Heliyon 7 (2021) e08384

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A statistical analysis of antigenic similarity among influenza A (H3N2) viruses

Emmanuel S. Adabor *

School of Technology, Ghana Institute of Management and Public Administration, Accra, Ghana

ARTICLE INFO

Keywords:
Influenza virus
Antigenic similarity
Vaccine effectiveness
Machine learning
Statistical model

ABSTRACT

An accurate assessment of antigenic similarity between influenza viruses is important for vaccine strain recommendations and influenza surveillance. Due to the mechanisms that result in frequent changes in the antigenicities of strains, it is desirable to obtain an antigenic similarity measure that accounts for specific changes in strains that are of epidemiological importance in influenza. Empirically grounded statistical models best achieve this. In this study, an interpretable machine-learning model was developed using distinguishing features of antigenic variants to analyze antigenic similarity. The features comprised of cluster information, amino acid sequences located in known antigenic and receptor-binding sites of influenza A (H3N2). In order to assess validity of parameters, accuracy and relevance of model to vaccine effectiveness, the model was applied to influenza A (H3N2) viruses due to their abundant genetic data and epidemiological relevance to influenza surveillance. An application of the model revealed that all model parameters were statistically significant to determining antigenic similarity between strains. Furthermore, upon evaluating the model for predicting antigenic similarity between strains, it achieved 95% area under Receiver Operating Characteristic curve (AUC), 94% accuracy, 76% precision, 97% specificity, 68% sensitivity and a diagnostic odds ratio (DOR) of 83.19. Above all, the model was found to be strongly related to influenza vaccine effectiveness to indicate the correlation between vaccine effectiveness and antigenic similarity between vaccine and circulating strains in an epidemic. The study predicts probabilities of antigenic similarity and estimates changes in strains that lead to antigenic variants. A successful application of the methods presented in this study would complement the global efforts in influenza surveillance.

1. Introduction

The impact of seasonal influenza infections on populations and economies of countries cannot be underestimated. In particular, these seasonal influenza epidemics are estimated to cause about 3–5 million cases of severe illnesses and about 290,000 to 650,000 deaths globally in each year [1]. The virus frequently changes amino acids located in specific regions of its hemagglutinin protein resulting in antigenic variants which enable it to escape neutralizing antibodies during infections [2, 3]. Although vaccinations are effective medical interventions to protect against these infections, the immunity they provide may decline over time because of the frequent changes in the antigenicity of influenza viruses. Therefore, influenza vaccines are reviewed regularly to make appropriate recommendation of strains to be considered for future vaccines [4]. The vaccine strain recommendation depends on an accurate characterization of strains and determination of antigenic relatedness (similarity) between strains. Usually, a determination is made of antigenic similarity between circulating and vaccine strains through the network of National Influenza Centres established by the World Health Organization (WHO) around the world for influenza surveillance [5]. Traditional methods for measuring antigenic similarity between influenza strains involve the use of Hemagglutination Inhibition (HI) assay results (titres) [6, 7, 8].

Although the direct use of HI titres has been the conventional approach, the need for improved alternatives has necessitated methods based on amino acid sequences [9, 10, 11]. Furthermore, amino acid sequences are thought to be more sensitive to antigenic changes in influenza strains and correlated with the outcomes of refined HI assays [12, 13]. These have motivated the developments of some computational methods for determining antigenic similarity between strains based on amino acid sequences. Most of the computational approaches have been applied to influenza A (H3N2) virus (strain) because of its abundant genetic data and epidemiological importance [14, 15, 16, 17, 18, 19]. Phylogenetic trees have been used to analyze antigenic properties of

* Corresponding author.
E-mail address: healmes@gmail.com.

https://doi.org/10.1016/j.heliyon.2021.e08384
Received 29 March 2021; Received in revised form 21 April 2021; Accepted 10 November 2021

strains [3]. Other computational models enhance the analysis of strains by predicting antigenic variants [14].

In particular, previous modelling approaches incorporated scoring and regression-based methods to predict antigenic variants of influenza A (H3N2) viruses [14]. Other authors using linear models for predicting antigenicity of the influenza A (H3N2) viruses obtained lower accuracies compared to machine learning-based methods [15]. In other applications, a bootstrapped ridged regression model and statistical mechanics approach to modelling have been employed to predict antigenicity of strains [16, 20]. These methods have a common regularization parameter, lasso, that boosted prediction accuracies in their applications [21].

