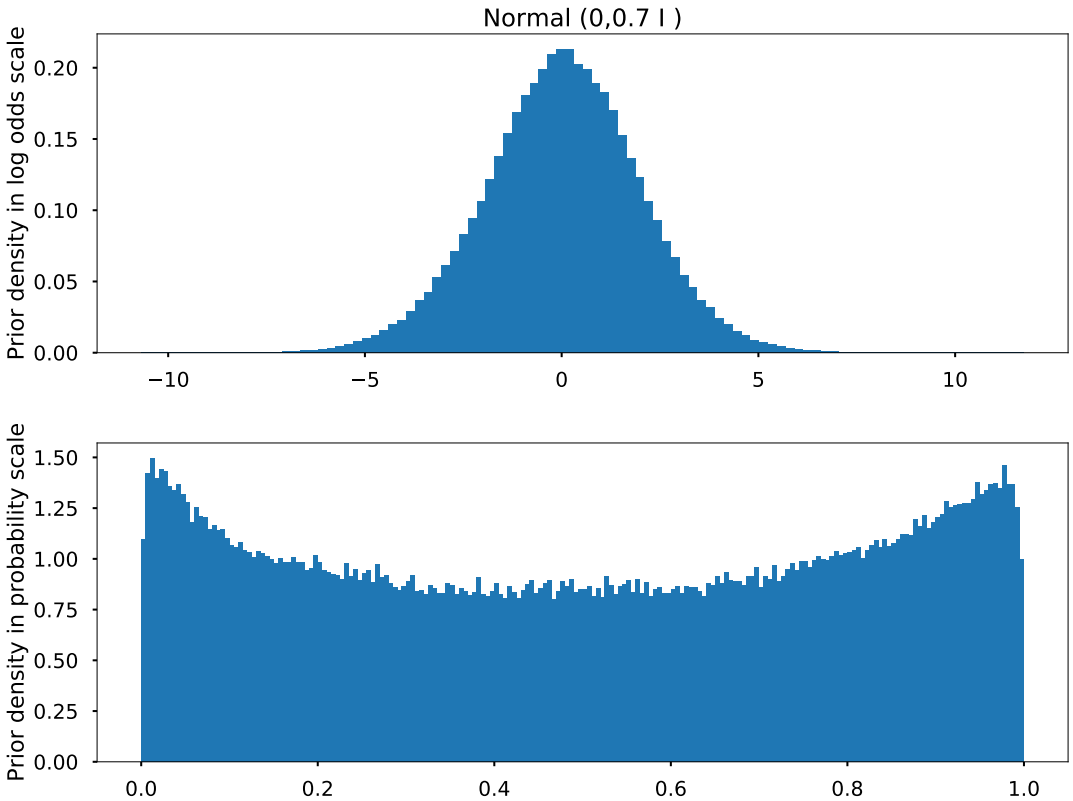

Figure S.1: The prior distribution of the regression coefficients transformed from the log odds scale to the probability scale via the logit function, after multiplication with 1000 samples from the training set predictors of 2017. The covariance matrix 0.7 $\mathcal{I}$ was chosen for the model as it was somewhat uniform across the probability scale, with some skewness towards the boundaries.

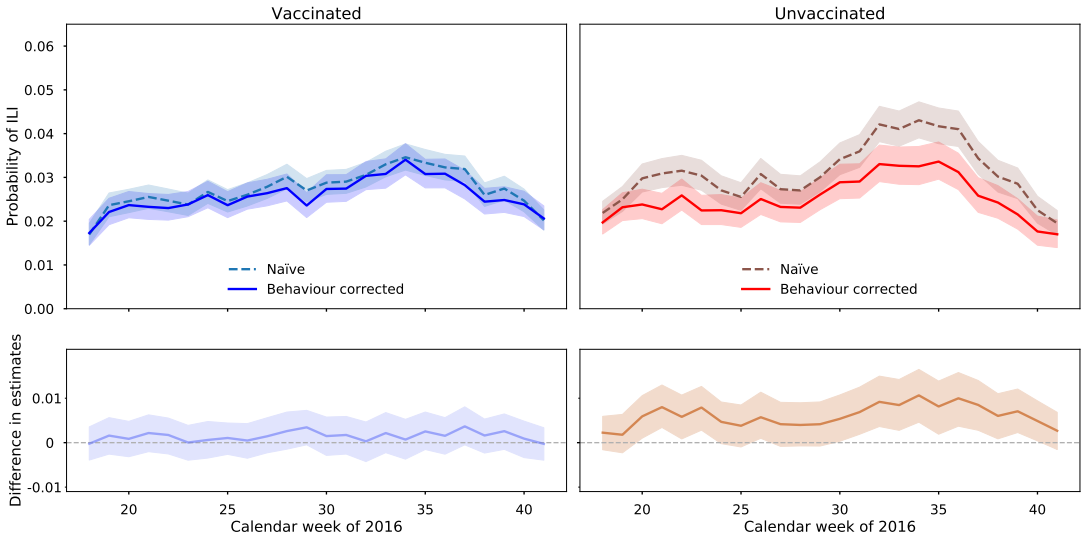

Figure S.2: Comparison of model posterior estimate of the prevalence rate of ILI in the population to the naïve estimate for the winter of 2016, and the difference between the two distributions over time for both vaccinated and unvaccinated individuals. Lines represent the median of the distribution and shaded regions are 95% credible intervals of the posterior distributions.

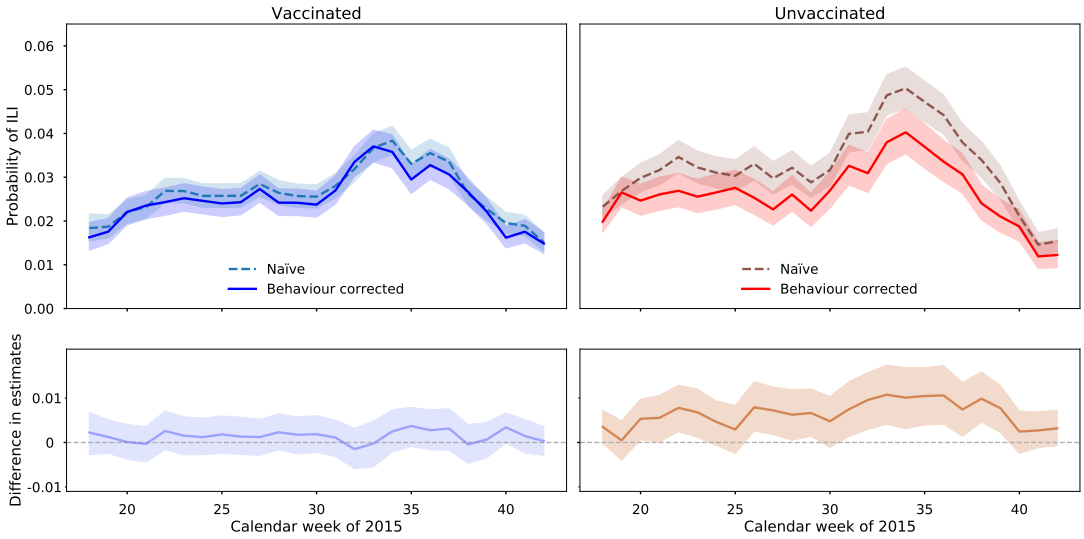

Figure S.3: Comparison of model posterior estimate of the prevalence rate of ILI in the population to the naïve estimate for the winter of 2015, and the difference between the two distributions over time for both vaccinated and unvaccinated individuals. Lines represent the median of the distribution and shaded regions are 95% credible intervals of the posterior distributions.

