

1    **Forecasting influenza incidence as an ordinal variable using machine learning**

2    Haowei Wang<sup>1,2</sup>, Kin On Kwok<sup>3,4,5</sup>, Steven Riley<sup>1,2,\*</sup>

3    <sup>1</sup>     School of Public Health, Imperial College London, UK

4    <sup>2</sup>     MRC Centre for Global Infectious Disease Analysis and Abdul Latif Jameel Institute  
5    for Disease and Emergency Analytics, Imperial College London, UK

6    <sup>3</sup> JC School of Public Health and Primary Care, The Chinese University of Hong Kong,  
7    Hong Kong Special administrative regions, China

8    <sup>4</sup>     Stanley Ho Centre for Emerging Infectious Diseases, The Chinese University of  
9    Hong Kong, Hong Kong Special Administrative Region, China

10   <sup>5</sup>     Hong Kong Institute of Asia-Pacific Studies, The Chinese University of Hong Kong,  
11   Hong Kong Special Administrative Region, China

12   \*Corresponding authors: Steven Riley, [s.riley@imperial.ac.uk](mailto:s.riley@imperial.ac.uk), School of Public Health,  
13   Imperial College London, Norfolk Place, London, W2 1PG

14

## 15 Abstract

16 Many mechanisms contribute to the variation in the incidence of influenza disease, such as  
17 strain evolution, the waning of immunity and changes in social mixing. Although machine  
18 learning methods have been developed for forecasting, these methods are used less  
19 commonly in influenza forecasts than statistical and mechanistic models. In this study, we  
20 applied a relatively new machine learning method, Extreme Gradient Boosting (XGBoost), to  
21 ordinal country-level influenza disease data. We developed a machine learning forecasting  
22 framework by adopting the XGBoost algorithm and training it with surveillance data for over  
23 30 countries between 2010 and 2018 from the World Health Organisation's FluID platform.  
24 We then used the model to predict incidence 1- to 4-week ahead. We evaluated the  
25 performance of XGBoost forecast models by comparing them with a null model and a  
26 historical average model using mean-zero error (MZE) and macro-averaged mean absolute  
27 error (mMAE). The XGBoost models were consistently more accurate than the null and  
28 historical models for all forecast time horizons. For 1-week ahead predictions across test  
29 sets, the mMAE of the XGBoost model with an extending training window was reduced by  
30 78% on average compared to the null model. Although the mMAE increased with longer  
31 prediction horizons, XGBoost models showed a 62% reduction in mMAE compared to the  
32 null model for 4-week ahead predictions. Our results highlight the potential utility of machine  
33 learning methods in forecasting infectious disease incidence when that incidence is defined  
34 as an ordinal variable. In particular, the XGBoost model can be easily extended to include  
35 more features, thus capturing complex patterns and improving forecast accuracy. Given that  
36 many natural extreme phenomena, such as floods and earthquakes, are often described on  
37 an ordinal scale when informing planning and response, these results motivate further  
38 investigation of using similar scales for communicating risk from infectious diseases.

39

## 40 **Author Summary**

41 Accurate and timely influenza forecasting is essential to help policymakers improve influenza  
42 preparedness and responses to potential outbreaks and allocate medical resources  
43 effectively. Here, we present a machine learning framework based on Extreme Gradient  
44 Boosting (XBoost) for forecast influenza activity. We used publicly available weekly  
45 influenza-like illness (ILI) incidence data in 32 countries. The predictive performance of the  
46 machine learning framework was evaluated using several accuracy metrics and compared  
47 with baseline models. XGBoost model was shown to be the most accurate prediction  
48 approach, and its accuracy remained stable with increasing prediction time horizons. Our  
49 results suggest that the machine learning framework for forecasting ILI has the potential to  
50 be adopted as a valuable public health tool globally in the future.

## 51 Introduction

52 Influenza forecasting plays a critical role in helping healthcare planners and policymakers to  
53 improve response to large seasonal epidemics and to mitigate their impact in terms of  
54 morbidity and mortality as well as social and economic impacts. Before the COVID-19  
55 pandemic, a contagious respiratory illness caused by influenza viruses posed continual  
56 threats globally, causing an estimated 1 billion cases and up to 650000 deaths annually  
57 [1,2]. Accurate and timely forecasting of influenza epidemics in terms of the start of the  
58 epidemic, the time and size of the peak, and the duration of the epidemic enable  
59 policymakers to take effective interventions and optimise the allocation of healthcare  
60 resources. For example, to conduct the maximum number of elective surgical procedures  
61 prior to opening up space in intensive care wards for community-acquired pneumonia  
62 patients. However, reliable influenza forecasts remain a substantial challenge due to the  
63 variation of dominant influenza strains and environmental factors affecting the outbreak  
64 intensity. [3–5].

65 Different analytical methods have been used as the basis for forecast models of influenza  
66 disease. These methods can be classified into two broad categories: mechanistic models  
67 and statistical methods [5–7]. Mechanistic models attempt to reproduce key features of the  
68 underlying mechanism of transmission. Typical examples include classic compartmental  
69 models [8–11] and agent-based models (ABMs) [12–15]. Statistical approaches are  
70 phenomenological and do not attempt to reproduce the transmission mechanism, such as  
71 autoregressive integrated moving average (ARIMA) [16,17] and Gaussian Process  
72 Regression (GPR) [18,19].

73 Machine learning (ML) models have gained much attention in recent years and have  
74 gradually become the third category of models. Given that machine learning is sometimes

75 differentiated from traditional statistics for its focus on prediction, it is not surprising that  
76 these methods are being applied to forecasting influenza. Traditional statistical models have  
77 a longstanding emphasis on inference, which is achieved through creating and fitting project-  
78 specific probability models. They are often less well-suited to data with large sample sizes  
79 and input variables [20]. By contrast, ML methods are data-driven and avoid making prior  
80 assumptions about underlying correlations and rather employ algorithms to identify patterns  
81 in the data [21,22]. In addition, ML methods are flexible in taking different types of input  
82 variables and a huge number of observations into consideration to improve predictive  
83 performance. The usage of some popular machine learning and deep learning methods in  
84 influenza forecasting has been discussed, such as Long Short Term Memory (LSTM) [23],  
85 Support Vector Machine (SVM) [24–27], and neural networks [28–31]. These methods show  
86 consistent and high forecasting accuracy but also suffer from the risk of over-fitting.

87 Many natural extreme phenomena, such as floods and earthquakes, are often described on  
88 an ordinal scale, but infectious disease incidence is usually described as a count or a  
89 percentage of a population. The use of the moving epidemic method is a notable exception,  
90 which is now routinely used to compare and communicate the current state of epidemics  
91 across nearby connected populations [32]. However, influenza forecasting has remained  
92 focused on non-ordinal target observations [33], limiting the range of analytical methods that  
93 can be applied.

94 Extreme Gradient Boosting (XGboost) is a decision-tree-based ensemble machine learning  
95 method employing the gradient boosting algorithm [34]. It has demonstrated good prediction  
96 performance on a wide range of problems in different industries, including finance, physics  
97 and clinical research (e.g. patient diagnosis) [35–38].

98 In this paper, we aim to construct an XGBoost model to make short-term predictions of the  
99 weekly influenza incidence as an ordinal variable, ranging from 1- to 4-week ahead of the

100 current time. We also evaluate the performance of XGBoost by accuracy metrics and  
101 compare it with baseline models.

## 102 Methods

### 103 *Data*

104 We used a global web-based collection tool for epidemiological indicators and data on  
105 influenza, Flu Informed Decisions (FluID) dataset from the World Health Organization (WHO)  
106 [39]. This platform collects weekly Influenza-Like Illness (ILI) incidence data from WHO  
107 member countries and regions, which is either submitted on a weekly basis or obtained by  
108 the WHO from regional networks such as EUROFlu [2].

109 The FluID data was available from ISO year 2010 week 1 to 2017 week 52 and included  
110 data from 146 countries initially. To ensure sufficient data for model training, countries with  
111 less than 50% of the data were excluded. Further countries that did not have at least 10  
112 weeks of data for each year between 2010 and 2017 were also excluded. After applying  
113 these criteria were applied to the dataset, 32 countries remained, primarily in Europe and  
114 North America.

115 The key field in the data was ILI incidence, which was an integer ranging from 0 to 44965 (in  
116 week 52 of 2017 in the USA). This field was transformed into an ordinal variable by  
117 discretizing into N different bins with equal intervals, using N = 10 as the default. The highest  
118 incidence data for each country was used to add 10% as the upper boundary. The range  
119 from 0 to the upper limit was then divided into ten equal ordinal intervals, each of which was  
120 mapped to an ordinal value from 1 to 10, 1 represented the lowest incidence level, while 10  
121 represented the highest level. Note that each country's classification was based on its own

122 influenza incidence level, so the same category in two different countries corresponded to  
123 different absolute levels of incidence.

#### 124 *Models*

125 We used XGBoost to establish a model to predict short-term weekly influenza incidence  
126 levels [34]. XGBoost is an implementation of the gradient-boosted decision trees and has  
127 been developed to improve computation speed and predictive performance for a variety of  
128 problems, including classification and regression [40]. Gradient boosting is an algorithm  
129 where new models are added in an adaptive way based on the residuals or errors of  
130 predictions from prior models and then combined to make the final prediction. Boosting is an  
131 ensemble method that creates a strong prediction model by iteratively combining a number  
132 of weak classifiers. New weak classifiers are added to correct the errors made by existing  
133 models, and every new model is added sequentially until no further improvements can be  
134 achieved or until a maximum number of models is added. Gradient boosting uses a gradient  
135 descent algorithm to minimize the loss when adding new models, i.e., at every optimization  
136 step, only models that reduce the residual or errors are added. XGBoost uses second-order  
137 Taylor expansion to minimize the loss function and added regularization terms to prevent  
138 overfitting.

139 To further validate the effectiveness of XGBoost for influenza forecasting, we constructed  
140 two additional baseline models for comparison: 1) null model, in which the prediction of the  
141 target week is the same as the most recent available observation week, and 2) historical  
142 average model in which the prediction of the target week is the average of observations of  
143 the same weeks in other years. The usual definition for historical average models is to mean  
144 value over prior observations for that week of the year [41]. Here, we use a slightly different  
145 categorical definition of the incidence level with the highest frequency across the same week  
146 in the other years in the study. The aim of the historical average model is similar to that of

147 the null model, which is to provide a baseline for comparison with the accuracy of the  
148 XGBoost model, reflecting how users might informally forecast the incidence levels for the  
149 next 1 to 4 weeks.