Furthermore, using 12 structural and physiochemical features of the influenza hemagglutinin (HA) protein, a naïve Bayes model was developed to predict antigenic clusters of influenza A (H3N2) viruses [22]. This development followed a suggestion from earlier researchers who recommended that the best vaccine strategies should target antigenic clusters [23]. In particular, this was supported by the findings that the antigenic evolution pattern of influenza A (H3N2) formed chronological clusters based on the years of characterization of strains and that the evolutions of the strains were portrayed as a sequence of antigenic cluster replacements [23, 24, 25]. The need for improving the determination of antigenic relatedness or clusters is underscored from a previous analysis of clusters in which some cluster members were reassigned into different groups after adjusting pairwise antigenic similarity measure [26]. This followed a study that showed that the main parameters, HI titres, for clustering in those studies were affected by both antigenic and non-antigenic factors [27]. Although Bayesian inference methods have been proposed to decouple non-antigenic factors from HI titres to enhance the determination of antigenic similarity [28], models that are based on amino acid changes in known antigenic sites of the HA protein of influenza provide better determinations of antigenic variants [29].

In order to better determine antigenic similarity of influenza A (H3N2) strains, this paper incorporates amino acid changes in antigenic and receptor-binding sites of the HA protein to develop an interpretable statistical classification model to predict antigenic similarity between pairs of strains. Using the number of amino acid changes in both antigenic and receptor-binding sites as features, a supervised machine learning method based on logistic regression was developed. The logistic regression analysis allows direct computational assertions that model components (features) are determinants of antigenicity of influenza strains. In addition, this approach provides numerical estimates of the impacts of model components on the antigenic relatedness between strains. These properties of the model make it interpretable and useful for analyzing antigenic similarity between strains. In order to assess the impact of the model on vaccine strain recommendation, the study establishes a relationship between the model outcomes and vaccine effectiveness. Labelled datasets for testing the methods were curated from influenza A (H3N2) antigenic clusters analyzed in previous studies [23]. The evaluation of the method reveals its potential to accurately decipher antigenic similarity between influenza strains and to support the global surveillance of influenza.

2. Materials and methods

2.1. Materials and data

In order to obtain the required data for supervised machine learning technique used in this study, 11 antigenic clusters of influenza A (H3N2) strains were obtained from previous studies by Smith and colleagues [23]. By this clustering, 231 influenza A (H3N2) strains with complete amino acid sequences were found to be members of the 11 antigenic clusters. The amino acid sequences of the HA of influenza A (H3N2) viruses were obtained from the National Center for Biotechnology Information (NCBI) protein database [30]. A multiple sequence alignment was performed using Clustal Omega [31]. The influenza A (H3N2) strains, GenBank sequence accession numbers and clusters are presented as supplementary files (Supplementary Table S1). Here, pair of strains belonging to the same antigenic clusters are considered as antigenically similar and pair of strains from different clusters are considered antigenically distinct. These definitions of groupings are based on previous analysis of influenza A (H3N2) viruses [22, 23]. Applying these to the current strains resulted in 2,832 antigenically similar pairs and 23,733 antigenically distinct pairs of strains. That is, in all, 26,565 pairs of strains were used in this study (Supplementary Table S2).

The features of the data used for developing the proposed learning model comprised of number of differences in amino acid residues located on designated positions of antigenic and receptor binding sites of HA proteins of influenza A (H3N2) strains. There are 131 amino acid positions on five designated antigenic sites, epitopes (A, B, C, D and E), and 14 amino acid positions on the receptor-binding site [32, 33, 34, 35]. Changes in amino acids located in these sites accompany different antigenic variants of influenza A (H3N2) [23,32]. These amino acid changes in the antigenic and receptor-binding sites of the HA protein of influenza were identified as suitable parameters for the model.

The corresponding amino acids in these sites in the HA proteins of the strains considered in this study were identified and selected after multiple sequence alignment. The pairwise differences in the selected amino acid residues for all the relevant sites were obtained using R code (available in Supplementary Material). The code is implemented in R version 3.6.3. Datasets obtained for the six relevant features (five antigenic and receptor-binding sites) were used to build the model for predicting antigenic similarities among pairs of strains.

Models based on the designated antigenic sites in the HA proteins of influenza A (H3N2) strains are reported to be better predictors of antigenic similarities and have been used in such analytical studies in the past [29]. Furthermore, the designated antigenic and receptor-binding sites in the HA proteins of influenza A (H3N2) strains are considered because the number of amino acid changes in these regions are associated with antigenic variants of strains [23, 32]. Therefore, they make the approach presented in this study more sensitive to antigenic drifts.

2.2. Statistical methods

The goal of the modelling is to establish a method for classifying any pair of influenza strains as similar or distinct using the relevant amino acid substitutions accompanying variants of a strain. This study develops a logistic regression model, as it is more suitable for interpreting (binary) outcomes of classification models. Unlike the well-known least squares regression, which cannot predict a qualitative response, the logistic model predicts qualitative response variable with meaningful model parameters and a probability of prediction. These features of the logistic regression model make it more interpretable and preferable compared to other machine learning techniques that can be applied to the classification task modelled in this study. In general, for a set of n features, $X_1, X_2, X_3, \ldots, X_n$, a logistic function defines the probability of an instance, say X, belonging to a category, say i, as Eq. (1):

$$P(X) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n}}, \quad (1)$$

where parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are effects of corresponding features $X_1, X_2, X_3, \ldots, X_n$ on determining the class of X with a probability value, $P(X)$.

