RESEARCH ARTICLE

Optimizing influenza vaccine composition: A machine learning approach

Hari Bandi1 | Dimitris Bertsimas2

1 Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
2 Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Correspondence
Dimitris Bertsimas, Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
Email: dbertsim@mit.edu

History
Accepted by Sanjay Mehrotra, healthcare management.

Abstract
We propose a holistic framework based on state-of-the-art methods in machine learning and optimization to prescribe influenza vaccine composition that are specific to a region, or a country based on historical data concerning the rates of circulation of predominant viruses. First, we develop a tensor completion formulation to predict rates of circulation of viruses for the next season based on historical data. Then, taking into account the uncertainty in the predicted rates of circulation of predominant viruses, we propose a novel robust prescriptive framework for selecting suitable strains for each subtypes of the flu virus: Influenza A (H1N1 and H3N2) and B viruses for production. Through numerical experiments, we show that our proposed vaccine compositions could potentially lower morbidity by 11–14% and mortality by 8–11% over vaccine compositions proposed by World Health Organization (WHO) for Northern Hemisphere, and lower morbidity by 8–10% and mortality by 6–9% over vaccine compositions proposed by U.S. Food and Drug Administration (FDA) for United States, and finally, lower morbidity by 10–12% and mortality by 9–11% over vaccine compositions proposed by European Medicines Agency (EMA) for Europe.

KEYWORDS
robust optimization, tensor completion, vaccines

1 INTRODUCTION

Influenza (flu) is a highly contagious respiratory viral disease and the seasonal flu epidemics affect about 5–15% of the world’s population, and cause an estimated 3–5 million cases of severe illnesses and up to half a million annual deaths worldwide. The flu viruses can be segregated into four types, namely influenza A, B, C, and D. Influenza A and B viruses circulate and therefore are mainly responsible for seasonal flu epidemics, whereas influenza C viruses are not detected frequently and usually cause mild infections; influenza D viruses affect cattle and thus do not present a serious public health risk. Influenza A viruses are classified on the basis of their two surface proteins: hemagglutinin and neuraminidase, into 18 different subtypes of hemagglutinin and 11 different subtypes of neuraminidase viruses. Together, all subtypes of influenza A and influenza B viruses are further classified into various strains based on their antigenic properties (response to antibodies).

The flu shot (vaccine), which contains two strains of the influenza A virus (H1N1 and H3N2) and one or two strains of the B virus is the first line of defense against seasonal epidemics. The influenza B viruses have two distinct lineages, therefore in addition to a vaccine containing three virus-strains, manufacturers also produce a vaccine containing four virus-strains vaccine which includes two influenza B strains to cover both lineages. Most individuals have some level of prior immunity. However, new strains with mutations in their protein regions that are not recognized by human antibodies frequently arise and it is well established that these strains have an advantage over existing dominant strains to effectively escape from host immunity. This continuous process of evolution results in a rapid turnover of the viral population and poses a great challenge to producing an effective vaccine.

The flu vaccine is annually reformulated due to rapid emergence of new strains, and prepared at least 6–8 months in advance of the upcoming flu season in order to have enough
time for production and distribution. Every year, the vaccine compositions for the Northern and Southern Hemispheres are reviewed and updated as necessary by the World Health Organization (WHO) through a global surveillance and response system. The WHO monitors and collects data on antigenic characterization, genetic variations, prevalence rates, and geographic distributions of virus variants across the world. Although, antigenic characterization of circulating viruses by standard ferret antibodies is the main determinant in vaccine strain selection, many approaches have been proposed to partially explain mutations in the virus strains using genomic data in the literature. Predicting emerging virus strains is a complex problem due to the uncertain nature of the continuous evolution of the virus strains. Moreover, predicting the fate of strains currently circulating in the population is difficult for two reasons. First, multiple strains carrying different combinations of mutations co-circulate and therefore compete with one another for potential hosts. Second, antigenic characterization via ferret antibodies is different from that of human post-vaccination antibodies because humans and ferrets have different immune systems as well as very different prior exposure to influenza viruses (Agor & Özaltun, 2018).

In order to tackle these challenges, we employ a variety of state-of-the-art methods in machine learning and optimization to prescribe influenza vaccine composition that are specific to a region, or a country based on historical data concerning the rates of circulation of predominant viruses. Below, we briefly outline our approach.

1. **Tensor completion**: We adapt an algorithm proposed in Bertsimas and Pawlowski (2019) to predict rates of circulation of viruses in a season. The historical rates of circulation of viruses are available in the form of a three-dimensional matrix \( M \in \mathbb{R}^{n \times m \times T} \), where \( n \) is the number of countries participating in WHO’s Global Influenza Surveillance and Response System (GISRS), \( m \) is the number of viruses in the data set and \( T \) is the number of flu seasons. Such a three-dimensional matrix is known as a three-dimensional tensor.

   About 88% of the tensor entries are missing. This is primarily due to the fact that GISRS continuously monitors novel emerging strains and adds them to the database over time. Therefore, for newly identified viruses, we observe missing entries for all earlier flu seasons. For example, a particular virus first identified in 2009 will have missing entries in all the preceding flu seasons. Moreover, as viruses keep undergoing mutations over time, not all viruses in the data set may be circulating in a particular flu season. Therefore, we observe a lot of missing entries either due to the viruses undergoing mutations or some inherent bias in the testing sample. For example, a sample tested in a particular laboratory may miss out on some emerging strains which might be crucial to identify and predict future circulating viruses.

   In order to predict rates of circulation of viruses in the future, we propose to approximate the observed data with a low-rank tensor. The intuition behind a low-rank approximation is that a small number of latent factors corresponding to different geographical locations and viruses influence the rates of circulation of viruses in a flu season. In the Netflix challenge, low-rank matrix factorization was found to produce most accurate results in comparison with other methods and that only a small number of latent factors influence the system Bell et al. (2007). We thus propose to estimate a low-rank tensor to approximate the observed data such that the low-rank component of the tensor varies slowly across time. The reason that we impose such a constraint on the low-rank component of the tensor is that we assume that the weights which influence the system do not change drastically across consecutive flu seasons.