#### 150 *Forecast and features*

151 Our target was to predict ordinal categories for each week at the country level. We first did  
152 regular machine learning forecasting, in which we trained the model with a training set of  
153 fixed durations and then assessed the model's performance on a testing set. We designed  
154 our analysis to use 5 years of training data to predict one year of outcomes. The first training  
155 set used data from 2010 to 2014, the second training set used data from 2011 to 2015, and  
156 so on (S1 Table).

157 For predictors in the short-term forecasting model we used: categorical incidence levels of  
158 the prior  $n$  week and prior  $(n + 1)$  weeks, the month of the year (from January to December)  
159 and the season (spring, summer, autumn, and winter), where  $n$  represents the  $n$ -week  
160 ahead forecast. In a 1-week ahead forecast, the model uses the incidence levels of the prior  
161 1 week and the prior 2 weeks; similarly, the incidence levels of the prior 4 weeks and prior 5  
162 weeks will be used as predictors in the 4-week ahead prediction.

163 In addition to the fixed training window approach, we also trained the XGboost algorithm with  
164 an extending window. In the fixed window approach, we train the XGBoost model only once  
165 with the fixed training set and then predict the test set. Since the data were collected on a  
166 weekly basis, we are able to update our model by including the new coming data and  
167 retraining it to see if model performance can be improved. Thus we train the model with an  
168 extended training set for each week, which is called extending window approach in this  
169 paper. For example, if we are going to predict the week  $i$  ( $i > 2$ ) of the year 2015, our  
170 training set used data from 2010 until 2014 for all weeks and data from 2015 for weeks 1 to  
171 week  $(i - 1)$ . Our test data were data in 2015 week  $i$ . The baseline models were trained with



172 the fixed window approach only. Weeks with missing values for predictors were removed  
173 from both the training set and the test set.

174 XGBoost provides a large number of hyperparameters to help achieve optimal performance.  
175 They are mainly divided into three types, general parameters, booster parameters, and  
176 learning task parameters [42]. Booster parameters are closely related to the performance of  
177 the model, which is the focus of hyperparameter tuning. Grid search was used to final optimal  
178 values for hyperparameters max\_depth, min\_child\_weight, subsample, colsample\_bytree,  
179 learning rate and gamma. Grid search can be challenging and time-consuming due to the many  
180 parameters to optimize, even with XGBoost's rapid convergence. Our grid search for  
181 hyperparameters was carried out as below:

- 182 1) Firstly, find the optimal gamma and eta (learning rate) at the same time since they have  
183 an impact on the performance of the model. The values searched for gamma are 0.1,  
184 0.2, 0.5, 1, 1.5, 2, and 10, while those for eta are 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, and 0.3.  
185 We ran all possible combinations of these two hyperparameter values to tune, and the  
186 one with the best performance was selected as the optimal value for gamma and eta,  
187 respectively.
- 188 2) With the optimal values of gamma and eta obtained in the previous step, a grid search  
189 was conducted for max\_depth, and min\_child\_weight range from 0.1 to 1.
- 190 3) Made a grid-search over subsample and colsample\_bytree simultaneously range from  
191 0.1 to 1.

192 K-fold cross-validation is used during the tuning process to assess the model's performance with  
193 different combinations of hyperparameter values. Since there is a dependency between weekly  
194 ILLI incidence and future values that cannot be used to forecast past values, the traditional K-fold  
195 method that randomly splits the training set into K folds is not applicable to our data. Instead, we  
196 used cross-validation on an extending basis (S1 Fig). We selected data from 2010 to 2014 as  
197 the overall training set for the cross-validation, but during the process of rolling cross-validation,

198 we started with data from 2010 as the first fold of the training set and checked the accuracy of  
199 prediction for the data from 2011. Data from 2010 and 2011 then formed the second fold of the  
200 training set, and data from 2012 became the test for accuracy check. The accuracy metric used  
201 in the tuning process is macro-averaged mean absolute error (mMAE) which will be defined in  
202 detail later. This process was repeated until 2014 became the last test set. The final optimal  
203 values of the hyperparameter were given in S2 Table.

#### 204 *Accuracy metrics*

205 We used two of the most commonly used metrics in ordinal classification problems [43],  
206 macro-averaged mean absolute error (mMAE) and mean zero-one error (MZE) ) to evaluate  
207 the model performance. Both metrics are defined as negatively oriented penalties that we  
208 aim to minimize: the lower the score, the better the forecast is considered.

#### 209 Macro-averaged mean absolute error

210 Macro-averaged mean absolute error (mMAE) measure is adapted from the traditional class-  
211 based metric mean absolute error (MAE) which assesses the average deviation of the  
212 predicted class from the actual class, and it is defined as

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i^* - y_i|$$

213  
214 where  $y_i$  and  $y_i^*$  denotes the true class and the predicted class respectively. However, MAE  
215 averages effectiveness across individual observations, which does not reflect the  
216 imbalanced distribution of classes in our datasets if many observations are in the lowest  
217 class and very few are in higher classes. Instead, we used macro-averaged mean absolute  
218 error (mMAE) as one of our metrics to assess the performance of our models in terms of  
219 predicting ordered outcomes. mMAE is obtained by first computing the MAE on a per-class

220 basis and then averaging the results across the classes so that mMAE is insensitive to class  
221 imbalance. Let  $y_i$  be the true class and  $y_i^*$  be the predicted class of  $i$ -th week in the test set.  
222 Let  $N_k$  be the number of true cases with the class  $k$  where  $k \in \{1, 2, \dots, 9, 10\}$ . There are 10  
223 classes in our classification problem, i.e.,  $N_k = \sum_{i=1}^N I_{y_i=k}$  and  $N = \sum_{k=1}^{10} N_k$ . The  
224 Macro-averaged mean absolute error is calculated with the following formula:

$$mMAE = \frac{1}{10} \sum_{k=1}^{10} \frac{1}{N_k} \sum_{i=1}^N |y_i^* - y_i| I_{y_i=k}$$

225

226 , where

$$I_{y_i=k} = \begin{cases} 1, & \text{if } y_i = k \\ 0, & \text{if } y_i \neq k \end{cases}$$

227

## 228 Mean zero-one error

229 Our second metric is the mean zero-one error (MZE) which is more frequently known as the  
230 error rate of classifiers and is, essentially, the proportion of time that a correct prediction is  
231 made

$$\begin{aligned} MZE &= \frac{1}{N} \sum_{i=1}^N I_{y_i^* \neq y_i} \\ &= 1 - Acc \end{aligned}$$

232

233 where

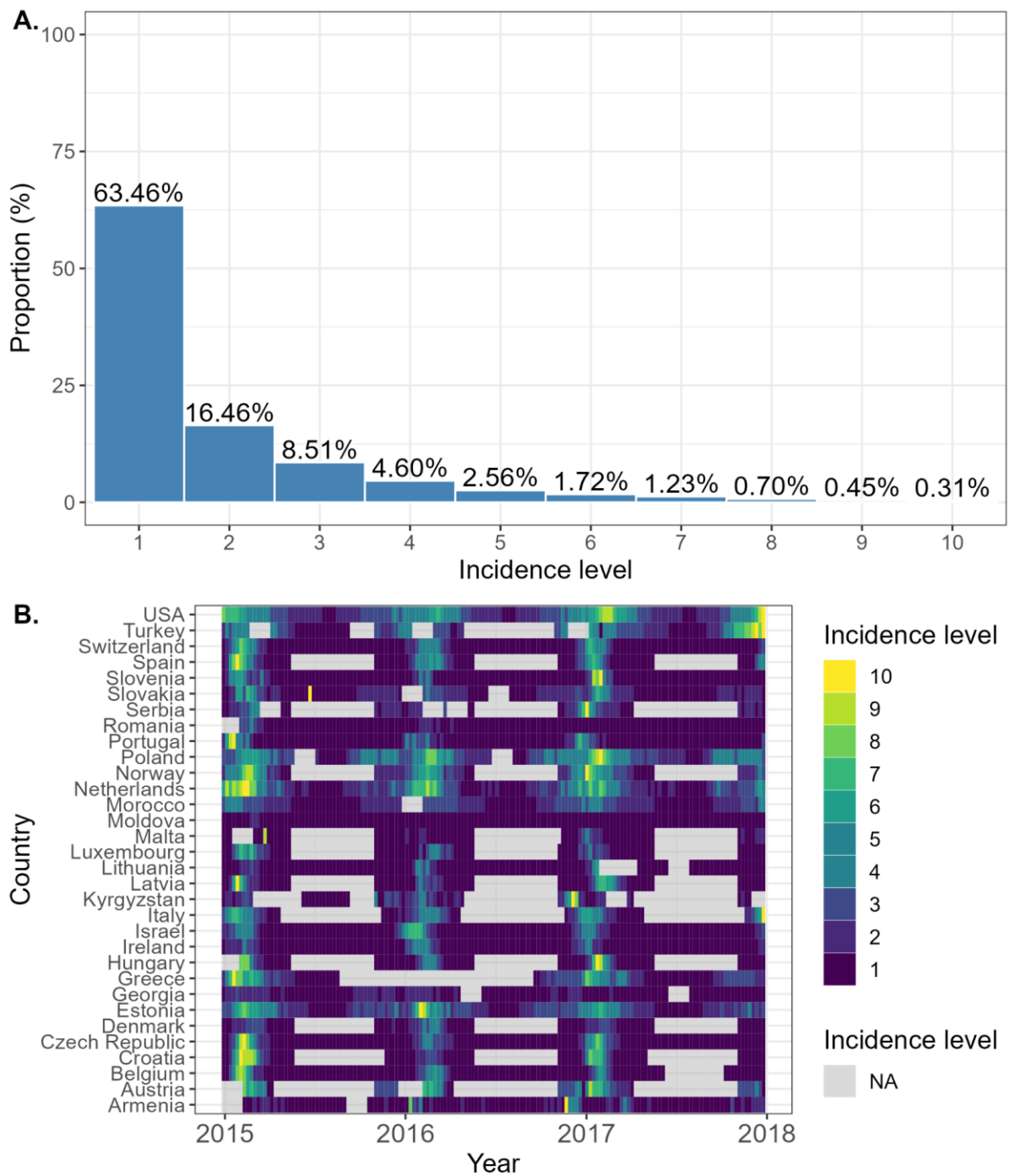
$$I_{y_i^* \neq y_i} = \begin{cases} 1, & \text{if } y_i^* \neq y_i \\ 0, & \text{if } y_i^* = y_i \end{cases}$$

234

235  $y_i$  and  $y_i^*$  denotes the true class and the predicted class respectively. *Acc* is the accuracy of  
236 the model, i.e., the number of correctly predicted categories divided by the total number of  
237 predictions. Therefore, MZE ranges between 0 and 1, and the lower the MZE, the higher  
238 accuracy indicates better predictive performance.