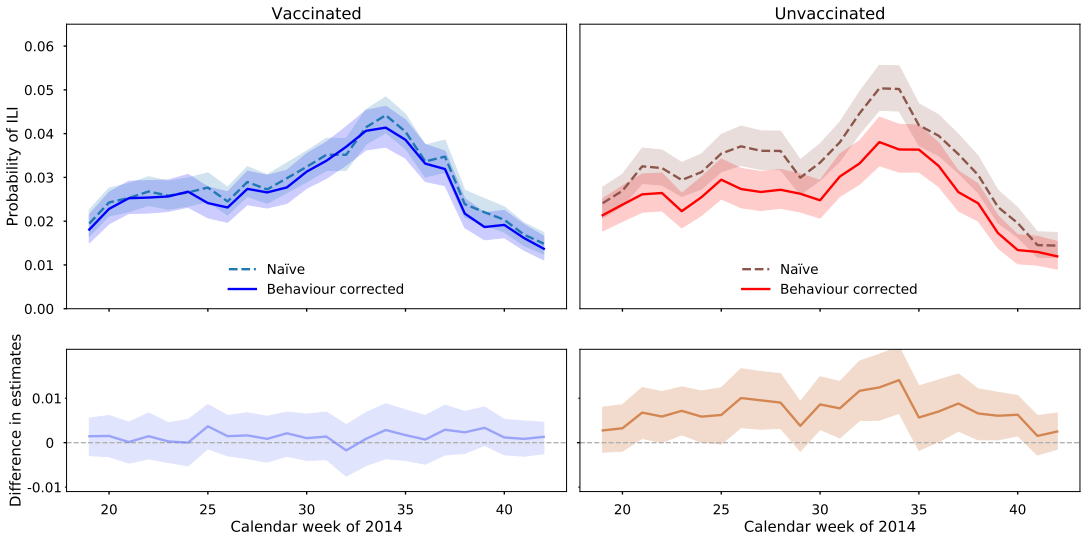

Figure S.4: Comparison of model posterior estimate of the prevalence rate of ILI in the population to the naïve estimate for the winter of 2014, and the difference between the two distributions over time for both vaccinated and unvaccinated individuals. Lines represent the median of the distribution and shaded regions are 95% credible intervals of the posterior distributions.

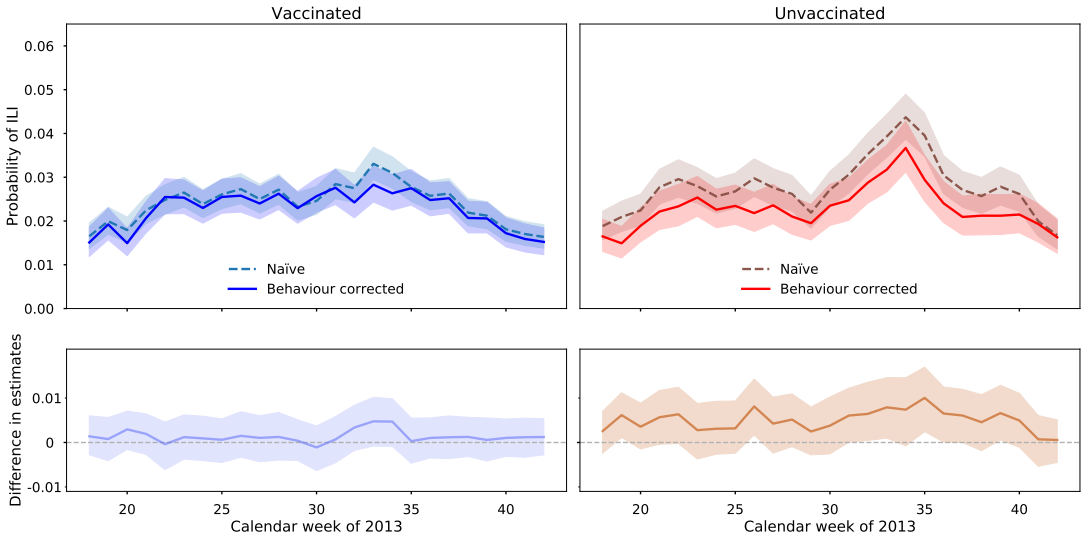

Figure S.5: Comparison of model posterior estimate of the prevalence rate of ILI in the population to the naïve estimate for the winter of 2013, and the difference between the two distributions over time for both vaccinated and unvaccinated individuals. Lines represent the median of the distribution and shaded regions are 95% credible intervals of the posterior distributions.

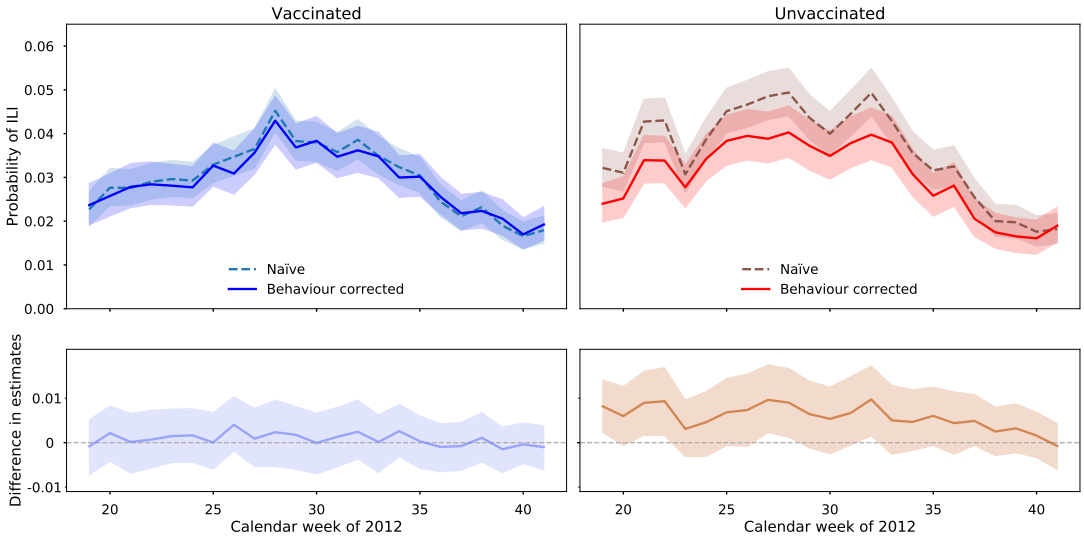

Figure S.6: Comparison of model posterior estimate of the prevalence rate of ILI in the population to the naïve estimate for the winter of 2012, and the difference between the two distributions over time for both vaccinated and unvaccinated individuals. Lines represent the median of the distribution and shaded regions are 95% credible intervals of the posterior distributions.