The parameters are estimated using the maximum likelihood method [36]. This ensures that the estimated coefficients maximizes or minimizes the probability that an instance, X, belongs to a given class, say i. In this study, an instance of a pair of strains can be either antigenically similar or antigenically distinct. That is, there are only two possible classes and these are the contexts of application considered. Once the effects are determined in Eq. (1), it can ascertain the class of any instance based on the value of the probability. Although other

regression-based approaches have been applied in the past using a different scoring scheme [14], this study is distinguished by the attributes or features of the antigenic data (antigenic and other receptor-binding sites) being incorporated as effects, the specific sequence information and the emphasis on the interpretation of model parameters in relation to the relatedness between strains. Here, the significance of the model parameters to determining antigenic similarity is assessed.

In this study, statistical testing of significance of all coefficients in the model was performed at 5% level of significance. The predictive performance of model was assessed in a seven-fold cross-validation. In the seven-fold cross-validation process, the dataset was randomly divided into seven equal subsets, one subset was held-out as a test set while the remaining six subsets were combined and used to train or develop model. Then, the held-out subset was used to test the model. The process was repeated for each subset. The predictions from all the seven subsets were combined in order to assess the overall performance of model. For a more standard evaluation of the accuracy of model, the area under a Receiver Operating Characteristic (ROC) curve is employed [37]. The ROC curve measures the discriminating power of a model showing the trade-off between sensitivity and false positive rate of the method. The performance of the method is based on the numbers of correct or incorrect predictions. In particular, a correctly predicted pair of antigenically similar strains is counted as True Positive (TP). A pair of antigenically similar strains wrongly predicted as antigenically distinct is counted as False Negative (FN). A correctly predicted pair of antigenically distinct strain is counted as True Negative (TN). A pair of antigenically distinct strains wrongly predicted as antigenically similar strains is counted as False Positive (FP). Confusion or error matrix is usually presented by machine learning algorithms to convey these indicators as in Table 1.

In this study, the entries of the confusion matrix were obtained using R code (available in the Supplementary Material). With these performance indicators, the antigenic relatedness model was evaluated using the following accuracy measures:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

$$Precision = \frac{TP}{TP + FP}, \tag{3}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{5}$$

$$Diagnostic\ odds\ ration\ (DOR) = \frac{TP \times TN}{FP \times FN}. \tag{6}$$

The higher these measures, the better the predictive model. Although accuracy, precision, specificity, and sensitivity can reach a maximum of 100%, it is desirable for the diagnostic odds ratio (DOR) to be greater than one (1) since it is an indicator of a better model's predictive performance. The methods and analysis were conducted using codes implemented in R (version 3.6.3). The r-code accompanying this manuscript can be found in the Supplementary Material.

Table 1. Error matrix.

	Predicted pair of distinct strains	Predicted pair of similar strains
Actual pair of distinct strains	True negative	False positive
Actual pair of similar strains	False negative	True positive

2.3. The model for predicting antigenic similarity between strains

The analysis of the antigenicity between two strains is crucial in influenza surveillance. Using the designated antigenic and receptor-binding sites as parameters for the model makes it possible to analyze the model properties and antigenic characterization of the influenza strains. Therefore, assuming X_1, X_2, X_3, X_4, X_5 and X_6 are, respectively, the numbers of differences in amino acids located on designated antigenic sites A, B, C, D and E, and receptor-binding sites of any pair of strains, then the antigenic similarity is determined by the relation:

$$\log\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \sum_{i=1}^{6} \beta_i x_i, \tag{7}$$

where $P(x)$ is the probability that the pair of strains are similar, $\beta_i^{'}$sare effects denoting the significance and extent of effect of each antigenic site or receptor binding site on the similarity of the two strains, and $\beta_0$is a constant which accounts for other factors that contribute to antigenic similarity between the strains.

From Eq. (7), the relationships between probability value and effects are not linear. However, it can be readily deduced that for a positive effect, sayβ_1, if the differences in amino acids located on the antigenic site A of any pair of strains increase, then $P(x)$ also increases. However, if the effect is negative, then increasing the differences in the amino acid residues at the antigenic site A will decrease the probability, $P(x)$, that the pair of strains under consideration is similar. This interpretation, which applies to every other model parameter, can be readily deduced from Eq. (1).