## 239 Results

240 Country-level forecasts were performed using ILI data from 32 countries, located in northern  
241 temperate regions (S2 Fig). Due to the incompleteness of data, countries located outside of  
242 temperate regions were excluded from the analysis. The countries in the northern  
243 hemisphere's temperate regions demonstrated a similar trend in influenza activity, with ILI  
244 incidence typically rising at the end of the year and reaching a peak level at the beginning of  
245 the following year (Fig 1B). During the 2010 - 2017 surveillance, after excluding weeks with  
246 missing values, the data availability ranged between 229 and 417 weeks, with each country  
247 contributing an average of 337 weeks of data (median = 391 weeks). Of these weeks, over  
248 60% fell into level 1, while 0.76% of the total weeks fell into levels 9 and 10 (Fig 1A).

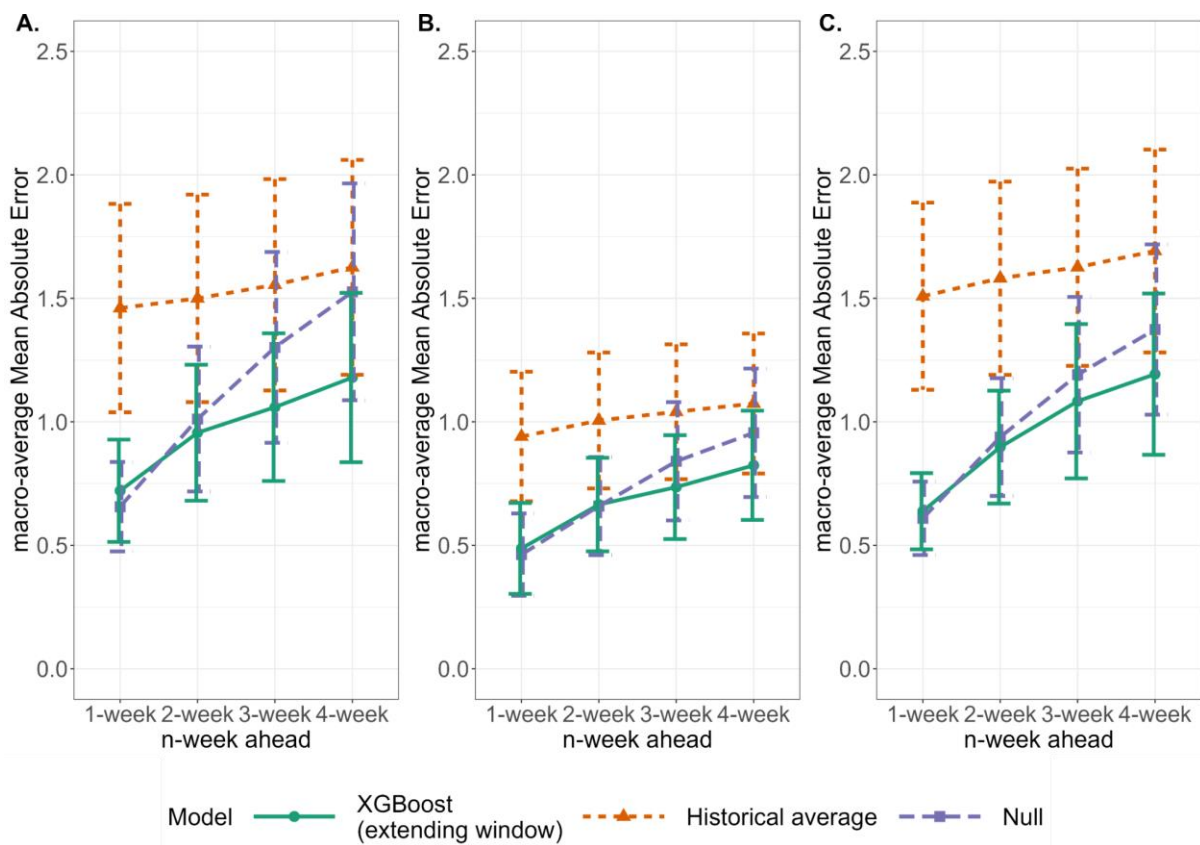


249

250 **Fig 1. Distribution of categorical incidence levels for weekly influenza-like illness for**  
 251 **countries included in the study. A.** overall frequency of every level appearing in 32  
 252 countries. **B.** The heat map shows of incidence level in each country by time. Grey-shaded  
 253 areas indicate that no available were data for those weeks.

254 *Predictive performance*

255 Overall, the XGBoost model with an extended window generated fewer errors and  
256 demonstrated more stable predictive performance than the XGBoost model with a fixed  
257 training window or the baseline models. The short-term prediction errors of different models  
258 are summarized in Table 1, by averaging the mMAE is averaged across countries, and  
259 presented for each prediction horizon (1-week ahead, 2-week ahead, etc. ). We separately  
260 compare the mMAEs of three test periods (2015 to 2017). From Table 1, for these three test  
261 periods, we show that prediction accuracy measured by mMAEs decreases with the  
262 increasing forecast length (i.e. one to four-week in advance), but both XGBoost models (with  
263 extending window and fixed window) uniformly outperform the two baseline models (Fig 2,  
264 Table 1). In particular, the mMAEs of the XGBoost model with an extended window remain  
265 below 1, on average across all countries, even as the forecast length increases, indicating  
266 that, on average, the predicted classes are either correct or only one class away from the  
267 true class. This XGBoost model has the lowest average MZEs for all prediction horizons  
268 compared to the XGBoost model with a fixed window and baseline models (S3 Fig, S3  
269 Table).



270

271 **Fig 2. Differences in macro-averaged mean absolute error (mMAE) by model.**

272 Comparison of mMAE for XGBoost (with an extended window) and baseline models while  
273 forecasting 1 to 4 weeks ahead. **A.** 2015, **B.** 2016, **C.** 2017.

274

275

276

277

278

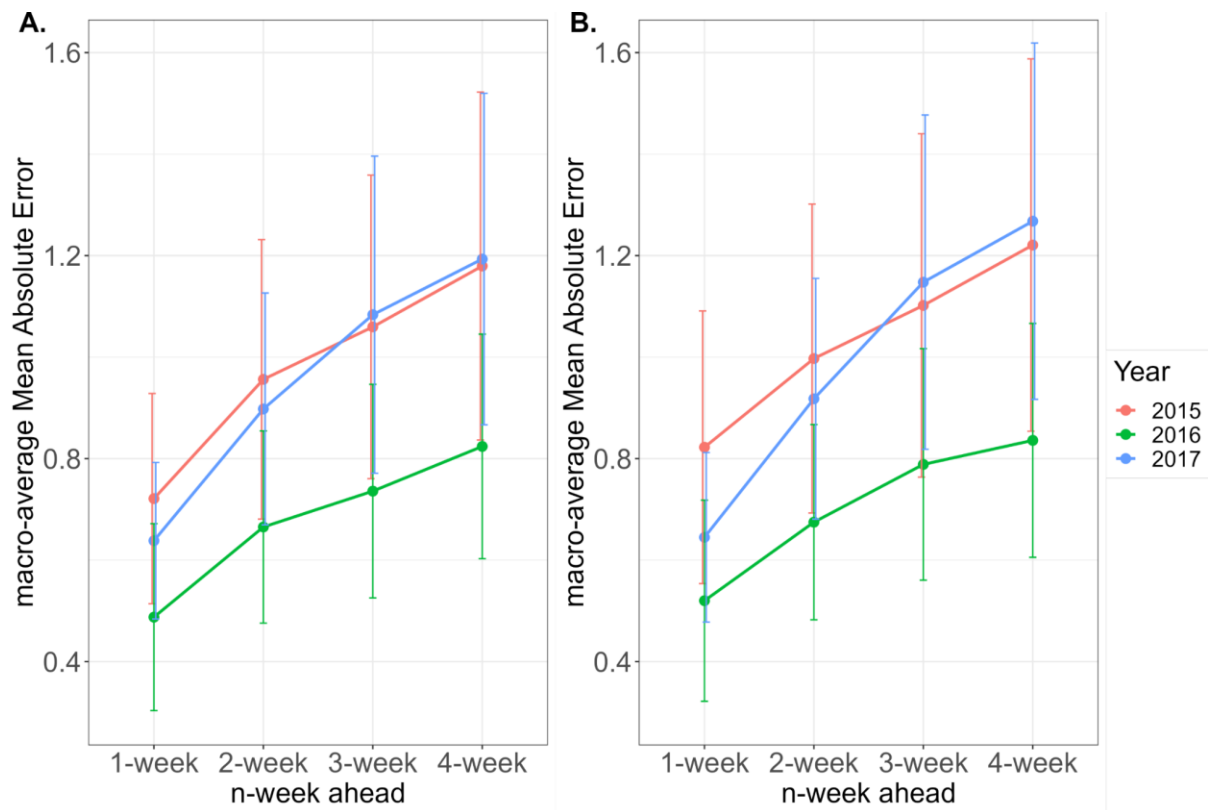
279

280 **Table 1. Overall macro-averaged mean absolute error (mMAE) of XGboost compared**  
 281 **with baseline models.** Overall mMAEs of 1- to 4-week ahead forecasts for each model are  
 282 calculated as the average of 32 countries' mMAEs by year.