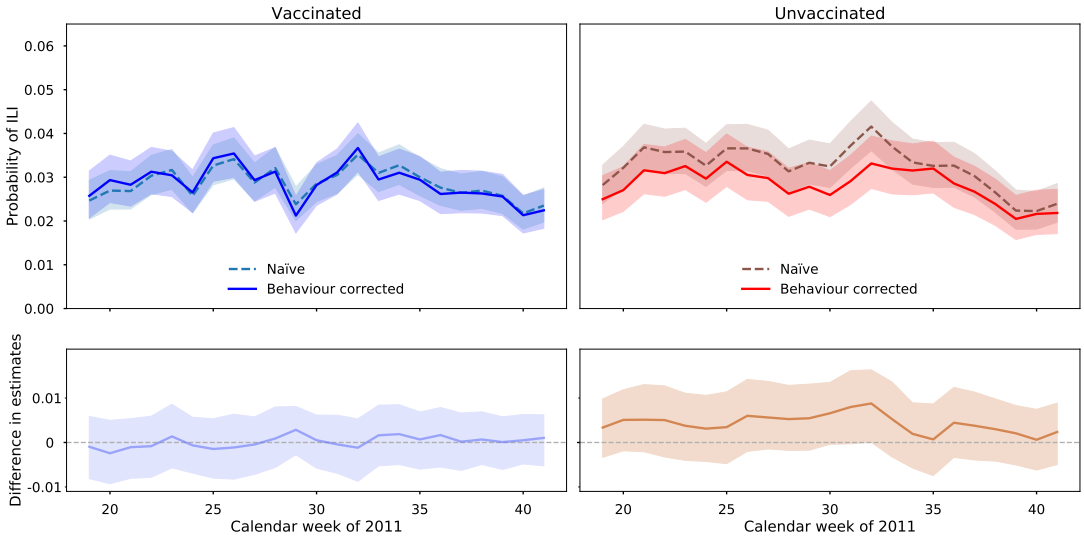

Figure S.7: Comparison of model posterior estimate of the prevalence rate of ILI in the population to the naïve estimate for the winter of 2011, and the difference between the two distributions over time for both vaccinated and unvaccinated individuals. Lines represent the median of the distribution and shaded regions are 95% credible intervals of the posterior distributions.

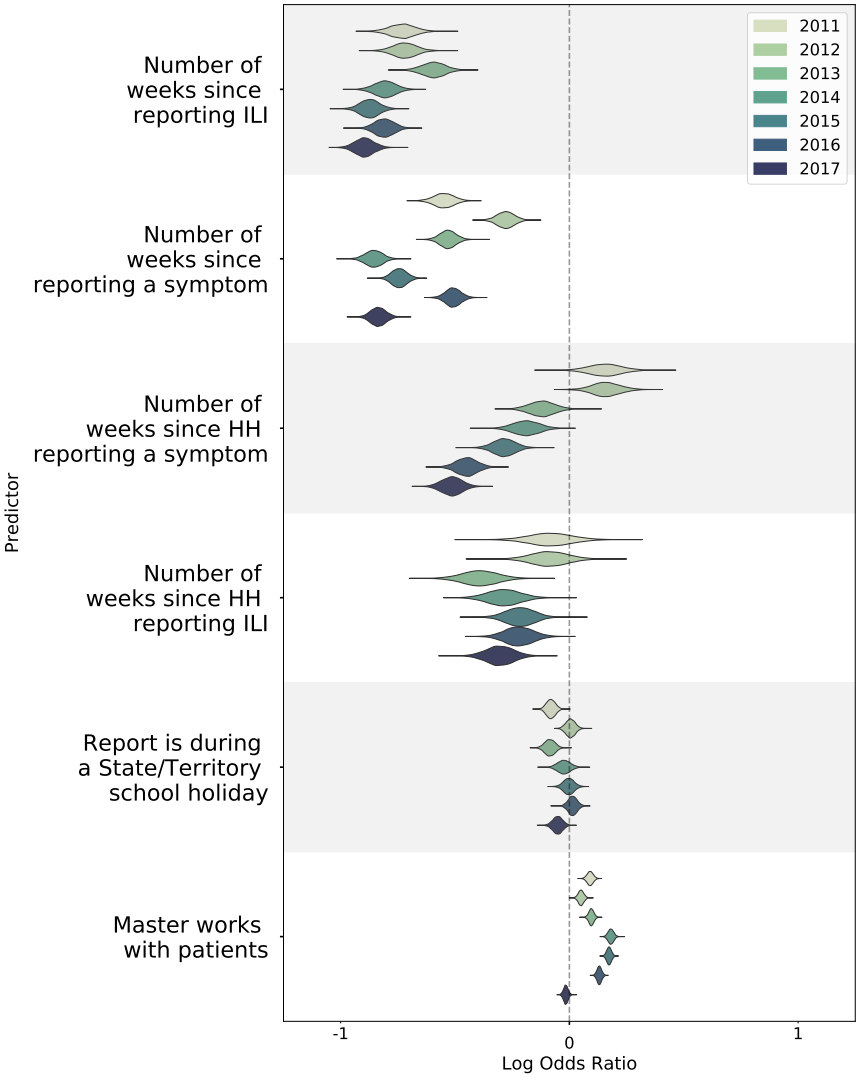

Figure S.8: The marginal posterior densities of the log odds ratio of the first half of non-chronological regression coefficients for predicting the probability of an individual reporting for all years.

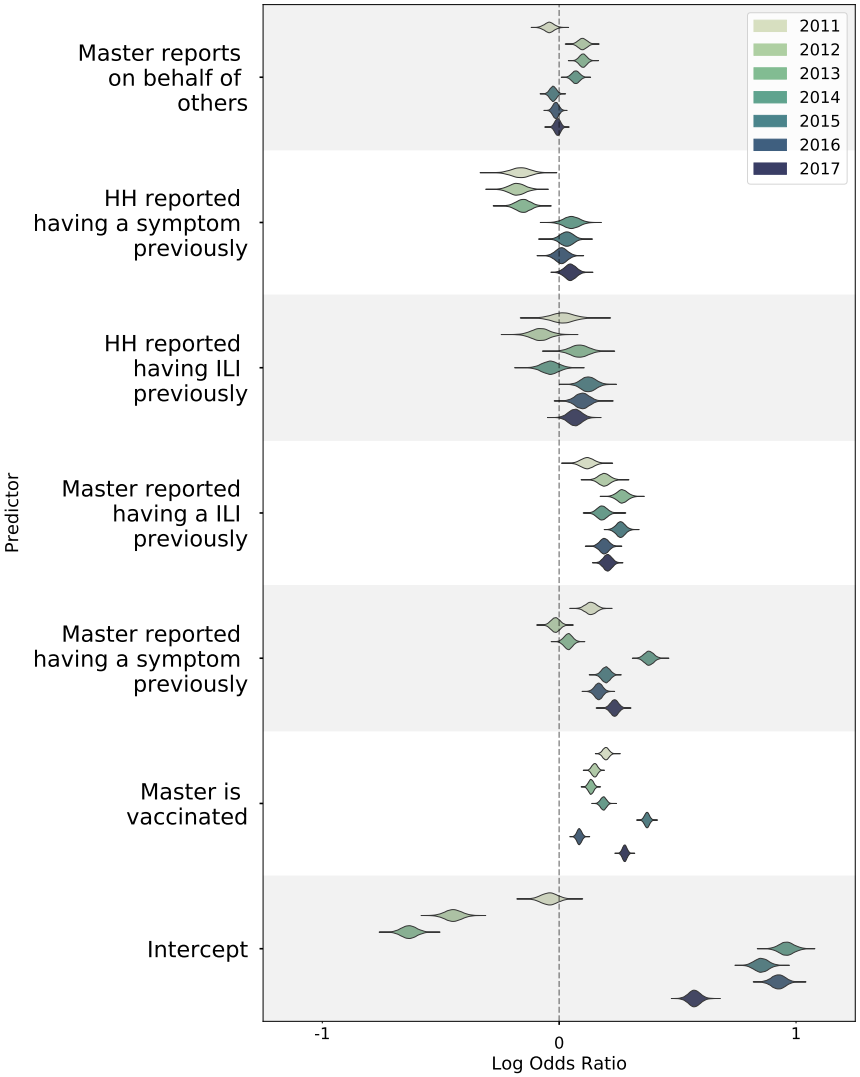

Figure S.9: The marginal posterior densities of the log odds ratio of the second half of non-chronological regression coefficients for predicting the probability of an individual reporting for all years.