2.4. Correlation analysis of influenza vaccine effectiveness and antigenic similarity model

Correlation analysis is concerned with the analysis of linear association between two variables. A quantitative measure of the strength of linear relationship between the two variables is correlation coefficient. In this study, the variables are vaccine effectiveness and antigenic relatedness model. Vaccine effectiveness (VE) provides estimate of the fraction of influenza cases prevented by vaccination in an influenza season. It is defined by:

$$VE = (1 - RR) \times 100, \tag{8}$$

where RR is the risk ratio defined by:

$$RR = \frac{ad}{bc}, \tag{9}$$

a is the total number of vaccinated individuals diagnosed with influenza-like illness in the particular VE study, b is total number of vaccinated individuals taking part in the study, c is the number of unvaccinated individuals diagnosed with influenza-like illness in the study, and d is the total number of unvaccinated individuals taking part in the study. The risk ratio is the relative risk of contracting influenza-like illness in the vaccinated individuals compared to the unvaccinated. Although this method for evaluating influenza vaccine effectiveness has been commonly used in the past, a more recent vaccine effectiveness study design, test-negative design, has been used over the last 16 years to obtain estimates. This recent method assumes that influenza vaccine only provides protection against influenza but does not affect non-influenza causes of influenza-like illness. By this method, the VE is given by Eq. (10):

$$VE = \left(1 - \frac{O_{positive}}{O_{negative}}\right) \times 100\%, \tag{10}$$

where $O_{positive}$ is odds of vaccination among study participants testing positive for influenza and $O_{negative}$ is odds of vaccination among those

testing negative. Usually, adjusted odds ratios of the positive and negative cases are determined using logistic models where adjustments are made for covariates such as age, study site, calendar time, presence of high-risk health conditions, race, among others. The VE estimates incorporating these adjustments are referred to as adjusted VE. In this study, adjusted VE were used to obtain estimates for the last 16 years, 2004–2005 to 2019–2020 influenza seasons (with details provided in the Supplementary Material). The variable of antigenic similarity model measures antigenic similarity between any pair of strain, which is given by the probability, P(X), defined by Eq. (1).

In this study, VE data was searched from literature and it was found to be normally distributed. However, the probability data for the antigenic similarity model variable was not normally distributed. Therefore, a more robust and suitable measure of correlation, Spearman's rank correlation coefficient, was used for the correlation analysis. This is given by Eq. (11):

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{11}$$

where d_i is the difference between i-th pair of ranks and n is the total number of observations.

3. Results

3.1. An application of model to influenza A (H3N2) data

In order to illustrate validity and interpretation of model, it is applied to the antigenic data of influenza A (H3N2) viruses considered in this study. The modelling made it possible to evaluate the contribution and significance of amino acid changes in each of the five designated antigenic and receptor binding sites of the HA protein of influenza A (H3N2). More specifically, it was found that all the parameters of the antigenic similarity model had statistically significant contributions to the similarity or dissimilarity of any pair of strains and that the model is accurate (Table 2). These results are important and consistent with expectations that antigenic variant of epidemiological importance usually has at least four amino acid changes in at least two of the designated antigenic sites [32]. Among the five designated antigenic sites, contributions of changes in amino acid residues in sites A and B to antigenic similarity between the influenza strains were higher compared to contributions of the other antigenic sites (Table 2). These findings are consistent with other experimental studies that found amino acid changes in antigenic sites A and B to be the major drivers of antigenic drift in influenza A (H3N2) [38, 39]. Here, the modelling distinguishes itself by providing details of the extent of contributions of these antigenic changes in each site that may lead to antigenic variants of the virus.

Table 2. Estimates of model parameters and significance.

Site	Parameter (β_i)	Significance (p-value)	**Lower limit	***Upper limit
Antigenic site A	-1.03	<2e-16	-1.09	-0.96
Antigenic site B	-0.47	<2e-16	-0.51	-0.43
Antigenic site C	0.3	<2e-16	0.24	0.37
Antigenic site D	0.05	0.01	0.01	0.08
Antigenic site E	0.31	<2e-16	0.26	0.35
Receptor-binding site	-0.29	8.28e-16	-0.36	-0.22
*Constant	1.69	<2e-16	1.57	1.81

* Model constant accounting for other factors that contribute to antigenic variations among strains.
** Lower limit of 95% Wald confidence interval of parameter estimates.
*** Upper limit of 95% Wald confidence interval of parameter estimates.

The antigenic sites, A and B, are crucial for neutralizing antibodies and changes in the amino acids in these epitopes facilitate the escape of the virus from neutralizing antibodies [40]. The results revealed that amino acid changes in these epitopes decreased the chances of antigenic similarity between pairs of strains. The (negative) signs and magnitudes of coefficients of amino acid changes in the antigenic sites A and B (Table 2) explain this. The results provide empirical evidence in support of previous findings that these epitopes are the most valuable epitopes in directing the evolution of influenza A (H3N2) viruses reported in previous studies [41, 42]. The findings of this study are consistent with a recent study which found that key amino acid substitutions in antigenic sites A and B were responsible for major antigenic changes in influenza A (H3N2) [43]. A further study by Koel and colleagues suggests that the magnitudes of antigenic effects of some amino acid substitutions in these sites are context-dependent [44]. Additionally, differences in amino acid residues in receptor-binding sites of pairs of strains also decrease the probability of antigenic relatedness of those strains. This is a consequence of the negativity of the corresponding effect of the receptor-binding site (Table 2).