2. **Optimal regression trees**: In order to quantify the efficacy of a vaccine, we train regression trees using the optimal regression trees (ORTs) algorithm proposed in Bertsimas and Dunn (2019) to predict both the morbidity and mortality rates as a percentage of the population using information about how effectively a vaccine-strain hinders a virus’ ability to attack healthy red-blood cells. For this purpose, we use a distance metric also called antigenic distance between a vaccine-strain and a virus for each subtype of the influenza virus, and define a weighted distance (see Cai et al., 2012) between a given vaccine-strain and all predominant circulating viruses as the antigenic distance between each of these pairs weighted by the rates of circulating of the corresponding virus during a flu season.

3. **Robust optimization**: Using the same weighted distance as a metric of performance, we propose a novel robust prescriptive problem that minimizes the worst-case weighted distance given some uncertainty about the rates of circulation of viruses in the upcoming flu season in order to inform vaccine composition. The uncertainty in the rates of circulation of viruses for the upcoming flu season is quantified by restricting the low-rank component of the tensor decomposition to not deviate from its counterpart from the previous time period. We reformulate the corresponding robust prescriptive problem as a Second Order Cone Program (SOCP), and show that it is both practically and theoretically tractable.

### 1.1 Related work

In this section, we review some related work in literature that propose methods to model evolution of the viruses in order to inform strain selection for the seasonal influenza vaccine.

Wilson and Cox (1990) studied evolution of various virus strains and suggested that a drift variant of epidemiologic importance usually contains at least four amino acid substitutions located at more than two of the epitope regions (part of an antigen molecule to which an antibody attaches itself) on the HA1 polypeptide. Lee and Chen (2004) showed that the number of amino acid changes in the 131 amino acid positions around the epitope sites had the highest correlation with the antigenic distance and the best performance for predicting antigenic difference between any two virus
Our contributions

The key contributions of this paper are summarized below.

1. We propose a novel tensor completion formulation that estimates a low-rank representation of the circulation rates of viruses across various regions around the world while restricting the low-rank component to not deviate much from its counterpart from a previous time period. Leveraging algorithms for tensor completion, we develop new algorithms to solve the corresponding restricted tensor completion problem.

2. We train optimal regression trees to predict both morbidity and mortality rates using weighted distances between a given vaccine-strain and all predominant circulating viruses for each subtype of the influenza virus during a flu season. The $R^2$ which is a measure of goodness-of-fit varied around 0.67–0.75 for regression trees predicting morbidity rates and 0.64–0.78 for regression trees predicting mortality rates.
3. We propose a set based uncertainty for the rates of circulation of viruses based on the low-rank component of the matrix factorization and formulate a robust prescriptive problem to choose vaccine composition that minimizes a worst-case weighted distance between the chosen vaccine-strains and viruses that are predicted to circulate in the future. We show that this problem can be reformulated as a SOCP and show that it is both practically and theoretically tractable.

4. Through a retrospective study, we illustrate the effectiveness of our approach in terms of a weighted distance between the vaccine chosen composition and observed predominant circulating viruses during a flu season in comparison to vaccine compositions proposed by WHO, FDA, and EMA. Using ORTs, we also show that the vaccine compositions from the robust prescriptive problem could potentially lower morbidity by 11–14% and mortality by 8–11% over vaccine compositions proposed by WHO for Northern Hemisphere, and similarly, morbidity by 8–10% and mortality by 6–9% over vaccine compositions proposed by FDA for United States, and finally, morbidity by 10–12% and mortality by 9–11% over vaccine compositions proposed by EMA for Europe.

1.3 Outline
The rest of the paper is structured as follows. In Section 2, we describe the data we used and the methods we employed for predicting the rates of circulation of viruses for the upcoming flu season given historical data. We also review the optimal regression trees algorithm which we use to train regression trees in order to quantify efficacy of a vaccine in terms of morbidity and mortality rates. In Section 3, we present a novel robust prescriptive formulation to optimize influenza vaccine composition given some uncertainty about the rates of circulation of viruses for an upcoming flu season. In Section 4, we present the influenza vaccine compositions prescribed by the robust prescriptive problem and compare them with the ones proposed by WHO, FDA, and EMA in terms of weighted distance between the vaccine-strains and the observed predominant viruses. Finally, in Section 5, we compare the efficacy of the vaccines using predictions about morbidity and mortality rates using ORTs introduced in Section 2.

1.4 Notation
Lowercase and uppercase bold letters denote vectors and matrices, respectively. For a tensor $\mathbf{M} \in \mathbb{R}^{p \times q \times r}$, we denote the slices of the tensor as $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^\tau \in \mathbb{R}^{p \times q}$. A tensor unfolding, is an operation which essentially flattens the tensor into a matrix. The mode-1 unfolding of a tensor $\mathbf{M} \in \mathbb{R}^{p \times q \times r}$, denoted by $\mathbf{M}_{(1)}$, is the $p \times qr$ matrix whose columns are the columns of $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^\tau$. Similarly the mode-2 unfolding, denoted by $\mathbf{M}_{(2)}$, is the $q \times pr$ matrix whose columns are the transposed rows of $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^\tau$. The Frobenius norm of a matrix $\mathbf{U} \in \mathbb{R}^{m \times n}$ denoted by $\| \mathbf{U} \|_F$ is given as $(\sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}^2)^{1/2}$. We denote the set $\{1, 2, \ldots, n\}$ by $[n]$.

2 DATA AND METHODS
In this section, we describe the data and methods used in our analysis to predict rates of circulation of predominant viruses in a flu season, and to predict morbidity and mortality rates given composition of an influenza vaccine and the observed rates of circulation of viruses in a flu season. In Section 2.1, we describe the data that we use in our analysis. In Section 2.2, we propose a novel tensor completion problem and present an algorithm to model evolution of rates of circulation of viruses through different seasons. Finally, in Section 2.3, we review ORTs which we use to train regression trees for predicting morbidity and mortality rates using factors such as weighted distance between the vaccine strain and circulating viruses for each subtype of the influenza virus.

2.1 Data
Here, we describe the data that we use to inform influenza vaccine composition. The WHO has a network of laboratories around the world that contribute to the GISRS system which monitors and tracks properties of predominant circulating viruses using a test called Hemagglutinin Inhibition (HI) test. Below, we provide a short description of data compiled through GISRS, NIAID Influenza Research Database (IRD), the Centers for Disease Control and Prevention (CDC), and the European Centers for Disease Control and Prevention (ECDC).