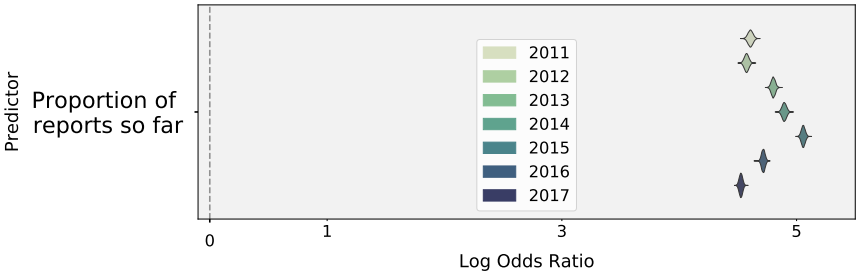

Figure S.10: The marginal posterior densities of the log odds ratio of the regression coefficients for the proportion of reports submitted on-time for predicting the probability of an individual reporting for all years.

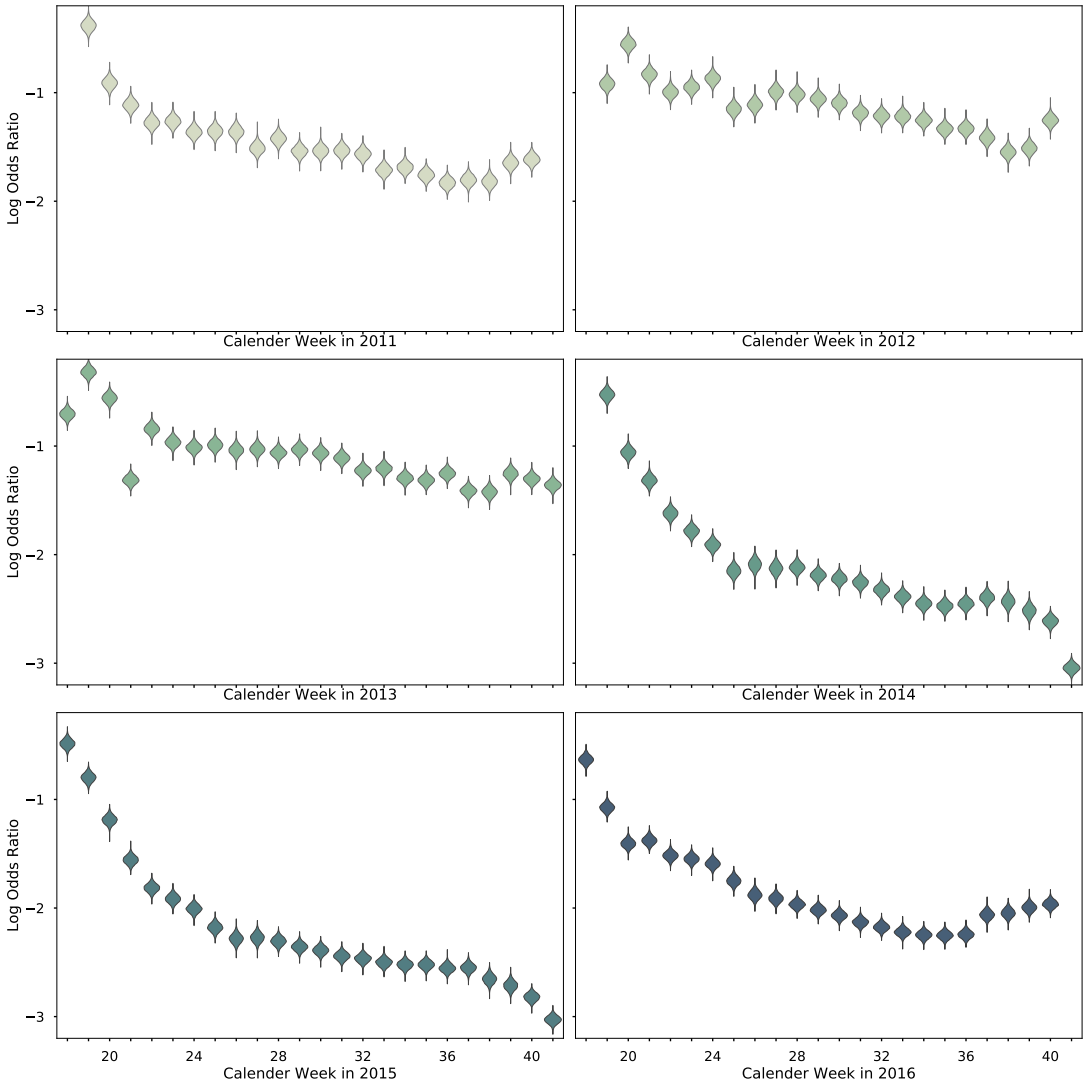

Figure S.11: The marginal posterior densities of the log odds ratio of chronological regression coefficients for predicting the probability of an individual reporting in a given week for years 2011 to 2016.