On the other hand, the probability of antigenic relatedness between pairs of strains are less affected by amino acid changes in antigenic sites C, D and E compared to antigenic sites A and B. This is due to the relatively lower magnitudes of coefficients corresponding to antigenic sites C, D, and E compared to antigenic sites A and B (Table 2). However, the antigenic sites C, D and E have positive coefficients suggesting that changes in those sites will not lead to increase antigenic dissimilarity between pair of strains. Interestingly, the findings that the changes in antigenic sites C, D and E did not lead to increasing antigenic dissimilarity are in line with previous studies [43]. In particular, it was reported that amino acid changes in the sites C, D and E did not cause major antigenic changes in the strains [43]. These previous findings affirm the statistical analysis that did not predict that changes in the sites C to E could increase antigenic dissimilarity. These consistent findings are noteworthy since both studies considered the evolution of influenza A (H3N2) over 35 years, 1968 to 2003. While the study by Koel and coworkers [43] used a representative virus for each cluster in their study, the current study considered all the viruses within the clusters.

3.2. Accuracy of predictions of antigenic similarity model

In order to evaluate performance of the model, it is tested in a 7-fold cross-validation using the data curated for the study. The proposed model performed better than a random model and has a value of 95% area under ROC curve (Figure 1). These results provide further support for the

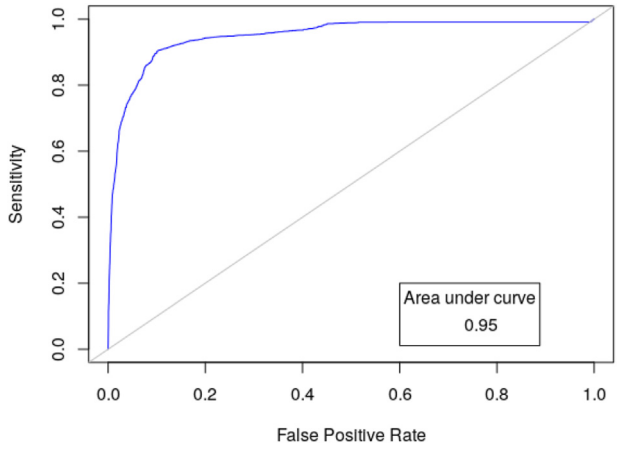

Figure 1. Receiver-operating characteristic (ROC) curve of antigenic similarity model. Area under ROC curve (AUC) of proposed model is appended on the plot. Results are produced from 7-fold cross-validation. The proposed model has higher AUC compared to the random model.

use the model for analyzing antigenic similarity between strains. Not only is the method interpretable, it also compares favourably well with other methods for predicting antigenic similarity among influenza A (H3N2) strains. In particular, the proposed model achieved 95% area under ROC curve (AUC), 94% accuracy, 76% precision, 97% specificity, 68% sensitivity and a diagnostic odds ratio (DOR) of 83.19. These were derived from performance indicators presented in the error matrix (Table 3). These performance scores are better compared to other methods based on Naive Bayes and statistical mechanics employed to predict antigenicity of influenza A (H3N2) strains, using different features [20, 45]. Those methods were reported to achieve accuracies that do not exceed 90% even though they were better than random model. However, it should be noted that while all antigenic and receptor binding sites are considered in this study, those previous studies considered regional band or artificial sites located on the HA1 protein as features for classification models, combinations of HI titres and amino acid sequences [20, 45]. In this study, features are changes in amino acid residues located on designated antigenic and receptor binding sites of the HA protein of influenza A (H3N2).

3.3. The model of antigenic similarity is linked to vaccine effectiveness

Furthermore, the model is assessed in relation to influenza vaccine effectiveness. This is important because the antigenic similarity between circulating and vaccine strains is evaluated at each influenza season to make recommendations for future vaccines. The outcome of the antigenic similarity between the circulating and the vaccine strains determines whether to maintain or substitute the vaccine strain in subsequent influenza seasons. In addition, besides factors such as age and immune status of vaccinated individuals, the antigenic similarity between circulating and vaccine strains also affect influenza vaccine effectiveness [46]. Therefore, a desirable property of a measure of antigenic similarity is that it significantly correlates with the vaccine effectiveness. In order to test this assertion on the model presented in this study, data on influenza vaccine effectiveness were searched from the literature (Supplementary Material).