1. Rates of circulation of predominant viruses: The WHO monitors the rates of circulation of predominant viruses (along with their properties) at country level for each flu season and this data is available with GISRS system (WHO, 2013). The historical rates of circulation of viruses are available in the form of a three-dimensional tensor (see Figure 1) $\mathbf{M} \in \mathbb{R}^{n \times m \times T}$, where $n$ is the number of countries contributing to the WHO’s GISRS system, $m$ is the number of viruses in the data set and $T$ is the number of flu seasons. Such a three-dimensional matrix is known as a three-dimensional tensor. Each entry in the tensor is given by

$$
\mathbf{M}_{cv} = \frac{N_{cv}^t}{\sum_{v=1}^{m} N_{cv}^t}, \quad c \in [n], v \in [m], t \in [T],
$$

where, $N_{cv}^t$ is the number of observed cases of virus $v$ found in tests performed in country $c$ for flu season $t$. We denote the set of virus-strains belonging to influenza A (H1N1) as $D_{H1N1}$, and similarly, for influenza A (H3N2) and B as $D_{H3N2}$ and $D_B$ respectively. Each slice of the tensor representing a flu season from 1987 until 2018 ($T = 32$), and each slice consists of observed rates of circulation of $m = 1206$ viruses in about $n = 132$ countries.
Formulation

2. Estimates for morbidity and mortality rates due to influenza, and the corresponding flu vaccine compositions:
We compiled various estimates of morbidity and mortality rates for United States from the CDC website, for Europe from the ECDC website, and for Northern and Southern hemispheres from GISRS and Iuliano et al. (2018). The data for influenza vaccine composition over various years from 1987 until 2018 was obtained from the NIAID Influenza Research Database (IRD) (Bao et al., 2008).

3. Antigenic properties of predominant circulating viruses:
At each of the laboratories that contribute to WHO’s GISRS, circulating viruses during a season are subject to a HI (Hemagglutinin Inhibition) test which measures the ability of the antibodies (injected by a vaccine) to block the Hemagglutinin (HA) protein of the virus being tested from attacking healthy red blood cells. This data was obtained from IRD and ImmPort (Bhattacharya et al., 2018). For each virus-strain $u$ and vaccine-strain $v$, we have a corresponding HI value denoted by $h_{uv}$, where $u$ belongs to either of $D_{H1N1}, D_{H3N2}$, or $D_B$. Then, a distance (Cai et al., 2012) between virus $u$ and vaccine-strain $v$ can be defined as follows,

$$d_{uv} = \log \left( \max_u (h_{uv}) / h_{uv} \right),$$

where, $\max_u (h_{uv})$ is the maximum HI value for vaccine-strain $v$ across all the viruses in the data set. Multiple measurements of virus-strain and vaccine-strain distances are available when antibodies raised against a viral strain are tested in multiple laboratories or at several time points, or when multiple antibodies are raised against the same strain. The resulting antigenic data set comprises of distances between $p = 1,377$ viruses and $q = 82$ vaccine strains.

2.2 Tensor completion

Here, we present our tensor completion formulation to model the evolution of rates of circulation of viruses across different geographies and flu seasons.

2.2.2 Tensor unfolding and alternating optimization

To estimate the latent spaces $U$ and $V$, we use the algorithm proposed in Bertsimas and Pawlowski (2019), which is based on Farias and Li (2019). In the first step, we construct the mode-1 unfolding $M_{(1)} \in \mathbb{R}^{n \times mT}$, which is an $n \times mT$ matrix whose columns are the columns of $M^1, M^2, \ldots, M^T$ (the order of the columns does not matter). We then compute $\hat{U}$ as the first $r$ left singular vectors of $M_{(1)}$. More precisely, assuming that $M_{(1)}$ admits the singular value decomposition $M_{(1)} = U_1 \Sigma_1 V_1^\top$, we set $\hat{U}$ to be the columns of $U_1$ corresponding to the $r$ largest singular values. See Figure 2 for an example. We denote this entire procedure with the shorthand

$$\hat{U} = \text{SVD } (M_{(1)}, r).$$

To estimate $V$, we apply a similar procedure using the mode-2 unfolding $M_{(2)}$, which is the $m \times nT$ matrix whose columns are the transposed rows of $M^1, M^2, \ldots, M^T$. Therefore, the estimate of $V$ is given by the $r$ largest singular vectors of mode-2 unfolding $M_{(2)}$ and we denote this procedure as follows

$$\hat{V} = \text{SVD } (M_{(2)}, r).$$
Given some estimates \(\{\hat{U}, \hat{V}\}\). Problem (3) reduces to a Second Order Cone Program (SOCP) in \(\{S^j\}_{j=1}^T\) as follows:

\[
\min_{\{S^j\}_{j=1}^T} \sum_{j=1}^T \| M^j - \hat{US}^j\hat{V}^T \|_F
\]

\[
\text{s.t. } \| S^j - S^{(j-1)} \|_F \leq \lambda, \quad t \in \{2, 3, \ldots, T\},
\]

Problem (4) is a SOCP in \(T^2\) variables and is tractable. Putting all of this together, we present Algorithm 1 to solve Problem (3) for a given value of rank \(r\), parameter \(\lambda\) and maximum number of iterations \(K\). We evaluate \(r\) and \(\lambda\) through cross validation.

**Algorithm 1.** Algorithm for tensor completion problem (3)

**Input:** Incomplete tensor \(X \in \mathbb{R}^{m \times n \times T}\), Rank \(r\), parameter \(\lambda\) and max iterations \(K\).

1: procedure TENSORCOMPLETION \((X, r, \lambda, K)\)
2: \(\hat{Y}_0 \leftarrow X\) and \(k \leftarrow 1\).
3: while \(k < K\) do
4: \(\hat{U}_k = \text{SVD}\left(\hat{Y}_{k-1,1}\right), r\)
5: \(\hat{V}_k = \text{SVD}\left(\hat{Y}_{k-1,2}, r\right)\)
6: \(\left\{\hat{S}^k\right\}_{j=1}^T = \text{Solve Problem (4) with estimates} (\hat{U}_k, \hat{V}_k)\) and parameter \(\lambda\)
7: \(\hat{Y}_k = \hat{U}_k\hat{S}^k\hat{V}_k, \quad t \in [T]\)
8: \(k = k + 1\)
9: end while
10: return \(\left(\hat{U}_K, \left\{\hat{S}^k\right\}_{j=1}^T, \hat{V}_k\right)\).
11: end procedure

### 2.3 Optimal regression trees

In order to predict the morbidity and mortality rates given a vaccine composition, we use a novel algorithm optimal regression trees (ORTs) proposed in Bertsimas and Dunn (2019) to train predictive trees that combine state-of-the-art performance (at par with gradient boosted trees) and interpretability. Such tree structures are based on a few decision splits on variables of high importance, and can readily model nonlinearities and interactions between variables.