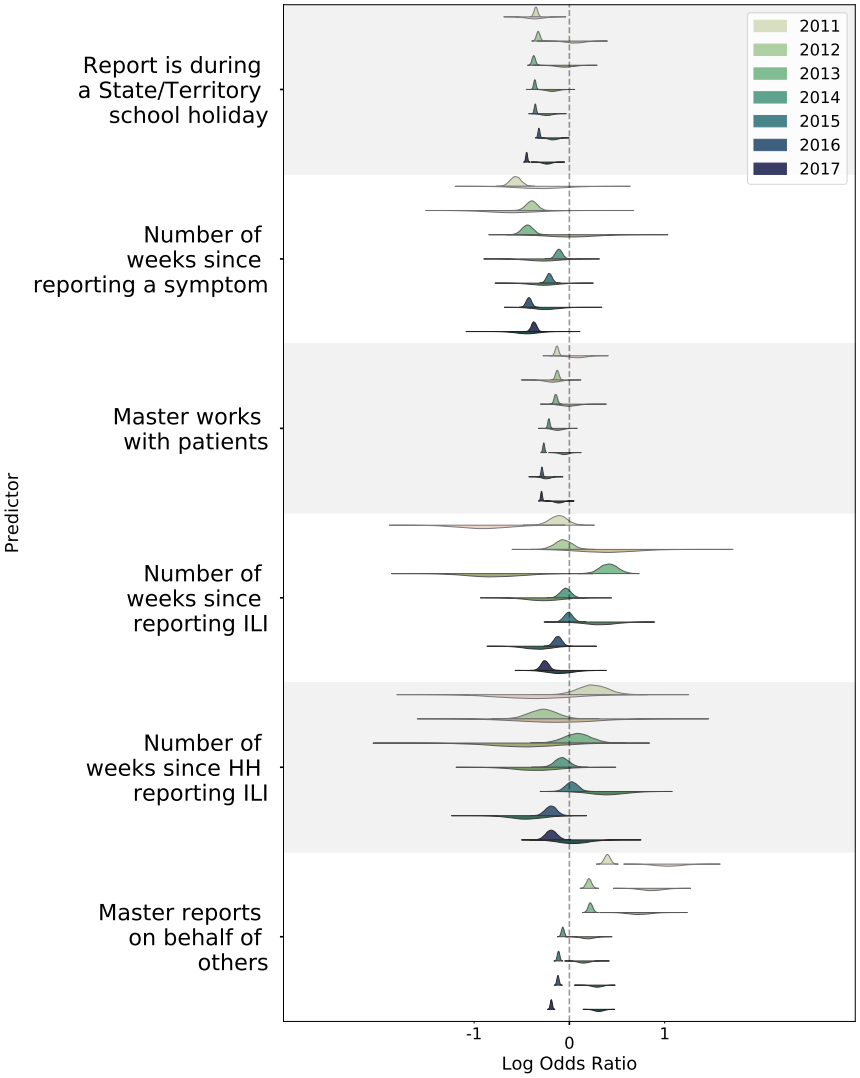

Figure S.12: The marginal posterior densities of the log odds ratio of the first half of regression coefficients for predicting the probability of an individual reporting on-time given they have reported and their ILI status, across all years examined. The upper density is the density of the regression coefficient given that the household has no participant with ILI, and the lower density is the density given the household has at least one participant with ILI.

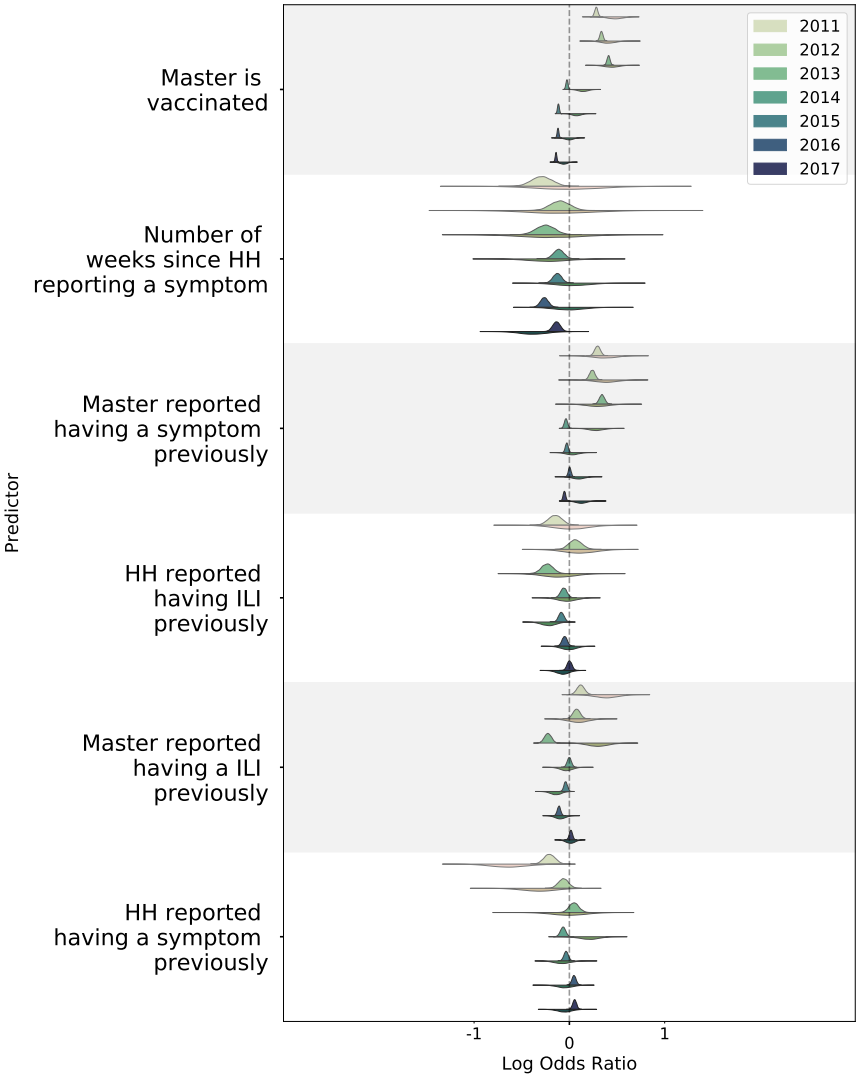

Figure S.13: The marginal posterior densities of the log odds ratio of the second half of regression coefficients for predicting the probability of an individual reporting on-time given they have reported and their ILI status, across all years examined. The upper density is the density of the regression coefficient given that the household has no participant with ILI, and the lower density is the density given the household has at least one participant with ILI.

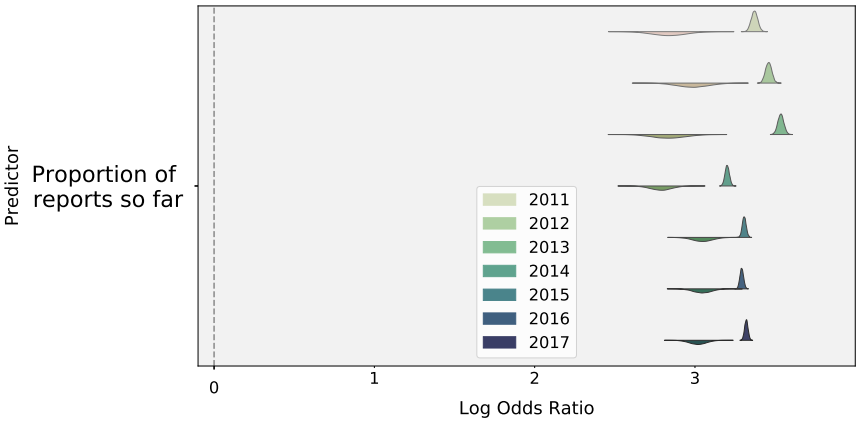

Figure S.14: The marginal posterior densities of the log odds ratio of the regression coefficients for the proportion of reports submitted on-time for predicting the probability of an individual reporting on-time given they have reported and their ILI status, across all years examined.