The vaccine effectiveness data consists of influenza seasons in which influenza A (H3N2) subtype dominated, the dominant influenza A (H3N2) that circulated, and the estimated vaccine effectiveness (Supplementary Material). While only pairs of vaccine and circulating strains, prior to the 2004–2005 influenza season, whose amino acid sequences are present in the data collected for this study were considered, all other vaccine and dominant circulating influenza A (H3N2) strains for the last 16 years were also included in this VE analysis. Regarding the antigenic similarity data from model, the probabilities of antigenic similarity estimated from the model were used (Supplementary Material). The antigenic similarity model was expected to correlate with vaccine effectiveness. As anticipated, a strong statistically significant correlation coefficient was established between the antigenic similarity model and vaccine effectiveness. It was found that the model strongly correlated with vaccine effectiveness prior to the 2004–2005 influenza season ($r = 0.76$, p-value $= 4.11e\text{-}04$, r-squared $= 0.58$), vaccine effectiveness from 2004-2005 to 2018–2019 influenza seasons ($r = 0.72$, p-value $= 0.02$, r-squared $= 0.51$), and vaccine effectiveness over all available seasons ($r = 0.71$, p-value $= 3.67e\text{-}05$, r-squared $= 0.50$). These results suggest that there exist a strong positive linear relationship between vaccine effectiveness and antigenic similarity between circulating and vaccine strains in recent years despite the changes in VE study designs. Furthermore, the results suggest that at least 50% of variations in the vaccine effectiveness could be explained by the antigenic similarity between circulating and vaccine strains. This value is significant since there are other factors that contribute to the vaccine effectiveness for any influenza season.

Table 3. Error matrix produced by the antigenic similarity model.

	Predicted pair of distinct strains	Predicted pair of similar strains
Actual pair of distinct strains	23136	597
Actual pair of similar strains	900	1932

4. Discussion

An accurate prediction of antigenic similarity between circulating and vaccine strains is paramount to vaccine strain recommendation. Therefore, a desirable property of a model or measure of antigenic similarity is to have its parameters and results to be directly applied to differences in characterized influenza strains. In this study, an interpretable model is proposed for analysis of antigenic similarity between strains. In particular, the model's parameters are differences in amino acid residues located on antigenic and receptor-binding sites of the HA protein of influenza A (H3N2) viruses. In addition, the interpretable nature of the model and its parameters permit a direct assessment of significance and nature of antigenic similarity between strains. These can be examined from both the magnitude and the sign of estimated parameters (Table 2). Moreover, the study established that all effects were statistically significant and that the model is adequate.

The statistical significance of all model parameters suggests that the parameters of the model have notable contributions to antigenic similarity between pair of influenza A (H3N2) strains. These results provide a quantitative assessment of association of amino acid changes in designated antigenic and receptor-binding sites to variations in antigenicities of strains. The model makes it possible to compare the levels of effects of amino acid changes at the individual sites considered in this study. In particular, amino acid changes in antigenic sites A and B have the greatest effect on antigenic similarity between strains compared to other sites considered in the study. This is indicated by the higher magnitudes of coefficients of the changes in sites A and B compared to sites C through E and the receptor-binding site (Table 2). These findings are in line with other studies that identified antigenic sites A and B to be most valuable epitopes capable of directing antigenic drift [47]. While increasing changes in amino acids located on antigenic sites A and B and the receptor-binding site may lead to decrease antigenic similarity between pairs of strains, the same cannot be said of antigenic sites C, D and E even though they are statistically significant determinants of antigenic similarity between pair of strains. This is due to the positive values of the coefficients of the changes in sites C to E (Table 2).

These results are consistent with previous studies that identified the need to change vaccine designs due to frequent changes in the sites A and B [48]. Nevertheless, a recent survey of antigenicity of influenza viruses suggested that a new epidemic strain would have changes in all antigenic sites in order to escape the human immune system [49]. Here, the proposed model showed that the designated sites considered in this study are relevant to antigenic similarity as they all had statistically significant coefficients. This is particularly important because of the frequent changes in the antigenicity of circulating influenza strains and the consequential vaccine updates [50, 51, 52].

Furthermore, the model is effective as it achieves 95% area under ROC curve when applied to influenza A (H3N2) data. This is in addition to the fact that it is linked to influenza vaccine effectiveness. Here, the interpretable model strongly correlated with vaccine effectiveness. Particularly, 51% variations in vaccine effectiveness since the 2004–2006 influenza seasons dominated by A/H3N2 is explained by the proposed model, which agrees in principle to the fact that vaccine effectiveness has other determinants. Overall, it suggests that a high antigenic similarity between vaccine and epidemic strains will lead to a high vaccine effectiveness given other factors of vaccine effectiveness.

5. Conclusion

In this study, using antigenic properties of influenza A (H3N2) strains, an interpretable machine-learning model was developed to analyze antigenic similarity between pair of strains. The model was successfully applied to influenza A (H3N2) data in as much as it had a desirable property of providing quantitative assessments of the antigenic characteristics that distinguishes antigenic variants. Furthermore, the model was linked to influenza vaccine effectiveness where it estimated the correlation between antigenic similarity between vaccine and epidemic strains and vaccine effectiveness in influenza seasons. Therefore, a successful application of the model of antigenic similarity could greatly improve prediction accuracies of antigenic future of influenza virus. In this way, recommendations for influenza vaccine design will be enhanced and an overview of how strains effectively bind together shall be ascertained. In this study, features considered for developing the model were derived from the HA protein of influenza A (H3N2). Therefore, a modification of the features for the model will be necessary in order to incorporate other proteins such as neuraminidase for antigenic analysis. This is a possible direction for future research.