| Year | Models                        | mMAE of n-week ahead |       |       |       |
|------|-------------------------------|----------------------|-------|-------|-------|
|      |                               | n = 1                | n = 2 | n = 3 | n = 4 |
| 2015 | XGBoost with extending window | 0.721                | 0.956 | 1.060 | 1.179 |
|      | XGBoost with fixed window     | 0.822                | 0.997 | 1.102 | 1.221 |
|      | Historical average model      | 3.316                | 3.449 | 3.554 | 3.705 |
|      | Null model                    | 2.908                | 1.848 | 2.447 | 2.908 |
| 2016 | XGBoost with extending window | 0.488                | 0.665 | 0.736 | 0.824 |
|      | XGBoost with fixed window     | 0.520                | 0.675 | 0.789 | 0.836 |
|      | Historical average model      | 3.338                | 3.200 | 3.270 | 3.247 |
|      | Null model                    | 2.751                | 1.826 | 2.429 | 2.751 |
| 2017 | XGBoost with extending window | 0.638                | 0.898 | 1.084 | 1.193 |
|      | XGBoost with fixed window     | 0.645                | 0.918 | 1.148 | 1.268 |
|      | Historical average model      | 3.484                | 3.604 | 3.734 | 3.878 |
|      | Null model                    | 2.663                | 1.720 | 2.221 | 2.663 |

283  
 284 The mMAEs for the extending window approach were consistently lower compared to the  
 285 XGBoost model with a fixed window (Fig 3, Table 1). The mMAEs for the fixed window  
 286 approach only remained below 1 for 1- to 4-week forecasts in 2016, but exceeded 1 for 3-  
 287 and 4-week ahead forecasts in 2015 and 2017. Although the two XGBoost models had  
 288 similar MZEs, the extending window approach still resulted in a lower MZE for each  
 289 prediction horizon (S4 Fig, S3 Table). The better performance of the extending window  
 290 approach can be attributed to its continuous addition of the predicted weeks' observations to  
 291 the training set, allowing the model to learn more data, resulting in smaller prediction errors.

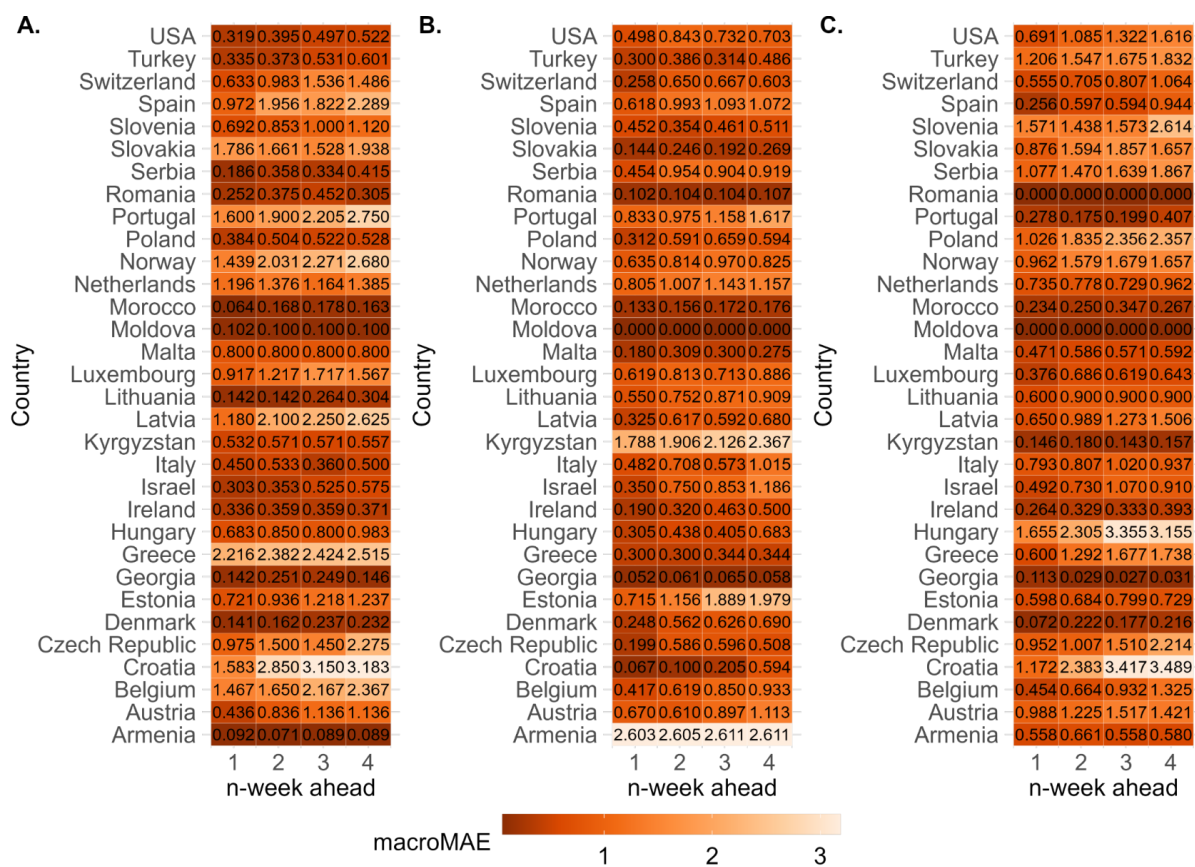




292

293 **Fig 3. Macro-averaged mean absolute error (mMAE) of XGBoost models.** Comparison  
294 of mMAEs for weekly forecasting ranging from 1-to 4-week ahead for the years 2015 to  
295 2017. **A.** The XGBoost model with an extended window approach. **B.** The XGBoost model  
296 with a fixed window approach.

297 There is substantial variation in the prediction accuracy, as indicated by mMAE, among the  
298 countries (Fig 4, S4 – S6 Table). For instance, Moldova, had remarkably low mMAEs from  
299 2015 to 2017, even achieving 0 in 2016 and 2017. Conversely, certain countries, such as  
300 Hungary and Norway, consistently exhibit much higher mMAEs compared to other countries.  
301 Similar results were found for MZEs (S5 Fig, S7 – S9 Table). The performance of XGBoost  
302 on the test set can be affected mainly by the quality of training data. By checking the missing  
303 data for countries, we discover that countries with high mMAEs often have more incomplete  
304 data than better-performing countries, implying that they have a smaller sample size of  
305 training sets, failing to capture complex epidemic patterns of influenza (Fig 2B).



306

307 **Fig 4. Macro-averaged mean absolute error (mMAE) for 32 countries.** mMAEs of

308 predictions by the XGBoost model with an extended window for the year **A.** 2015. **B.** 2016.

309 and **C.** 2017.

310 We investigated the influenza distribution and forecasts over a 3-year period in four

311 countries selected based on their mMAEs and data completeness and found different

312 qualitative drivers of accuracy. The plots in Fig 5 compare the predicted values from four

313 models with actual data for 1 to 4-week-ahead forecasts in Moldova, Switzerland, Estonia

314 and Hungary separately. Moldova had the lowest mMAE, but its influenza activity barely

315 fluctuated with 97.5% (390 out of 400 weeks) of the weeks remaining at level 1. Conversely,

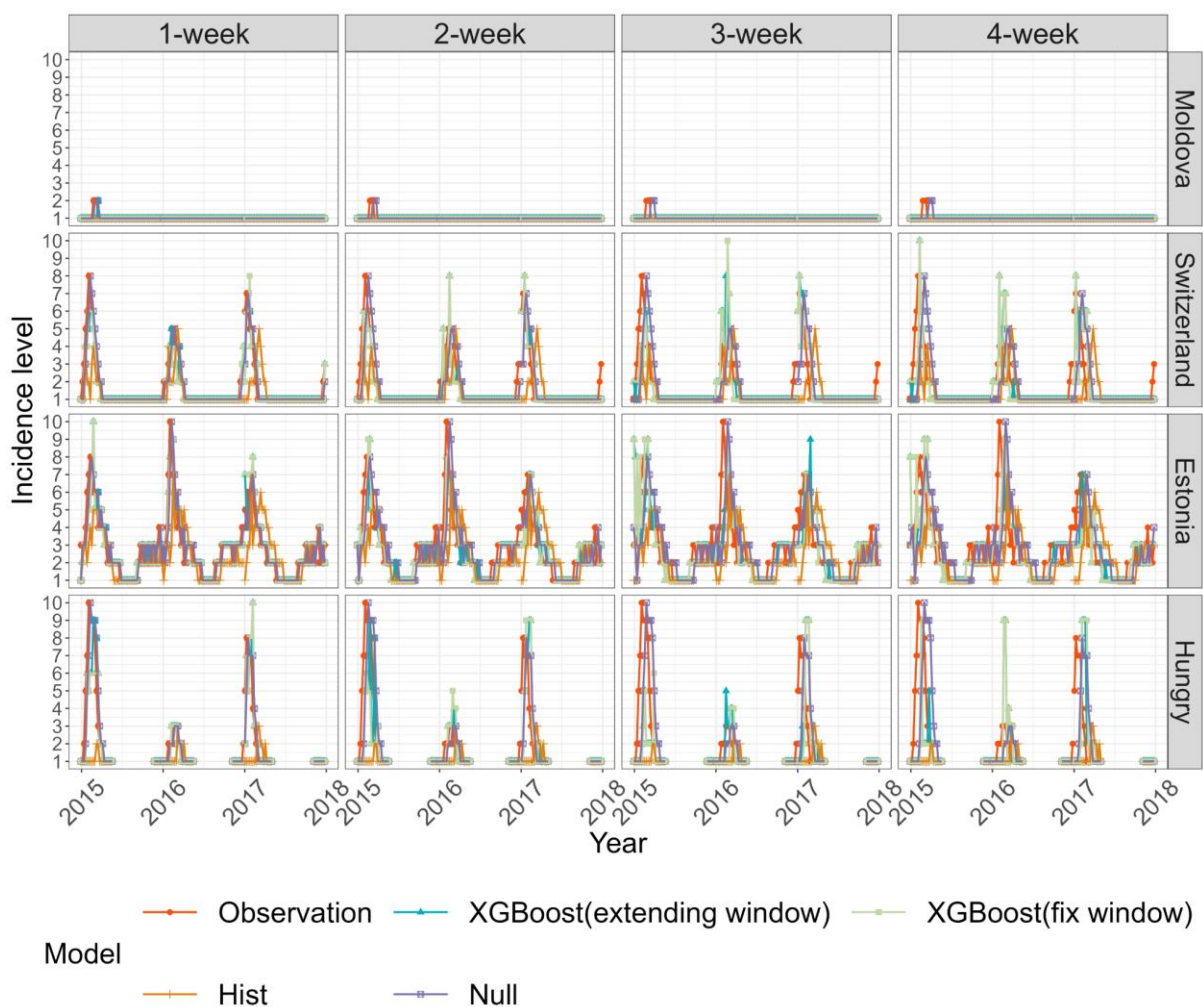
316 Switzerland and Estonia had 25.3% (82 out of 406 weeks) and 72.5% (290 out of 400

317 weeks) of the weeks, respectively, above the baseline level. All four models successfully

318 predict peak influenza season for 1-week and 2-week ahead forecasts in 2015 and 2016 in

319 Switzerland Estonia, but the historical average model failed to identify peaks for 3-week and  
 320 4-week ahead forecasts. The other three models could identify peaks up to 4-week ahead of  
 321 forecasts, with the XGboost with extending approach being the most accurate. The deviation  
 322 between actual observation and prediction is typically high when there is a sudden increase  
 323 or decrease in flu incidence levels. In Hungary, due to the limited sample size of the training  
 324 set and the complexities of influenza activity, a large deviation between actual observation  
 325 and prediction even for the 1-week forward forecast. The prediction of XGboost with an  
 326 extending window provided the closest prediction to the observed peak (Fig 6).