To predict morbidity and mortality rates, we used the following predictors that quantify the ability of the vaccine-strains to hinder viruses’ ability to attack healthy red blood cells,

1. Weighted distance between influenza A (H1N1) strain and the corresponding circulating viruses: \(w_{\text{H1N1}} = \sum_{u \in D_{\text{H1N1}}} d_{uv} r_u\), where, \(v\) is the chosen vaccine-strain, \(r_u\) is the normalized rate of circulation of virus \(u\) among all predominant viruses in a particular flu season.
2. Weighted distance between influenza A (H3N2) strain and the corresponding circulating viruses: \(w_{\text{H3N2}} = \sum_{u \in D_{\text{H3N2}}} d_{uv} r_u\).
3. Weighted distance between influenza B strain and the corresponding circulating viruses: \(w_B = \sum_{u \in D_B} d_{uv} r_u\).

During the training process, we tuned the parameters to maximize performance on a separate holdout set to avoid overfitting.

In Figure 3, we present optimal regression trees that were trained on data from United States from 1988 until 2018 to predict morbidity and mortality rates using weighted distances between the vaccine-strains proposed by FDA and the actual viruses that circulated during a particular flu season. The regression tree trained to predict morbidity has an accuracy (in terms of \(R^2\)) of 0.77 and the one for mortality has an accuracy of 0.75. In both the trees, \(w_{\text{H1N1}}\) appears as an important variable which is expected as it is well known that influenza A (H3N2) strains are highly volatile. Also, the positive coefficients of the weights \(w_{\text{H1N1}}, w_{\text{H3N2}},\) and \(w_B\) in the leaves of the regression trees signify that higher weighted distances between the vaccine-strains and predominant circulating viruses results in higher morbidity and mortality rates.

### 3 Optimizing the Vaccine Composition

In this section, we propose a set based uncertainty for the rates of circulation of viruses based on the low-rank component of the matrix factorization and formulate a robust prescriptive problem to choose vaccine composition that minimizes the worst-case weighted distance between the chosen vaccine-strain and the viruses that are predicted to circulate in the future. In Section 3.1, we describe a nominal formulation for choosing vaccine formulation a particular geographical location (prescribing country level, or region level influenza vaccine compositions) and in Section 3.2, we present the robust prescriptive counterpart of the nominal model.

#### 3.1 Nominal model

Given estimates of antigenic distances \(d_{uv}\) (see Eq. (2)) between virus “\(u\)” and vaccine-strain “\(v\)” and let, \(Y = (y_{iu}) \in \mathbb{R}^{n \times q}\) denote the predicted rates of circulation of viruses, our goal is to choose a suitable vaccine-strain which has the smallest weighted distance with the predicted circulating viruses in a flu season. In order to select such a vaccine-strain to be included in the vaccine for some location \(i\), we propose to solve the following optimization problem:

\[
\min_{z \in \{0,1\}^p} \sum_{u=1}^q \sum_{i=1}^n z_{iu} d_{uv} y_{iu}
\]

\[
\text{s.t. } Y = US_{T-1}V^T,
\]
FIGURE 3 Optimal regression trees to predict morbidity (in millions) and mortality (in thousands) in United States for flu seasons from 1988 till 2018. The variables DistH1N1, DistH3N2 and DistB denote weighted distances between the predominant circulating viruses and influenza a (H1N1), a (H3N2) and B strains, respectively. (a) Regression tree for predicting morbidity rates ($R^2 : 0.77$), (b) regression tree for predicting mortality rates ($R^2 : 0.75$).

3.2 Robust prescriptive model

Here, we present the robust prescriptive problem to choose vaccine composition for a particular geographical location. Given a low-rank decomposition of the Tensor $M^t = US^TV^t$, $t \in [T]$ containing rates of circulating viruses in different locations, we formulate a robust optimization problem to choose a vaccine strain that is robust to mutations during the time period the vaccines are manufactured and distributed.

Observe that Problem (5) is very sensitive to the predicted rates of circulation of viruses. Therefore, we propose using a set based uncertainty to capture any noise in the predictions from the tensor model. We define an uncertainty set for rates of circulation of viruses as follows,

$$\mathcal{U}_i(U, \hat{S}, V) = \left\{ Y : Y = USV^T, \|S - \hat{S}\|_{F_2} \leq \lambda \right\}.$$  

Therefore, instead of solving the nominal problem (5), we propose to solve the following robust optimization problem,

$$P_i = \min_{\epsilon \in (0,1)^q} \max_{Y \in \mathcal{U}_i(U, \hat{S}, V)} \sum_{u=1}^{p} \sum_{v=1}^{q} \zeta_v d_{uv} y_{iu},$$

s.t. $\sum_{v=1}^{q} z_v = 1,$

where the inner minimization problem can be reformulated as follows:

$$P_i = \max_{Y \in \mathcal{U}_i(U, \hat{S}, V)} \min_{\epsilon \in (0,1)^q} \sum_{u=1}^{p} \sum_{v=1}^{q} \zeta_v d_{uv} y_{iu},$$

s.t. $\sum_{v=1}^{q} z_v = 1,$

where the term $\sum_{u=1}^{q} d_{uv} y_{iu}$ in the objective function represents a weighted distance between all viruses $u \in [p]$ predicted to be circulating with frequencies $\{y_{iu}\}_{u=1}^{p}$ and some vaccine-strain $v$.