Declarations

Author contribution statement

Emmanuel Adabor: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Ghana Institute of Management and Public Administration.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The author declares no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2021.e08384.

References

[1] World Health Organization, Influenza (Seasonal). https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). (Accessed 2 November 2020).
[2] Carrat F, Flahault A. Influenza vaccine: the challenge of antigenic drift. Vaccine;25: 6852-6862.
[3] W.M. Fitch, R.M. Bush, C.A. Bender, N.J. Cox, Long term trends in the evolution of H(3) HA1 human influenza type A, Proc. Natl. Acad. Sci. U. S. A. 94 (15) (1997) 7712–7718.
[4] S.T. Layne, Human influenza surveillance: the demand to expand, Emerg. Infect. Dis. 12 (2006) 562–568.
[5] World Health Organization, WHO global influenza programme: survey on capacities of national influenza centres, January-June 2002, Wkly. Epidemiol. Rec. 77 (2002) 349–356.
[6] I. Archetti, F.L. Horsfall, Persistent antigenic variation of influenza A viruses after incomplete neutralization inovo with heterologous immune serum, J. Exp. Med. 92 (1950) 441–462.
[7] E.D. Kilbourne, B.E. Johansson, B. Grajower, Independent and disparate evolution in nature of influenza A virus haemagglutinin and neuraminidase glycoproteins, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 786–790.
[8] W. Ndifon, J. Dushoff, S.A. Levin, On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness, Vaccine 27 (2009) 2447–2452.
[9] J.B. Plotkin, J. Dushoff, S.A. Levin, Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus, Proc. Natl. Acad. Sci. U.S.A. 99 (9) (2002) 6263–6268.
[10] Y. Yao, X. Li, B. Liao, et al., Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method, Sci. Rep. 7 (1545) (2017).
[11] O. Hungnes, The role of genetic analysis in influenza virus surveillance and strain characterisation, Vaccine 20 (Suppl 5) (2002) B45–B49.
[12] J.S. Ellis, P. Chakraverty, J.P. Clewley, Genetic and antigenic variation in the haemagglutinin of recently circulating human influenza A(H3N2) viruses in the United Kingdom, Arch. Virol. 140 (11) (1995) 1889–1904.
[13] A. Hay, V. Gregory, A. Douglas, Y. Lin, The evolution of human influenza viruses, Philos. Trans. R. Soc. Lond. B Biol. Sci. 356 (1416) (2001) 1861–1870.
[14] Y.C. Liao, M.S. Lee, C.Y. Ko, C.A. Hsiung, Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus, Bioinformatics 24 (4) (2008) 505–512.
[15] W.D. Lees, D.S. Moss, A.J. Shepherd, A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2, Bioinformatics 26 (11) (2010) 1403–1408.
[16] H. Sun, J. Yang, T. Zhang, et al., Using sequence data to infer the antigenicity of influenza virus, Mbio 4 (4) (2013) e00230-13.
[17] M.S. Lee, M.C. Chen, Y.C. Liao, C.A. Hsiung, Identifying potential immunodominant positions and predicting antigenic variants of influenza A/H3N2 viruses, Vaccine 25 (48) (2007) 8133–8139.
[18] L. Steinbrück, T. Klingen, A. McHardy, Computational prediction of vaccine strains for human influenza A (H3N2) viruses, J. Virol. 88 (20) (2014) 12123–12132.
[19] E.K. Lee, H. Tian, H.I. Nakaya, Antigenicity prediction and vaccine recommendation of human influenza virus A (H3N2) using convolutional neural networks, Hum. Vaccines Immunother. 16 (11) (2020) 2690–2708.
[20] A.M. Degoot, E.S. Adabor, F. Chirove, W. Ndifon, Predicting antigenicity of influenza A viruses using biophysical ideas, Sci. Rep. 9 (2019) 10218.
[21] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. R. Stat. Soc. B Stat. Methodol. 58 (1996) 267–288.
[22] X. Du, L. Dong, Y. Lan, et al., Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation, Nat. Commun. 3 (709) (2012).
[23] D.J. Smith, A.S. Lapedes, J.C. de Jong, et al., Mapping the antigenic and genetic evolution of influenza virus, Science 305 (2004) 371–376.
[24] K. Koelle, S. Cobey, B. Grenfell, M. Pascual, Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans, Science 314 (2006) 1898–1903.
[25] E.S. Adabor, Anticipating time-dependent antigenic variants of influenza A (H3N2) viruses, Infect. Genet. Evol. 67 (2019) 67–72.
[26] Y. Li, D.L. Bostick, C.B. Sullivan, et al., Single haemagglutinin mutations that alter both antigenicity and receptor binding avidity influence influenza virus antigenic clustering, J. Virol. 87 (2013) 9904–9910.
[27] W. Ndifon, New methods for analyzing serological data with applications to influenza surveillance, Inf. Other Resp. Vir. 5 (2011) 206–212.
[28] E.S. Adabor, W. Ndifon, Bayesian inference of antigenic and non-antigenic variables from haemagglutination inhibition assays for influenza surveillance, R. Soc. open sci. 5 (2018) 180113.
[29] M.S. Lee, J.S.E. Chen, Predicting antigenic variants of influenza A/H3N2 viruses, Emerg. Infect. Dis. 10 (8) (2004) 1385–1389.
[30] NCBI Resource Coordinators, Database resources of the national center for Biotechnology information, Nucleic Acids Res. 44 (2016) D7–D19.
[31] F. Sievers, A. Wilm, D. Dineen, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011) 539.
[32] I.A. Wilson, N.J. Cox, Structural basis of immune recognition of influenza virus hemagglutinin, Annu. Rev. Immunol. 8 (1999) 737–771.
[33] D.C. Wiley, I.A. Wilson, J.J. Skehel, Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation, Nature 289 (1981) 373–378.
[34] W. Ndifon, N.S. Wingreen, S.A. Levin, Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines, Proc. Natl. Acad. Sci. U.S.A. 106 (21) (2009) 8701–8706.
[35] I.A. Wilson, J.J. Skehel, D.C. Wiley, Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution, Nature 289 (1981) 366–373.
[36] R. Hogg, J.W. McKean, A.T. Craig, Introduction to Mathematical Statistics, eighth ed., Pearson, US, 2019.
[37] W.L. Martinez, A.R. Martinez, Computational Statistics Handbook with MATLAB, second ed., Chapman & Hall/CRC, Boca Raton, FL, 2007.
[38] K. Sato, T. Morishita, E. Nobusawa, et al., Amino acid change on the antigenic region B1 of H3 haemagglutinin may be a trigger for the emergence of drift strain of influenza A virus, Epidemiol. Infect. 132 (2004) 399–406.
[39] R.M. Bush, W.M. Fitch, C.A. Bender, N.J. Cox, Positive selection on the H3 haemagglutinin gene of human influenza virus A, Mol. Biol. Evol. 16 (1999) 457–1465.
[40] L.E. Brown, J.M. Murray, D.O. White, D.C. Jackson, An analysis of the properties of monoclonal antibodies directed to epitopes on influenza virus hemagglutinin, Arch. Virol. 114 (1990) 1–26.
[41] R.M. Bush, C.A. Bender, K. Subbarao, N.J. Cox, W.M. Fitch, Predicting the evolution of human influenza A, Science 286 (1999) 1921–1925.
[42] E. Nobusawa, K. Omagari, S. Nakajima, K. Nakajima, Reactivity of humanconvalescent sera with influenza virus hemagglutinin protein mutants atantigenic site A, Microbiol. Immunol. 56 (2012) 99–106.
[43] B.F. Koel, D.F. Burke, T.M. Bestebroer, et al., Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution, Science 342 (6161) (2013) 976–979.