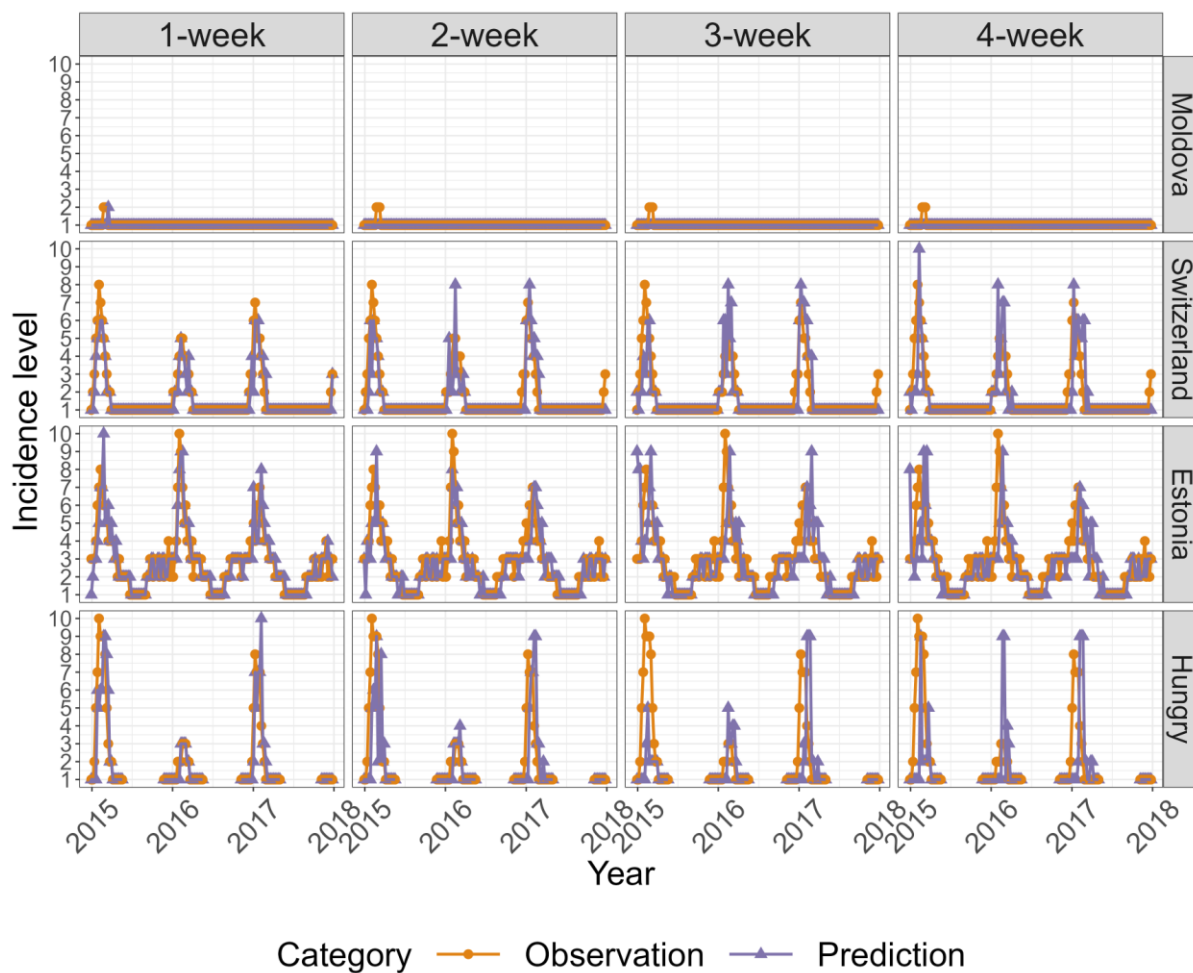
327



328

329 **Fig 5. Comparison of Actual and predicted incidence by country.** Countries Moldova,  
330 Switzerland, Estonia and Hungary are selected. Models include the two XGBoost models,  
331 the historical average model and the null model while forecasting 1- to 4-week ahead for the  
332 test period year 2015 to 2017.

333



334

335 **Fig 6. Comparison of actual and predicted incidence.** Predicted incidences are  
336 generated by the XGBoost model with an extended window for Moldova, Switzerland,  
337 Estonia and Hungry; for forecasts 1 to 4 weeks ahead for the test period year 2015 to 2017.

## 338 Discussion

339 In this work, we have proposed a potential machine learning model for influenza forecasting  
340 using a 10-unit ordinal variable to describe the case incidence. We evaluated two XGBoost  
341 machine learning models, one with a fixed window training set and the other with an  
342 extended window training set, and compared them to baseline models. We measured their  
343 accuracy using the mean zero-one error (MZE) and macro-averaged mean absolute error  
344 (mMAE). Our results showed that both XGBoost models outperformed the baseline models,  
345 with the extended window XGBoost model consistently achieving the highest accuracy.

346 This framework is novel because we defined the outcome on an ordinal scale, which is often  
347 how influenza incidence is communicated[32]. However, our results are not directly  
348 comparable to other forecasting studies, and the XGBoost model has not been validated  
349 against other data sources. This highlights the need to test its applicability in real-world  
350 public health practice.

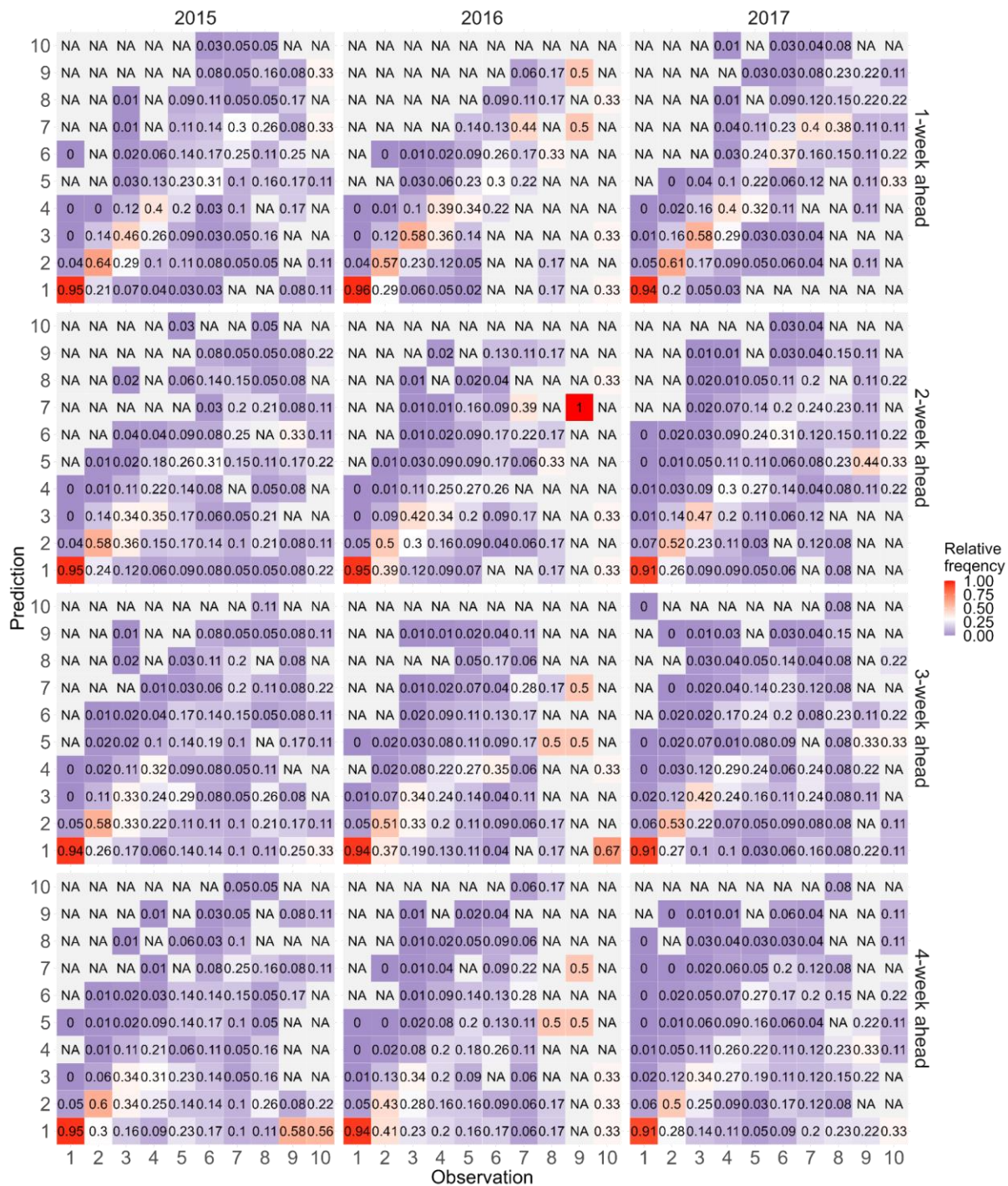
351 The use of a categorical outcome enabled us to apply machine learning algorithms. Although  
352 the XGBoost has advantages, it has not been tested against other data sources. Machine  
353 learning algorithms' performance can be limited by the quality of the training and testing  
354 data, and inconsistencies between different databases may reduce the generalisability of the  
355 XGBoost prediction model [44]. Further testing is needed to determine the applicability of the  
356 in real-world public health practice.