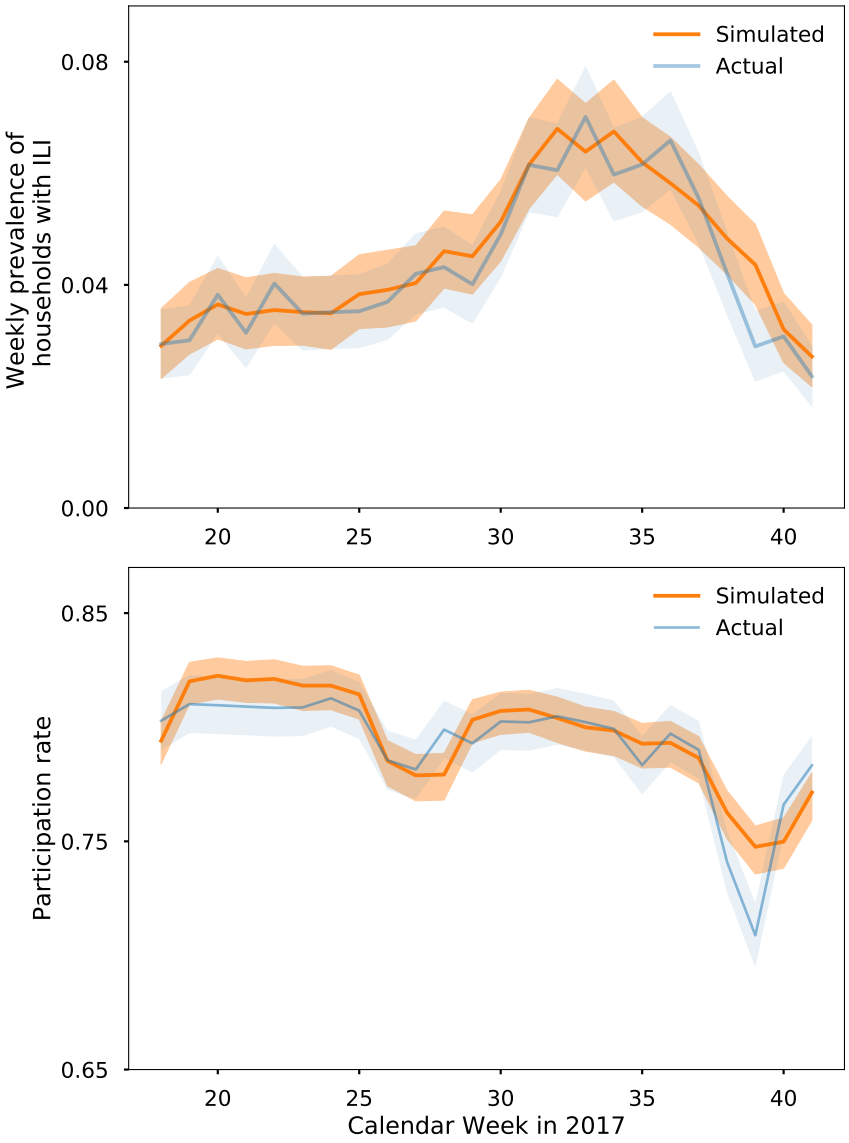

Figure S.15: Cross validation of model predictions with actual outcomes of test set for the 2017 season. Lines represent the median and shaded regions the 95% credible intervals. The model is able to generate the data observed in the test set with high probability.

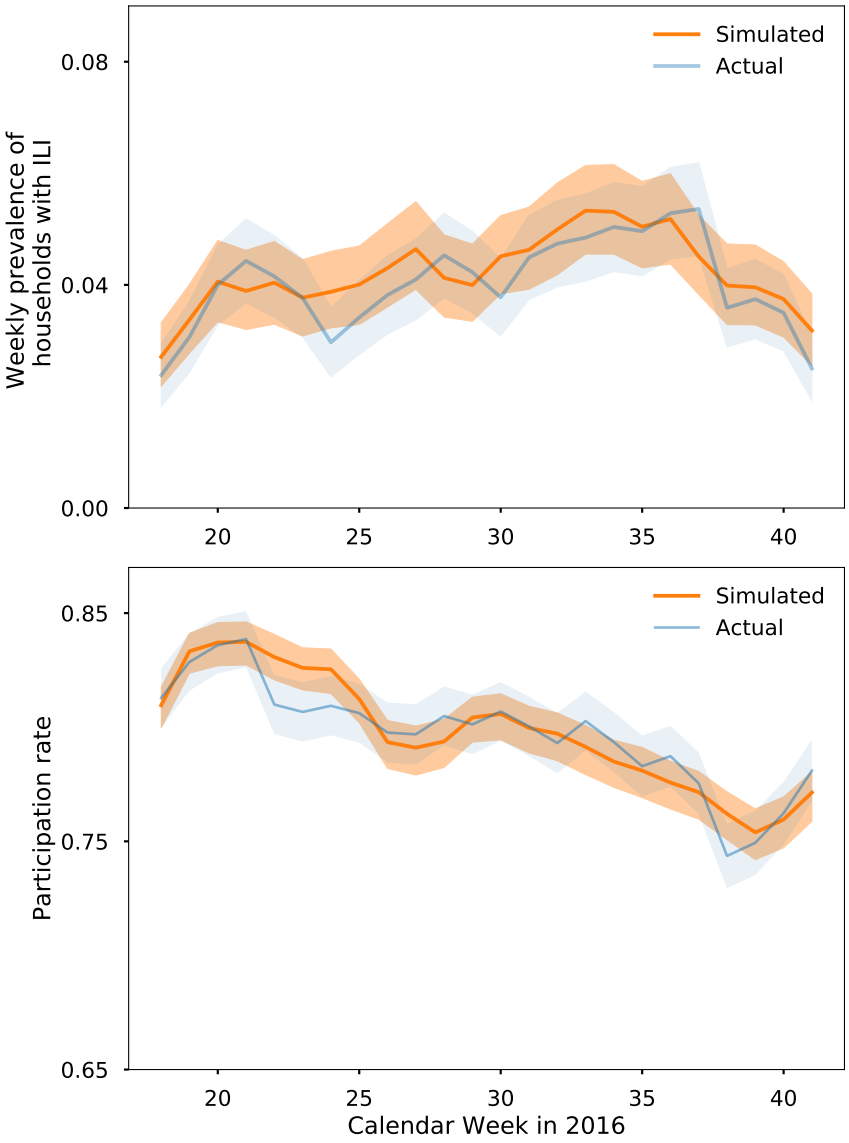

Figure S.16: Cross validation of model predictions with actual outcomes of test set for the 2016 season. Lines represent the median and shaded regions the 95% credible intervals. The model is able to generate the data observed in the test set with high probability.