Lemma 3.1 (Neumann (1928)). The min-max in Problem (6) can be interchanged, that is,

$$\min_{Y \in \mathcal{U}_i(U, \hat{S}, V)} \max_{\epsilon \in (0,1)^q} \sum_{u=1}^{p} \sum_{v=1}^{q} \zeta_v d_{uv} y_{iu} = \max_{Y \in \mathcal{U}_i(U, \hat{S}, V)} \min_{\epsilon \in (0,1)^q} \sum_{u=1}^{p} \sum_{v=1}^{q} \zeta_v d_{uv} y_{iu}.$$
\[ \| S - \hat{S} \|_2 \leq \lambda. \] (8)

Problem (8) is a Second Order Cone Program (SOCP) which can be further reduced by eliminating variables \( Y \), and can be solved using off-the-shelf solvers. The vaccine-strain prescribed by the model is given by \( \ell^\star \), where \( \ell^\star = \arg\min_{\ell} \sum_{i=1}^{p} d_{u_i} y_i^\star \), and the worst-case weighted distance of the proposed vaccine-strains to the circulating viruses is given by \( w^\star_i = \sum_{\mu=1}^{p} d_{u \mu} y_{i \mu}^\star \).

4 | RESULTS

In this section, we present the performance of vaccine compositions prescribed by the robust model in terms of predicted morbidity and mortality rates and compare it with that of WHO, FDA and EMA using optimal regression trees. The morbidity and mortality rates are predicted using weighted distances between the vaccine strains and the observed rates of circulation viruses for each subtype of the influenza virus.

4.1 | Morbidity and mortality rates

In Section 2, Figure 3 we present optimal regression trees that were trained on data compiled from FDA for flu seasons from 1988 until 2018 to predict morbidity and mortality rates using weighted distances between the vaccine-strains proposed by FDA and the actual viruses that circulated during a particular flu season. For this analysis, we trained multiple regression trees with a moving window size of 20 years starting from 1988 and used these models to predict morbidity and mortality rates for the next flu season given some vaccine compositions along with observed rates of circulation of viruses.

In Tables 1–4 we present a retrospective comparison of predicted mortality and morbidity rates based on the vaccine composition prescribed by WHO for Northern hemisphere and the CDC for United States with our prescriptions for flu seasons during 2009–2018 for the states of California, Texas, Florida, and New York respectively.

In Tables 5–7 we present a retrospective comparison of predicted mortality and morbidity rates based on the vaccine composition prescribed by WHO for Northern Hemisphere, FDA for United States and EMA for Europe with our prescriptions for flu seasons during 2009–2018 respectively. We compare the effectiveness of prescribed vaccine composition by training regression trees using optimal regression trees (ORTs) algorithm for predicting both morbidity and mortality using the following variables: (a) \( w_{H1N1} \), a weighted distance between influenza A (H1N1) strain and the corresponding circulating viruses, and similar weighted distances for influenza A (H3N2) and influenza B viruses (b) \( w_{H1N1} \), and (c) \( w_B \). During the training process, we tuned the parameters to maximize performance on a separate holdout set to avoid overfitting.
### TABLE 3  Retrospective 8-year comparison of number of severe illnesses and mortality under vaccine proposed CDC (Center for Disease Control and Prevention) versus robust prescriptive model for **Florida** using optimal regression trees (accuracy reported in terms of $R^2$)

| Season  | Severe illnesses (in millions) | Mortality (in thousands) |
|---------|-------------------------------|--------------------------|
|         | Observed CDC | Robust | Accuracy | Observed CDC | Robust | Accuracy |
| 2010–2011 | 3.33 | 3.25 | **3.02** | 0.75 | 2010–2011 | 3.38 | 3.41 | **3.31** | 0.65 |
| 2011–2012 | 3.19 | 3.22 | **2.49** | 0.73 | 2011–2012 | 3.87 | 3.04 | **3.23** | 0.69 |
| 2012–2013 | 3.55 | 3.22 | **3.1** | 0.77 | 2012–2013 | 3.51 | 3.36 | **3.28** | 0.71 |
| 2013–2014 | 2.42 | 3.44 | **2.85** | 0.79 | 2013–2014 | 3.1 | 3.37 | **2.92** | 0.71 |
| 2014–2015 | 3.58 | 2.64 | **2.5** | 0.68 | 2014–2015 | 3.07 | 2.94 | **2.91** | 0.77 |
| 2015–2016 | 2.89 | 3.7 | **2.69** | 0.65 | 2015–2016 | 3.3 | 3.23 | **3.01** | 0.74 |
| 2016–2017 | 3.65 | **2.78** | 3.26 | 0.67 | 2016–2017 | 3.3 | 4.11 | 3.52 | 0.79 |
| 2017–2018 | 3.39 | 3.36 | **3.33** | 0.76 | 2017–2018 | 3.68 | 3.19 | **3.07** | 0.78 |

### TABLE 4  Retrospective 8-year comparison of number of severe illnesses and mortality under vaccine proposed CDC (Center for Disease Control and Prevention) versus robust prescriptive model for **New York** using optimal regression trees (accuracy reported in terms of $R^2$)

| Season  | Severe illnesses (in millions) | Mortality (in thousands) |
|---------|-------------------------------|--------------------------|
|         | Observed CDC | Robust | Accuracy | Observed CDC | Robust | Accuracy |
| 2010–2011 | 2.25 | 1.99 | **1.99** | 0.73 | 2010–2011 | 2.67 | 2.81 | **2.51** | 0.75 |
| 2011–2012 | 2.22 | 2.13 | **1.96** | 0.69 | 2011–2012 | 2.42 | 2.21 | **2.08** | 0.68 |
| 2012–2013 | 2.43 | 2.58 | **1.91** | 0.7 | 2012–2013 | 2.66 | 2.84 | 2.74 | 0.76 |
| 2013–2014 | 2.41 | 2.41 | **1.98** | 0.8 | 2013–2014 | 2.79 | 3.13 | 3.09 | 0.67 |
| 2014–2015 | 2.2 | 2.4 | **2.17** | 0.78 | 2014–2015 | 2.64 | 2.58 | **2.51** | 0.68 |
| 2015–2016 | 2.01 | 2.2 | 2.16 | 0.73 | 2015–2016 | 2.36 | 2.56 | 2.48 | 0.64 |
| 2016–2017 | 2.5 | 2.16 | **1.81** | 0.77 | 2016–2017 | 2.64 | 2.39 | **2.05** | 0.7 |
| 2017–2018 | 1.95 | 2.51 | 2.24 | 0.73 | 2017–2018 | 2.23 | 2.47 | 2.42 | 0.75 |