357 Our baseline models provide a basic framework for assessing the potential of machine  
358 learning forecasts. The baseline models are intended to reflect the most basic of implicit  
359 forecast assumptions: they assume that incidence will remain unchanged and revert to its  
360 historical average. Our methods offer improvements over these assumptions, for example,  
361 the 4-week forecast horizon XGBoost model with an extended window being more accurate

362 than assuming the incidence would follow the historical average and more accurate than the  
363 default 1-week ahead model with no change assumption.

364 This work has several limitations. One of the challenges in working with infectious disease  
365 datasets is the limited period of data available for model estimation and evaluation,  
366 compared to other datasets used in machine learning. For each country in the study, only 5  
367 years of data were used for model training, leaving only 3 years for testing. Cross-validation  
368 was used to find the optimal hyperparameters instead of using an independent validation  
369 set. Additionally, countries with fewer weeks of data in the training set tend to have higher  
370 prediction errors, which suggests that accuracy will improve as more data becomes  
371 available. Furthermore, the results should be interpreted with caution, given the disruption to  
372 influenza transmission during the COVID-19 pandemic, as patterns may change as the  
373 transmission is re-established.

374 Our forecasts showed reduced accuracy in predicting turning points of the epidemic (Fig 7).  
375 When there is a sudden increase in incidence levels, they are often underestimated. On the  
376 other hand, a sudden decrease in incidence levels results in them being easily  
377 overestimated. This is especially important in a public health context where accurately  
378 predicting rare cases of much higher than usual incidence levels or earlier or later ends of  
379 the epidemic is crucial. Thus, improving the prediction stability at the extreme points of  
380 season patterns remains a priority in ongoing forecasting work.



381

382 **Fig 7. Distribution of predictions against each decile of observation for the XGBoost**  
 383 **model with an extended window.** Rows of heat maps are ordered from 1 week ahead (top)  
 384 to 4 weeks ahead (bottom). Columns of heat maps are ordered from 2015 (left) to 2017

385 (right). Values on the diagonal are the accuracy of forecasts, while other values represent  
386 the frequency at each level that was incorrectly predicted to be the other level.

## 387 Conclusion

388 Forecasting influenza as an ordinal outcome is a feasible task for machine learning. The  
389 widely used XGBoost model, even with a limited set of features, provides significantly more  
390 accurate predictions than the standard baseline models. With datasets that have longer  
391 history and comprehensive spatial coverage, it is possible to achieve more accurate  
392 forecasts. Similar to other epidemiological models, the framework can easily be expanded to  
393 include population serology and population mobility [45] or other relevant features as more  
394 data become available [46].



395 **Supporting information**

396 **S1 Table. Time ranges of training and test sets.**

397 (XLSX)

398 **S2 Table. Summary of optimal values of hyperparameters for XGBoost models.**

399 (XLSX)

400 **S3 Table. Overall mean-zero error (MZE) of XGboost compared with baseline models.**

401 Overall MZEs of 1- to 4-week ahead forecasts for each model are calculated as the average  
402 of 32 countries' MZEs by year.

403 (XLSX)

404 **S4 Table. Macro-average mean absolute error (mMAE) for the prediction by country in**  
405 **2015.**

406 (XLSX)

407 **S5 Table. Macro-average mean absolute error (mMAE) for the prediction by country**  
408 **in 2016.**

409 (XLSX)

410 **S6 Table. Macro-average mean absolute error (mMAE) for the prediction by country**  
411 **in 2017.**

412 (XLSX)

413 **S7 Table. Mean-zero error (MZE) for the prediction by country in 2015.**

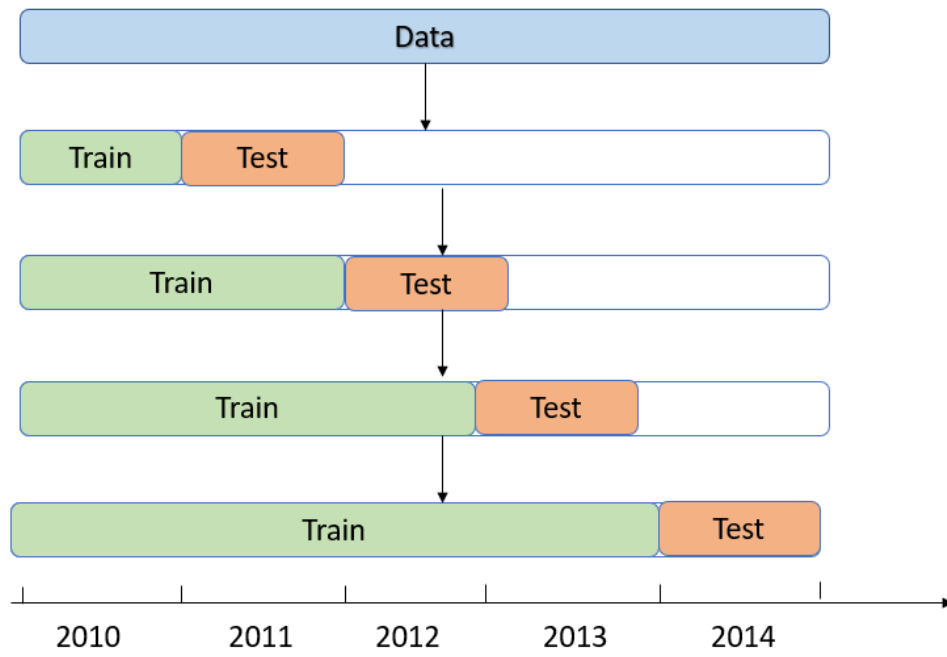
414 (XLSX)

415 **S8 Table. Mean-zero error (MZE) for the prediction by country in 2016.**

416 (XLSX)

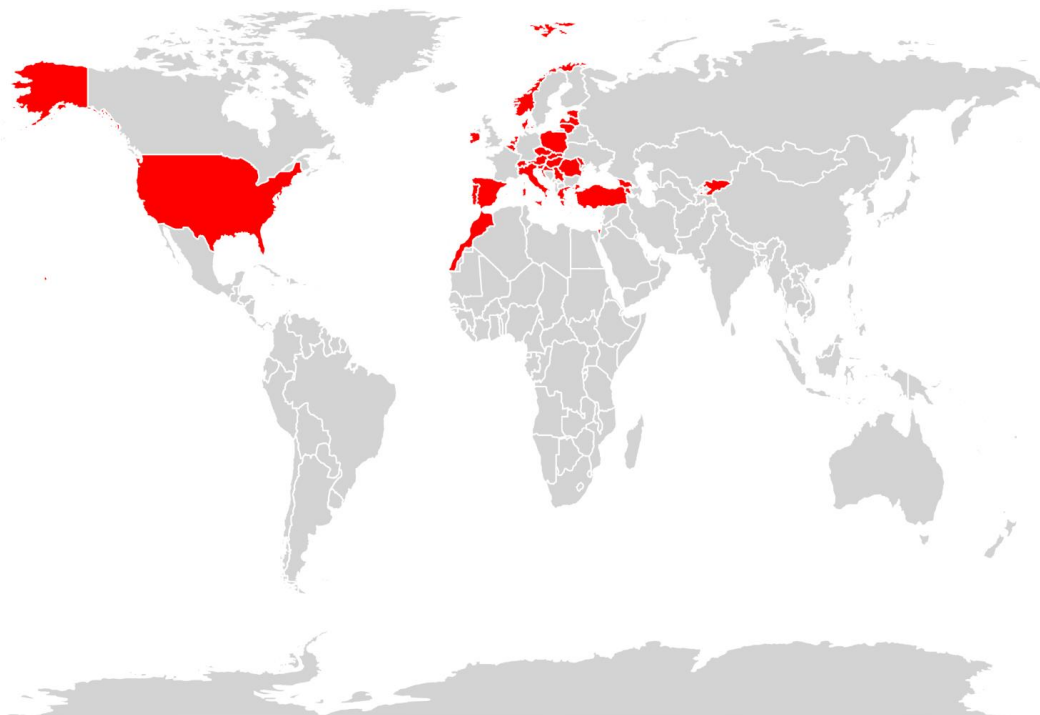
417 **S9 Table. Mean-zero error (MZE) for the prediction by country in 2017.**

418 (XLSX)



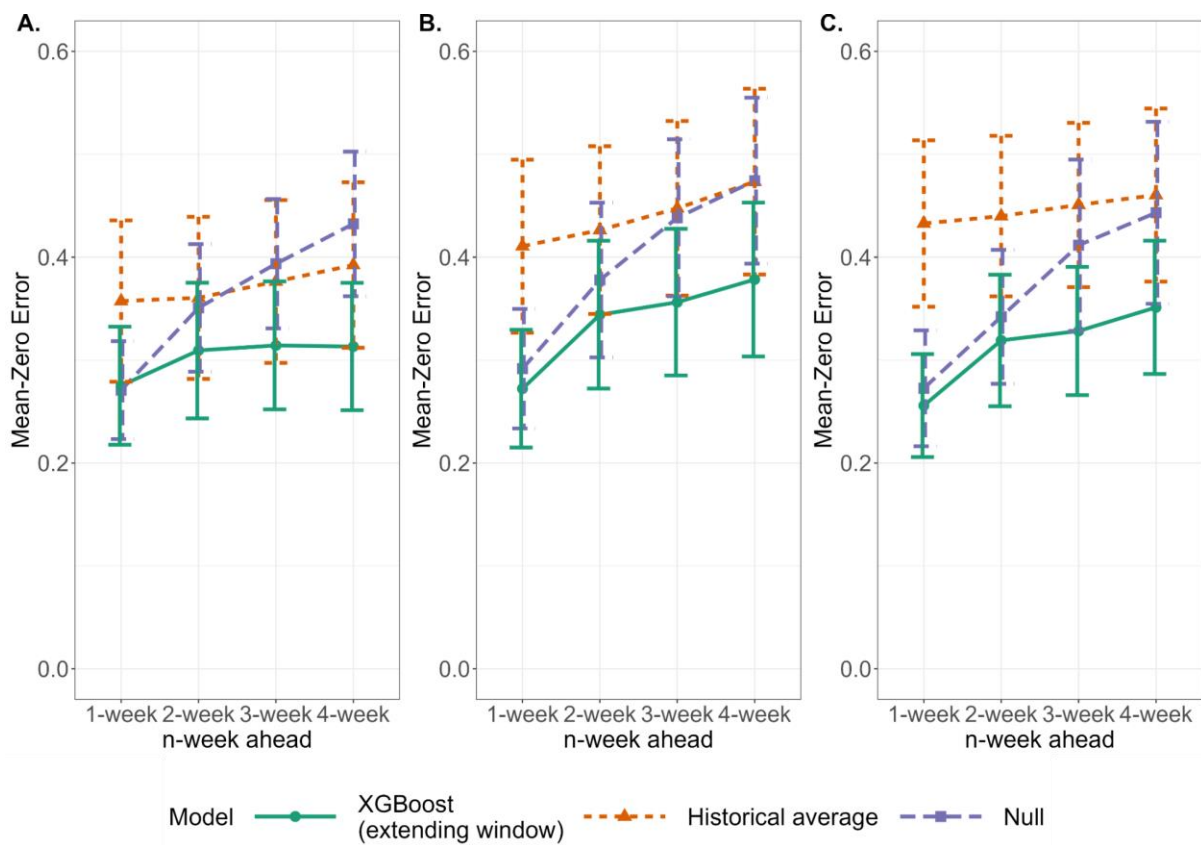
419

420 **S1 Fig. Time series split cross-validation.**



421

422 **S2 Fig. World map.** 32 Countries included in this study are marked in red. Samples are  
423 mainly from North America and Europe.