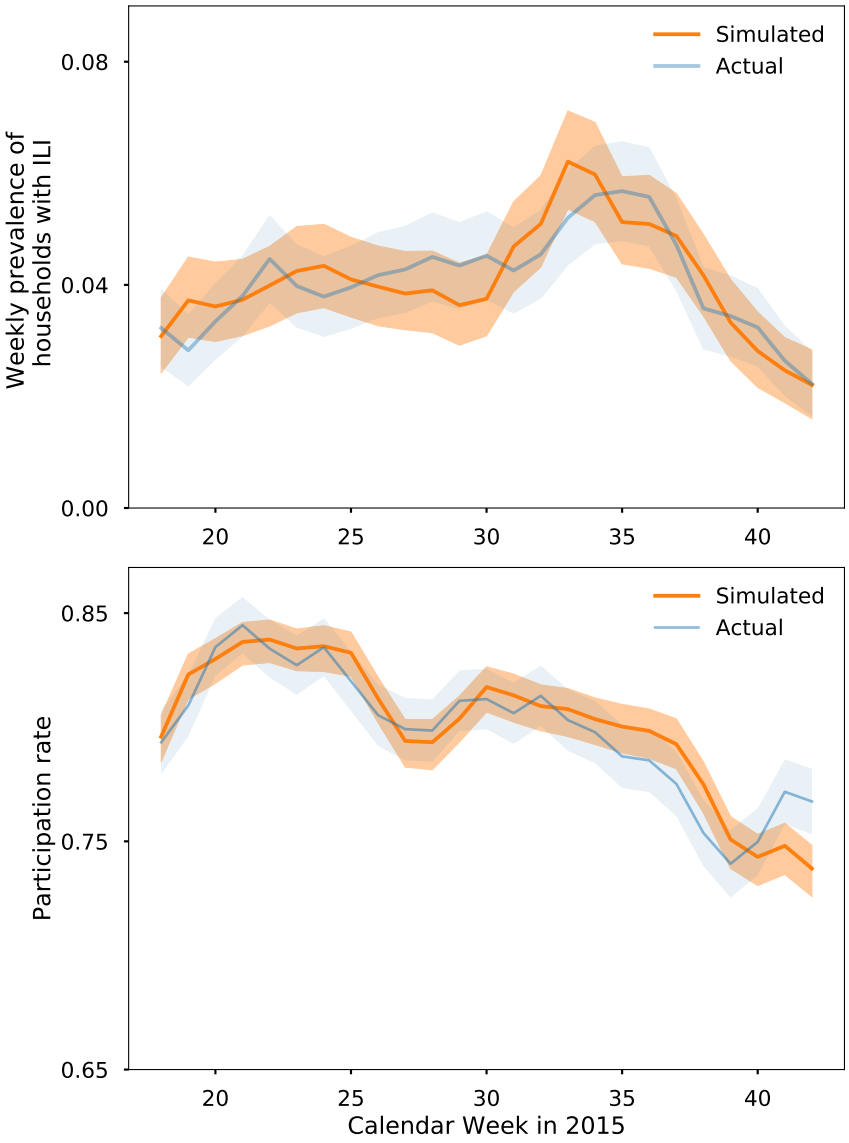

Figure S.17: Cross validation of model predictions with actual outcomes of test set for the 2015 season. Lines represent the median and shaded regions the 95% credible intervals. The model is able to generate the data observed in the test set with high probability.

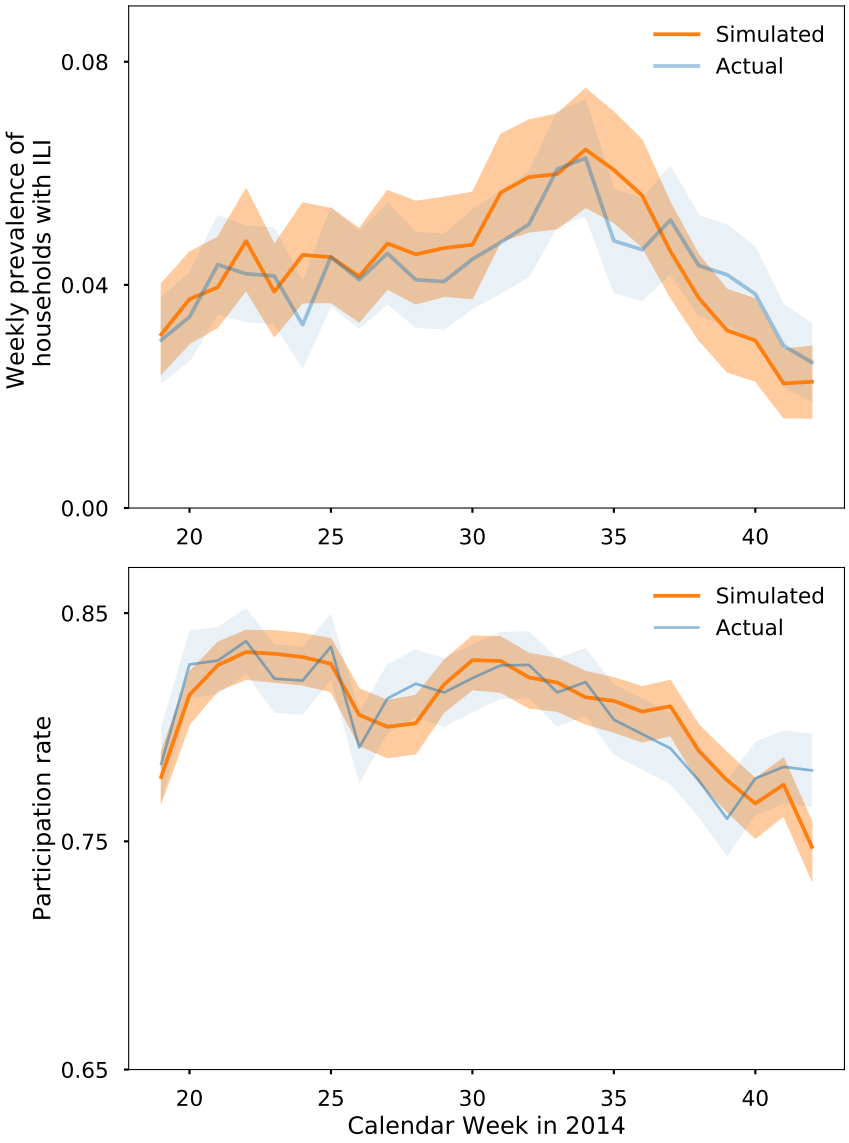

Figure S.18: Cross validation of model predictions with actual outcomes of test set for the 2014 season. Lines represent the median and shaded regions the 95% credible intervals. The model is able to generate the data observed in the test set with high probability.

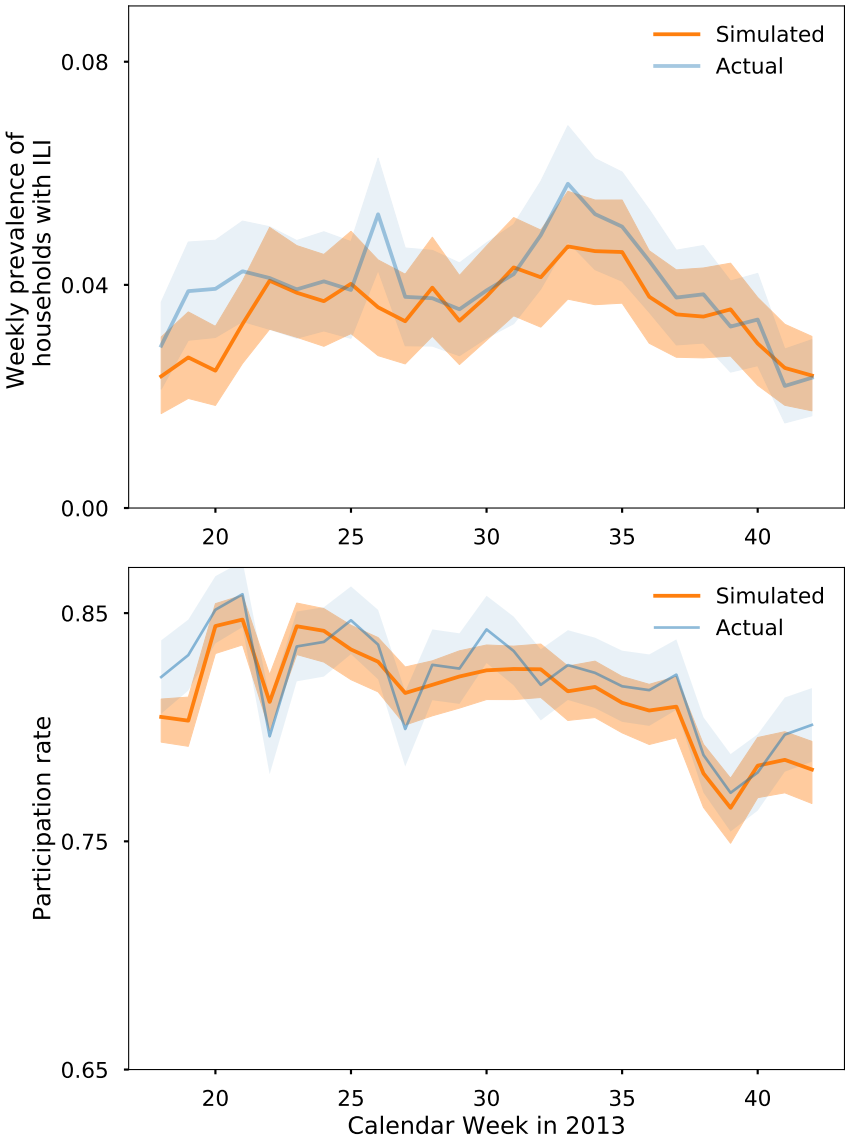

Figure S.19: Cross validation of model predictions with actual outcomes of test set for the 2013 season. Lines represent the median and shaded regions the 95% credible intervals. The model is able to generate the data observed in the test set with high probability.