### TABLE 5  Retrospective 8-year comparison of number of illnesses and mortality under vaccine proposed WHO versus robust prescriptive model for **Northern Hemisphere** using optimal regression trees (accuracy reported in terms of $R^2$)

| Season  | Illnesses (in millions) | Mortality (in thousands) |
|---------|-------------------------|--------------------------|
|         | Observed WHO | Robust | Accuracy | Observed WHO | Robust | Accuracy |
| 2010–2011 | 3.2 | 3.27 | **2.71** | 0.68 | 2010–2011 | 427 | 432.4 | **378.7** | 0.74 |
| 2011–2012 | 2.6 | 3.52 | **2.84** | 0.68 | 2011–2012 | 362 | 435.1 | **392.3** | 0.73 |
| 2012–2013 | 4.2 | 3.37 | 3.37 | 0.71 | 2012–2013 | 436 | 434.3 | 434.3 | 0.76 |
| 2013–2014 | 3.9 | 3.73 | **3.26** | 0.69 | 2013–2014 | 438 | 481.7 | **417.5** | 0.78 |
| 2014–2015 | 3.7 | 3.58 | **3.43** | 0.72 | 2014–2015 | 451 | 466.8 | **438.3** | 0.72 |
| 2015–2016 | 3.6 | 3.34 | **3.11** | 0.74 | 2015–2016 | 463 | 434.1 | 434.1 | 0.77 |
| 2016–2017 | 4.4 | 3.92 | **3.68** | 0.71 | 2016–2017 | 482 | 504.2 | **468.3** | 0.7 |
| 2017–2018 | 4.2 | 4.56 | **4.13** | 0.67 | 2017–2018 | 461 | 483.6 | **451.8** | 0.73 |
| 2018–2019 | 4.19 | 3.69 | **3.29** | 0.78 | 2018–2019 | 457 | 462.6 | **441.4** | 0.69 |

### TABLE 6  Retrospective 8-year comparison of number of severe illnesses and mortality under vaccine proposed FDA versus robust prescriptive model for **United States** using optimal regression trees (accuracy reported in terms of $R^2$)

| Season  | Severe illnesses (in millions) | Mortality (in thousands) |
|---------|-------------------------------|--------------------------|
|         | Observed FDA | Robust | Accuracy | Observed FDA | Robust | Accuracy |
| 2010–2011 | 21 | **24.4** | 26.1 | 0.78 | 2010–2011 | 37 | **36.7** | 37.1 | 0.72 |
| 2011–2012 | 9.3 | 24.7 | **20.8** | 0.73 | 2011–2012 | 12 | 31.4 | **29.6** | 0.76 |
| 2012–2013 | 34 | 30.6 | **28.2** | 0.75 | 2012–2013 | 43 | 39.2 | **36.7** | 0.71 |
| 2013–2014 | 30 | 29.3 | **26.4** | 0.81 | 2013–2014 | 38 | 40.3 | **38.4** | 0.7 |
| 2014–2015 | 30 | 32.7 | **28.5** | 0.83 | 2014–2015 | 51 | 44.2 | **27.4** | 0.73 |
| 2015–2016 | 25 | 28.7 | **25.9** | 0.77 | 2015–2016 | 25 | 35.8 | **33.1** | 0.74 |
| 2016–2017 | 30 | 31.6 | **28.8** | 0.75 | 2016–2017 | 51 | 43.8 | **41.6** | 0.71 |
| 2017–2018 | 49 | 34.6 | **31.7** | 0.77 | 2017–2018 | 79 | 49.6 | **46.4** | 0.75 |
| 2018–2019 | 35.5 | 37.6 | **32.9** | 0.74 | 2018–2019 | 34.7 | 34.8 | **32.9** | 0.77 |
In Table 5 we compare the effectiveness of vaccine compositions for Northern Hemisphere against WHO. We observed that in seven of the eight flu seasons, vaccine prescribed by the robust model had a smaller number of morbidity and mortality cases. On average, vaccine compositions prescribed by the robust model could potentially lower morbidity by 11–14% and mortality by 8–11% over vaccine compositions proposed by WHO for Northern Hemisphere.

Similarly, in Table 6 we compare the effectiveness of vaccine compositions for United States. Again, we observed that in seven of the eight flu seasons, vaccine compositions prescribed by the robust model had a smaller number of morbidity and mortality cases, and in only one season, we had greater morbidity and mortality cases. Vaccine compositions prescribed by the robust model could potentially lower morbidity by 8–10% and mortality by 6–9% over the ones proposed by FDA for United States.

Finally, in Table 7 we compare the effectiveness vaccine compositions for Europe. We observed that in seven of the eight flu seasons, vaccine compositions prescribed by the robust model had a smaller number of morbidity and mortality cases. Through ORTs, we show that vaccine compositions prescribed by the robust model could potentially lower morbidity by 10–12% and mortality by 9–11% over the ones proposed by EMA for Europe.

### 4.2 Prescribed vaccine formulations

Here, we compare the performance of the flu vaccine compositions prescribed by the robust model containing a type A (H1N1) virus in Section 4.2.1, a type A (H3N2) virus in Section 4.2.2, and a type B virus in Section 4.2.3 with the ones prescribed by WHO for Northern Hemisphere, FDA for United States, and EMA for Europe in terms of weighted distance $w_{\text{H1N1}}$.

In Table A1 we present vaccine strains prescribed by WHO for Northern Hemisphere, FDA for United States and EMA for Europe with the ones prescribed by the robust model, respectively for the flu seasons during 2009–2018 in terms of a weighted distance $w_{\text{H1N1}}$.

In Table 5 we compare the effectiveness of vaccine compositions for Northern Hemisphere against WHO. We observed that in four of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in four other seasons, the robust model prescribed the same vaccine strains as WHO.

In Table A4 we compare vaccine strains prescribed by the robust model with that of FDA for United States. We observe that in 3 of the 10 flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in five other flu seasons, the robust model and FDA proposed exactly same strains. Finally, in Table A5, we present vaccine strains prescribed by the robust model with that of EMA for Europe. We observe that in 5 of the 10 flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating virus-strains and in four other seasons, the robust model prescribed the same vaccine strains as EMA.

#### 4.2.2 Influenza type A (H3N2) virus

In Tables A2, A6 and A7 (in Appendix) we present a retrospective comparison of the H3N2-like virus strain as proposed by WHO for Northern Hemisphere, FDA for United States and EMA for Europe with the ones prescribed by the robust model, respectively for the flu seasons during 2009–2018 in terms of a weighted distance $w_{\text{H3N2}}$.