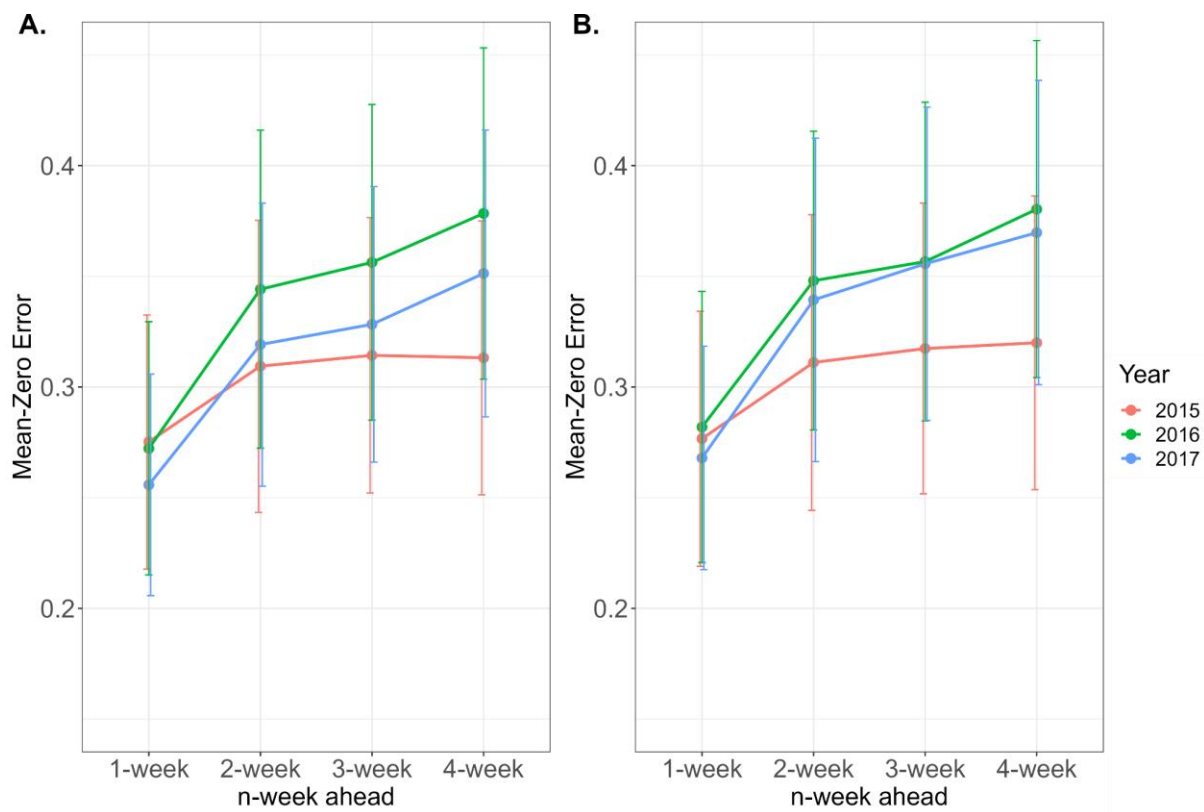
424

425 **S3 Fig. Differences in mean-zero error (MZE) by model.** Comparison of mean-zero error

426 (MZE) for XGBoost (with an extended window approach) and baseline models while

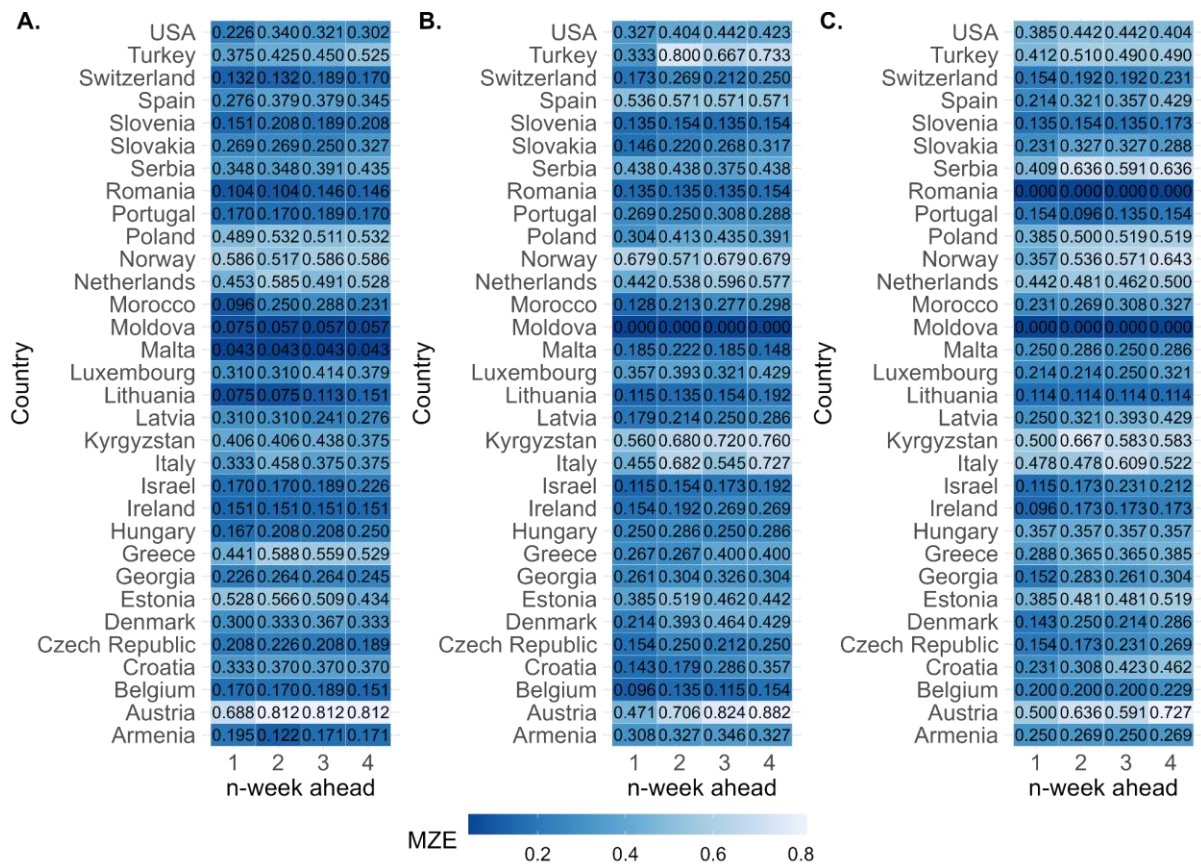
427 forecasting 1 to 4 weeks ahead for the test period year 2015 to 2017.

428



429

430 **S4 Fig. Mean-zero error (MZE) of XGBoost models.** Comparison of MZEs for weekly  
431 forecasting ranging from 1- to 4-week ahead in 2015 to 2017. **A.** The XGBoost model with  
432 an extended approach. **B.** The XGBoost model with a fixed window approach.



433

434 **S5 Fig. Mean-zero error (MZE) for 32 countries.** MZEs of predictions by the XGBoost

435 model with an extended window for the year **A.** 2015. **B.** 2016. and **C.** 2017.

436 **Acknowledgements**

437 The authors acknowledge funding from the MRC Centre for Global Infectious Disease  
438 Analysis (MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and  
439 the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO  
440 Concordat agreement, and also part of the EDCTP2 programme supported by the European  
441 Union. S.R. acknowledges the support from Wellcome Trust Investigator Award (UK,  
442 200861/Z/16/Z). KOK acknowledges funding from HMRF (INF-CUHK-1).