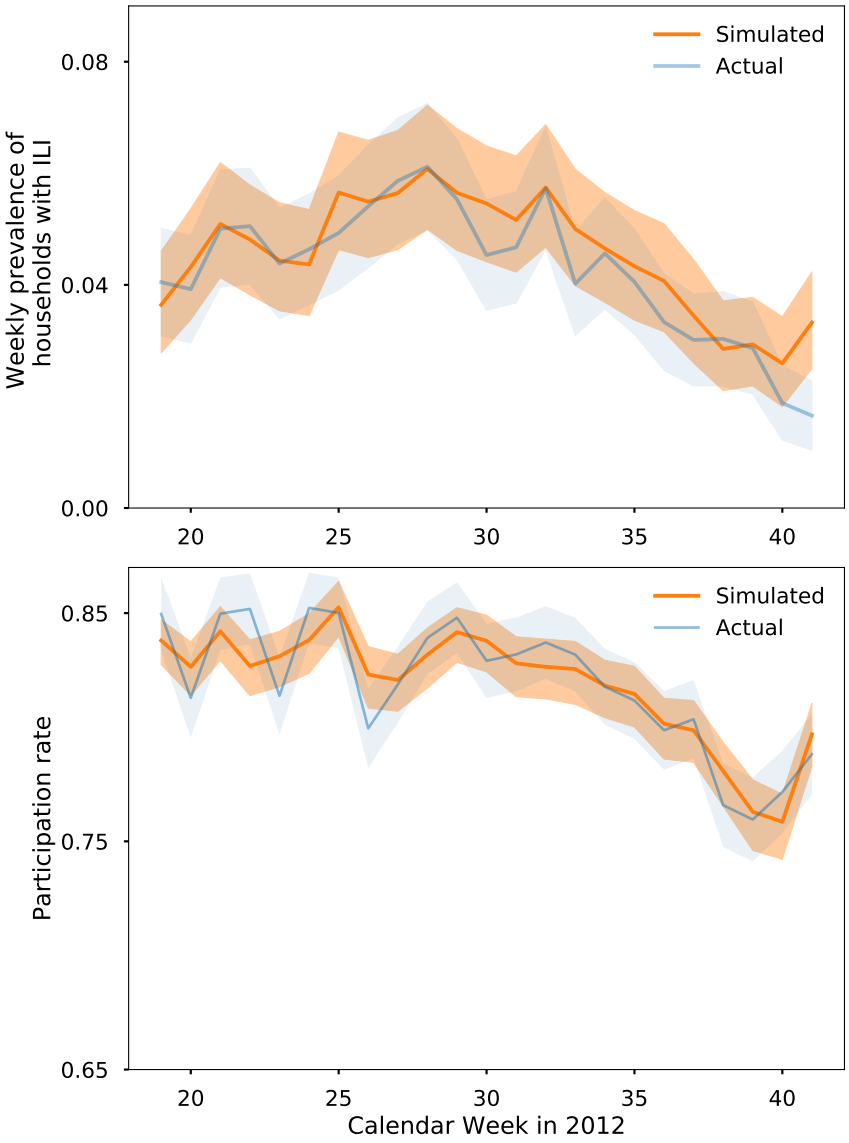

Figure S.20: Cross validation of model predictions with actual outcomes of test set for the 2012 season. Lines represent the median and shaded regions the 95% credible intervals. The model is able to generate the data observed in the test set with high probability.

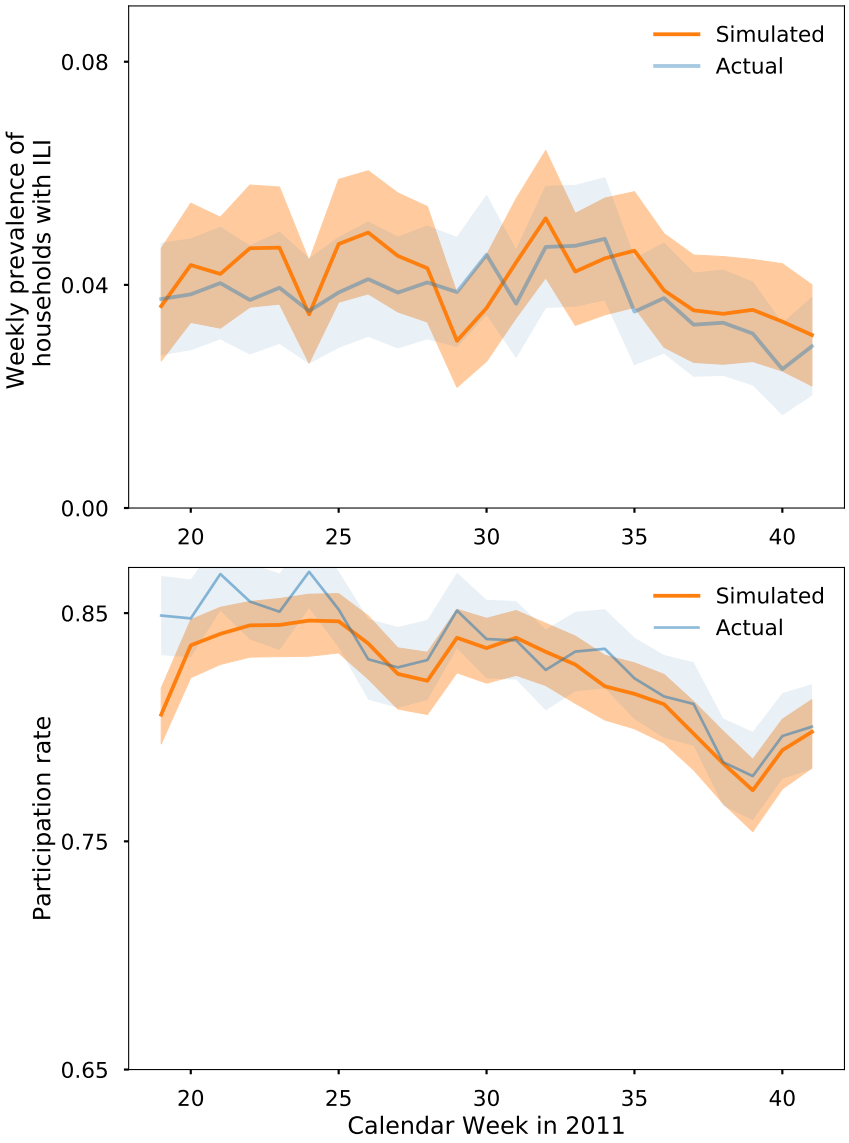

Figure S.21: Cross validation of model predictions with actual outcomes of test set for the 2011 season. Lines represent the median and shaded regions the 95% credible intervals. The model is able to generate the data observed in the test set with high probability.