In Table A2 we compare vaccine strains prescribed by the robust model with that of WHO for Northern Hemisphere. We observe that in 6 of the 10 flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in two other seasons, the robust model prescribed the same vaccine strains as WHO.

In Table A6 we compare vaccine strains prescribed by the robust model with that of FDA for United States. We observed...
that in 5 of the 10 flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in three other flu seasons, the robust model prescribed the same vaccine strains as FDA. Finally, in Table A7, we compare vaccine strains prescribed by the robust model with that of EMA for Europe. We observed that in 6 of the 10 flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating virus-strains and in two other seasons, the robust model prescribed exactly the same vaccine strains as EMA.

4.2.3 | Influenza type B virus

We present a retrospective comparison of the influenza B vaccine strain chosen by WHO, FDA and EMA with the ones prescribed by the robust model for the flu seasons during 2009–2018 in Tables A3, A8 (in Appendix), respectively in terms of a weighted distance $w_B$. We observed that in 6 of the 10 flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in three other seasons, the robust model and WHO proposed exactly same strains. In Table A8, we compare vaccine strains prescribed by the robust model with that of FDA for United States. We observed that in 6 of the 10 flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the actual circulating viruses and in three other flu seasons, the robust model prescribed exactly same strains as FDA.

5 | CONCLUSIONS

In this paper, we have proposed a holistic framework employing state-of-the-art methods in machine learning and optimization to prescribe influenza vaccine composition based on historical data regarding circulating viruses in the population compiled through WHO’s Global Influenza Surveillance and Response System (GISRS). Specifically, we proposed a novel tensor completion formulation that restricts low-rank component in the matrix factorization to not deviate from its counterpart from a previous time period. Using the estimates from tensor completion, we formulated a robust optimization problem to prescribe vaccine composition that is robust to rates of circulation of the viruses in region using a set based uncertainty on the low-rank component. Finally, we trained optimal regression trees to predict both morbidity and mortality rates using weighted distances between vaccine strains and circulating viral strains during a flu season. Through various numerical experiments, we showed that our proposed vaccine compositions could potentially lower morbidity and mortality by 11–14%, 8–11% respectively over vaccine compositions proposed by WHO for Northern Hemisphere, and similarly, morbidity and mortality by 8–10%, 6–9% over vaccine compositions proposed by FDA for United States, and finally, morbidity and mortality by 10–12%, 9–11% respectively over vaccine compositions proposed by EMA for Europe.

DATA AVAILABILITY STATEMENT

We will provide all data we used.

ORCID

Hari Bandi https://orcid.org/0000-0002-6220-1567
Dimitris Bertsimas https://orcid.org/0000-0002-1985-1003

REFERENCES

Agor, J. K., & Özaltm, O. Y. (2018). Models for predicting the evolution of influenza to inform vaccine strain selection. Human Vaccines & Immunotherapeutics, 14(3), 678–683.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., & Lipman, D. (2008). The influenza virus resource at the national center for biotechnology information. Journal of Virology, 82(2), 596–601.

Bell, R. M., Koren, Y., & Volinsky, C. (2007). The bellkor solution to the netflix prize. KorBell Team’s Report to Netflix. Retrieved from https://netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

Bertsimas, D., & Dunn, J. (2019). Machine learning under a modern optimization lens. Belmont: Dynamic Ideas.

Bertsimas, D. & Pawlowski, C. (2019). Tensor completion with noisy side information for the prediction of anti-cancer drug response, Retrieved from https://dbertsim.mit.edu/pdfs/papers/2020-colin-jmlr.pdf

Bhattacharya, S., Dunn, P., Thomas, C. G., Smith, B., Schaefer, H., Chen, J., Hu, Z., Zalouscys, K. A., Shankar, R. D., Shen-Orr, S. S., Thomson, E., Wiser, J., & Butte, A. J. (2018). Immport, toward repurposing of open access immunological assay data for translational and clinical research. Scientific Data, 5, 180015.

Cai, Z., Zhang, T., & Wan, X.-F. (2012). Antigenic distance measurements for seasonal influenza vaccine selection. Vaccine, 30(2), 448–453.

Cho, S.-H. (2010). The optimal composition of influenza vaccines subject to random production yields. Manufacturing & Service Operations Management, 12(2), 256–277.

Farias, V. F., & Li, A. A. (2019). Learning preferences with side information. Management Science, 65, 3131–3149.

He, J., & Deem, M. W. (2010). Low-dimensional clustering detects incipient dominant influenza strain clusters. Protein Engineering, Design & Selection, 23(12), 935–946.

Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., Cohen, C., Gran, J. M., Schanzer, D., Cowling, B. J., Wu, P., Kyncl, J., Ang, L. W., Park, M., Redlberger-Fritz, M., Yu, H., Espenhain, L., Krishnan, A., Emukule, G., … Mustaqim, D. (2018). Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. The Lancet, 391(10127), 1285–1300.

Kornish, L. J., & Keeney, R. L. (2008). Repeated commit-or-defer decisions with a deadline: The influenza vaccine composition. Operations Research, 56(3), 527–541.

Lee, M.-S., & Chen, J. S.-E. (2004). Predicting antigenic variants of influenza a/h3n2 viruses. Emerging Infectious Diseases, 10(8), 1385–1390.
Liao, Y.-C., Lee, M.-S., Ko, C.-Y., & Hsiung, C. A. (2008). Bioinformatics models for predicting antigenic variants of influenza a/h3n2 virus. *Bioinformatics, 24*(4), 505–512.

Luksza, M., & Lassig, M. (2014). A predictive fitness model for influenza. *Nature, 507*(7490), 57–61.

Neher, R. A., & Bedford, T. (2015). Nextflu: Real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics, 31*(21), 3546–3548.

Neher, R. A., Russell, C. A., & Shraiman, B. I. (2016). Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences, 113*(12), E1701–E1709.

Neher, R. A., Russell, C. A., & Shraiman, B. I. (2014). Predicting evolution from the shape of genealogical trees. *eLife, 3*, e03568.

Neumann, J. V. (1928). The theory of social games. *Mathematical Annals, 100*(1), 295–320.