443 **Author Contributions**

444 **Conceptualization:** Haowei Wang, Steven Riley

445 **Data curation:** Haowei Wang

446 **Formal analysis:** Haowei Wang

447 **Funding acquisition:**

448 **Methodology:** Haowei Wang

449 **Software:** Haowei Wang

450 **Supervision:** Steven Riley

451 **Validation:** Haowei Wang

452 **Visualization:** Haowei Wang

453 **Writing – original draft:** Haowei Wang

454 **Writing – review & editing:** Haowei Wang, Steven Riley, Kin On Kwok

455



## 456 References

- 457 1. Influenza (Seasonal). [cited 9 Mar 2022]. Available: [https://www.who.int/en/news-](https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal))  
458 [room/fact-sheets/detail/influenza-\(seasonal\)](https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal))
- 459 2. *Global\_Influenza\_Strategy\_2019\_2030\_Summary\_English.pdf*. Available:  
460 [https://www.who.int/influenza/Global\\_Influenza\\_Strategy\\_2019\\_2030\\_Summary\\_Englis](https://www.who.int/influenza/Global_Influenza_Strategy_2019_2030_Summary_English.pdf)  
461 [h.pdf](https://www.who.int/influenza/Global_Influenza_Strategy_2019_2030_Summary_English.pdf)
- 462 3. Viboud C, Vespignani A. The future of influenza forecasts. *Proceedings of the National*  
463 *Academy of Sciences of the United States of America*. 2019. pp. 2802–2804.
- 464 4. Ali ST, Cowling BJ. Influenza Virus: Tracking, Predicting, and Forecasting. *Annu Rev*  
465 *Public Health*. 2021;42: 43–57.
- 466 5. Chretien J-P, George D, Shaman J, Chitale RA, McKenzie FE. Influenza forecasting in  
467 human populations: a scoping review. *PLoS One*. 2014;9: e94130.
- 468 6. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. A systematic review of  
469 studies on forecasting the dynamics of influenza outbreaks. *Influenza Other Respi*  
470 *Viruses*. 2014;8: 309–316.
- 471 7. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Nonmechanistic forecasts  
472 of seasonal influenza with iterative one-week-ahead distributions. *PLoS Comput Biol*.  
473 2018;14: e1006134.
- 474 8. Yang W, Lau EHY, Cowling BJ. Dynamic interactions of influenza viruses in Hong Kong  
475 during 1998-2018. *PLoS Comput Biol*. 2020;16: e1007989.
- 476 9. Kramer SC, Shaman J. Development and validation of influenza forecasting for 64  
477 temperate and tropical countries. *PLoS Comput Biol*. 2019;15: e1006742.
- 478 10. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts  
479 during the 2012–2013 season. *Nat Commun*. 2013;4: 1–10.
- 480 11. Kermack WO, McKendrick AG, Walker GT. A contribution to the mathematical theory of  
481 epidemics. *Proceedings of the Royal Society of London Series A, Containing Papers of*  
482 *a Mathematical and Physical Character*. 1927;115: 700–721.
- 483 12. Laskowski M, Demianyk BCP, Witt J, Mukhi SN, Friesen MR, McLeod RD. Agent-based  
484 modeling of the spread of influenza-like illness in an emergency department: a  
485 simulation study. *IEEE Trans Inf Technol Biomed*. 2011;15: 877–889.
- 486 13. Arduin H, Domenech de Cellès M, Guillemot D, Watier L, Opatowski L. An agent-based  
487 model simulation of influenza interactions at the host level: insight into the influenza-  
488 related burden of pneumococcal infections. *BMC Infect Dis*. 2017;17: 382.
- 489 14. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza forecasting  
490 with Google Flu Trends. *PLoS One*. 2013;8: e56176.
- 491 15. Spaeder MC, Stroud JR, Song X. Time-series model to predict impact of H1N1  
492 influenza on a children’s hospital. *Epidemiol Infect*. 2012;140: 798–802.

- 493 16. Achrekar H, Gandhe A, Lazarus R, Yu S-H, Liu B. Predicting Flu Trends using Twitter  
494 data. 2011 IEEE Conference on Computer Communications Workshops (INFOCOM  
495 WKSHPs). 2011. pp. 702–707.
- 496 17. Perrotta D, Tizzoni M, Paolotti D. Using Participatory Web-based Surveillance Data to  
497 Improve Seasonal Influenza Forecasting in Italy. Proceedings of the 26th International  
498 Conference on World Wide Web. Republic and Canton of Geneva, CHE: International  
499 World Wide Web Conferences Steering Committee; 2017. pp. 303–310.
- 500 18. Chen S, Xu J, Wu Y, Wang X, Fang S, Cheng J, et al. Predicting temporal propagation  
501 of seasonal influenza using improved gaussian process model. *J Biomed Inform.*  
502 2019;93: 103144.
- 503 19. Predicting Spatio–Temporal Propagation of Seasonal ...Predicting Spatio-Temporal  
504 Propagation of Seasonal. <https://www.aaai.org> › AAI16 › paper ›  
505 download<https://www.aaai.org> › AAI16 › paper › download<https://ojs.aaai.org> ›  
506 index.php › AAI › article › view<https://ojs.aaai.org> › index.php › AAI › article › view.  
507 Available:  
508 <https://www.aaai.org/ocs/index.php/AAI/AAI16/paper/download/11998/12177>
- 509 20. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods.*  
510 2018;15: 233–234.
- 511 21. Bzdok D. Classical Statistics and Statistical Learning in Imaging Neuroscience. *Front*  
512 *Neurosci.* 2017;11: 543.
- 513 22. Bzdok D, Krzywinski M, Altman N. Points of Significance: Machine learning: a primer.  
514 *Nat Methods.* 2017;14: 1119–1120.
- 515 23. Venna SR, Tavanaei A, Gottumukkala RN, Raghavan VV, Maida AS, Nichols S. A  
516 Novel Data-Driven Model for Real-Time Influenza Forecasting. *IEEE Access.* undefined  
517 2019;7: 7691–7701.
- 518 24. A SVM-based prediction method for H5N1 Avian Influenza.  
519 <https://citeseerx.ist.psu.edu> › viewdoc › download<https://citeseerx.ist.psu.edu> › viewdoc ›  
520 download. Available:  
521 <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.842.4765&rep=rep1&type=pdf>  
522
- 523 25. Nsoesie EO, Oladeji O, Abah ASA, Ndeffo-Mbah ML. Forecasting influenza-like illness  
524 trends in Cameroon using Google Search Data. *Sci Rep.* 2021;11: 1–11.
- 525 26. Liang F, Guan P, Wu W, Huang D. Forecasting influenza epidemics by integrating  
526 internet search queries and traditional surveillance data with the support vector machine  
527 regression model in Liaoning, from 2011 to 2015. *PeerJ.* 2018;6: e5134.
- 528 27. Singh S, Kaur H. Influenza prediction from social media texts using machine learning. *J*  
529 *Phys Conf Ser.* 2021;1950: 012018.
- 530 28. Wu Y. DL4Epi: Deep Learning for Epidemiological Predictions. Github;  
531 doi:10.1145/3209978.3210077
- 532 29. Aiken EL, Nguyen AT, Viboud C, Santillana M. Toward the use of neural networks for  
533 influenza prediction at multiple spatial resolutions. *Sci Adv.* 2021;7.

- 534 doi:10.1126/sciadv.abb1237
- 535 30. Li Z, Luo X, Wang B, Bertozzi AL, Xin J. A Study on Graph-Structured Recurrent Neural  
536 Networks and Sparsification with Application to Epidemic Forecasting. Optimization of  
537 Complex Systems: Theory, Models, Algorithms and Applications. Springer International  
538 Publishing; 2020. pp. 730–739.
- 539 31. Hu H, Wang H, Wang F, Langley D, Avram A, Liu M. Prediction of influenza-like illness  
540 based on the improved artificial tree algorithm and artificial neural network. *Sci Rep*.  
541 2018;8: 4895.
- 542 32. Vega T, Lozano JE, Meerhoff T, Snacken R, Mott J, Ortiz de Lejarazu R, et al. Influenza  
543 surveillance in Europe: establishing epidemic thresholds by the moving epidemic  
544 method. *Influenza Other Respi Viruses*. 2013;7: 546–558.
- 545 33. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative  
546 multiyear, multimodel assessment of seasonal influenza forecasting in the United  
547 States. *Proc Natl Acad Sci U S A*. 2019;116: 3146–3154.
- 548 34. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the  
549 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data  
550 Mining. New York, NY, USA: Association for Computing Machinery; 2016. pp. 785–794.
- 551 35. Wang J, Cheng Q, Dong Y. An XGBoost-based multivariate deep learning framework  
552 for stock index futures price forecasting. *Kybernetes*. 2022;ahead-of-print.  
553 doi:10.1108/K-12-2021-1289
- 554 36. Qin R. The Construction of Corporate Financial Management Risk Model Based on  
555 XGBoost Algorithm. *Journal of Mathematics*. 2022;2022. doi:10.1155/2022/2043369
- 556 37. Zhang X-X, Deng T, Jia G-Z. Nuclear spin-spin coupling constants prediction based on  
557 XGBoost and LightGBM algorithms. *Mol Phys*. 2020;118: e1696478.
- 558 38. Ogunleye A, Wang Q-G. XGBoost Model for Chronic Kidney Disease Diagnosis.  
559 *IEEE/ACM Trans Comput Biol Bioinform*. 2020;17: 2131–2140.
- 560 39. FluID. [cited 28 Jul 2022]. Available: [https://www.who.int/teams/global-influenza-](https://www.who.int/teams/global-influenza-programme/surveillance-and-monitoring/fluid)  
561 [programme/surveillance-and-monitoring/fluid](https://www.who.int/teams/global-influenza-programme/surveillance-and-monitoring/fluid)
- 562 40. demo at master · dmlc/xgboost. Github; Available: <https://github.com/dmlc/xgboost>
- 563 41. Forecasting: Principles and Practice (2nd ed). [cited 30 Sep 2022]. Available:  
564 <https://otexts.com/fpp2/>
- 565 42. XGBoost Parameters — xgboost 2.0.0-dev documentation. [cited 3 Nov 2022].  
566 Available: <https://xgboost.readthedocs.io/en/latest/parameter.html>
- 567 43. Baccianella S, Esuli A, Sebastiani F. Evaluation Measures for Ordinal Regression. 2009  
568 Ninth International Conference on Intelligent Systems Design and Applications. 2009.  
569 pp. 283–287.
- 570 44. Jain A, Patel H, Nagalapatti L, Gupta N, Mehta S, Guttula S, et al. Overview and  
571 Importance of Data Quality for Machine Learning Tasks. Proceedings of the 26th ACM  
572 SIGKDD International Conference on Knowledge Discovery & Data Mining. New York,

- 573 NY, USA: Association for Computing Machinery; 2020. pp. 3561–3562.
- 574 45. Liang J-B, Yuan H-Y, Li K-K, Wei W-I, Wong SYS, Tang A, et al. Path to normality:  
575 Assessing the level of social-distancing measures relaxation against antibody-resistant  
576 SARS-CoV-2 variants in a partially-vaccinated population. *Comput Struct Biotechnol J*.  
577 2022;20: 4052–4059.
- 578 46. Kwok KO, Cowling B, Wei V, Riley S, Read JM. Temporal variation of human  
579 encounters and the number of locations in which they occur: a longitudinal study of  
580 Hong Kong residents. *J R Soc Interface*. 2018;15. doi:10.1098/rsif.2017.0838