[44] B.F. Koel, D.F. Burke, S. van der Vliet, et al., Epistatic interactions can moderate the antigenic effect of substitutions in haemagglutinin of influenza H3N2 virus, J. Gen. Virol. 100 (5) (2019) 773–777.
[45] Y. Peng, D. Wang, J. Wang, et al., A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures, Sci. Rep. 7 (2017) 42051.
[46] H.Q. McLean, B.D.W. Chow, J.J. VanWormer, J.P. King, E.A. Belongia, Effect of statin use on influenza vaccine effectiveness, J. Infect. Dis. 214 (8) (2016) 1150–1158.
[47] L. Popova, K. Smith, A.H. West, et al., Immunodominance of antigenic site B over site A of hemagglutinin of recent H3N2 influenza viruses, PLoS One 7 (7) (2012), e41895.
[48] J.W. Huang, J.M. Yang, Changed epitopes drive the antigenic drift of influenza A (H3N2) viruses, BMC Bioinf. 12 (Suppl 1) (2011) S31.
[49] G.M. Air, Influenza virus antigenicity and broadly neutralizing epitopes, Curr. Opin. Virol. 11 (2015) 113–121.
[50] R.G. Webster, W.G. Laver, G.M. Air, G.C. Schild, Molecular mechanisms of variation in influenza viruses, Nature 296 (1982) 115–121.
[51] E.C. Holmes, E. Ghedin, N. Miller, et al., Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses, PLoS Biol. 3 (2005), e300.
[52] N.J. Cox, C.A. Bender, The molecular epidemiology of influenza viruses, Semin. Virol. 6 (1995) 359–370.