Özaltın, O. Y., Prokopyev, O. A., & Schaefer, A. J. (2018). Optimal design of the seasonal influenza vaccine with manufacturing autonomy. *INFORMS Journal on Computing, 30*(2), 371–387.

Perelson, A. S., & Oster, G. F. (1979). Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of Theoretical Biology, 81*(4), 645–670.

Steinbrück, L., Klingen, T., & McHardy, A. (2014). Computational prediction of vaccine strains for human influenza a (h3n2) viruses. *Journal of Virology, 88*(20), 12123–12132.

Steinbrück, L., & McHardy, A. C. (2012). Inference of genotype–phenotype relationships in the antigenic evolution of human influenza a (h3n2) viruses. *PLoS Computational Biology, 8*(4), e1002492.

Wilson, I. A., & Cox, N. J. (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annual Review of Immunology, 8*(1), 737–787.

World Health Organization, et al. (2013). *Global influenza surveillance and response system*. In Summary of neuraminidase amino acid substitutions associated with reduced inhibition by neuraminidase inhibitors. NAI.

Wu, J. T., Wein, L. M., & Perelson, A. S. (2005). Optimization of influenza vaccine selection. *Operations Research, 53*(3), 456–476.

### How to cite this article: Bandi H, Bertsimas D. Optimizing influenza vaccine composition: A machine learning approach. *Naval Research Logistics* 2021;68:857–870. https://doi.org/10.1002/nav.21974

### APPENDIX A: A VACCINE COMPOSITIONS

See Tables A1–A8.

#### TABLE A1 Retroactive 10-year comparison of vaccine-strains (for influenza type A (H1N1)-like virus) proposed by WHO versus robust prescriptive model for Northern Hemisphere

| Year      | Weighted distance | Proposed vaccine WHO | Proposed vaccine Robust |
|-----------|-------------------|-----------------------|-------------------------|
| 2010–2011 | 3.64              | A/California/07/2009  | A/California/07/2009    |
| 2011–2012 | 3.38              | A/California/07/2009  | A/California/07/2009    |
| 2012–2013 | **3.19**          | A/Victoria/361/2011   | A/Victoria/210/2009    |
| 2013–2014 | 3.49              | A/Hong Kong/4801/2014 | A/Singapore/INFIMH-16-0019/2016 |
| 2014–2015 | 3.03              | A/California/07/2009  | A/California/07/2009    |
| 2015–2016 | 3.18              | A/California/07/2009  | A/California/07/2009    |
| 2016–2017 | 3.67              | A/California/07/2009  | A/California/07/2009    |
| 2017–2018 | 3.71              | A/Michigan/45/2015    | A/Michigan/45/2015     |
| 2018–2019 | **3.37**          | A/Michigan/45/2015    | A/Michigan/45/2015     |

Bold values identify a lower weighted distance whenever the columns WHO and Robust do not match.

#### TABLE A2 Retroactive 10-year comparison of vaccine-strains (for influenza type A (H3N2)-like virus) proposed by WHO versus robust prescriptive model for Northern Hemisphere

| Year      | Weighted distance | Proposed vaccine WHO | Proposed vaccine Robust |
|-----------|-------------------|-----------------------|-------------------------|
| 2010–2011 | 3.86              | A/Perth/16/2009       | A/California/7/2009     |
| 2011–2012 | 3.41              | A/Perth/16/2009       | A/Perth/16/2009         |
| 2012–2013 | **3.32**          | A/Victoria/361/2011   | A/Victoria/210/2009    |
| 2013–2014 | 3.65              | A/Victoria/361/2011   | A/Texas/50/2012        |
| 2014–2015 | 3.42              | A/Texas/50/2012       | A/Wisconsin/15/2009    |
| 2015–2016 | 3.15              | A/Switzerland/9715293/2013 | A/Norway/466/2014 |
| 2016–2017 | 3.54              | A/Hong Kong/4801/2014 | A/Hong Kong/4801/2014 |
| 2017–2018 | 4.08              | A/Hong Kong/4801/2014 | A/Singapore/INFIMH-16-0019/2016 |
| 2018–2019 | **3.36**          | A/Singapore/INFIMH-16-0019/2016 | A/Switzerland/8060/2017 |

Bold values identify a lower weighted distance whenever the columns WHO and Robust do not match.
TABLE A3  Retrospective 10-year comparison of vaccine-strains (for influenza type B virus) proposed by WHO versus robust prescriptive model for Northern Hemisphere

| Year       | Weighted distance | Proposed vaccine | Robust          |
|------------|-------------------|------------------|-----------------|
|            | WHO   | Robust           | WHO             | Robust          |
| 2010–2011  | 3.53  | 3.53             | B/Brisbane/60/2008 | B/Brisbane/60/2008 |
| 2011–2012  | 3.34  | 3.17             | B/Brisbane/60/2008 | B/Wisconsin/01/2010 |
| 2012–2013  | 3.21  | 3.21             | B/Wisconsin/01/2010 | B/Wisconsin/01/2010 |
| 2013–2014  | 3.43  | 3.22             | B/Massachusetts/02/2012 | B/Massachusetts/02/2012 |
| 2014–2015  | 3.31  | 3.14             | B/Massachusetts/02/2012 | B/Brisbane/33/2008 |
| 2015–2016  | 3.24  | 3.12             | B/Phuket/3073/2013 | B/Brisbane/9/2014 |
| 2016–2017  | 3.23  | 3.40             | B/Brisbane/60/2008 | B/Utah/09/2014 |
| 2017–2018  | 3.54  | 3.15             | B/Brisbane/60/2008 | B/Phuket/3073/2013 |
| 2018–2019  | 3.42  | 3.16             | B/Phuket/3073/2013 | B/Colorado/06/2017 |

Bold values identify a lower weighted distance whenever the columns WHO and Robust do not match.

TABLE A4  Retrospective 10-year comparison of vaccine-strains (for H1N1-like virus) proposed by FDA versus our model for United States

| Year       | Weighted distance | Proposed vaccine | Robust          |
|------------|-------------------|------------------|-----------------|
|            | FDA   | Robust           | FDA             | Robust          |
| 2010–2011  | 3.34  | 3.34             | A/California/07/2009 | A/California/07/2009 |
| 2011–2012  | 3.09  | 3.09             | A/California/07/2009 | A/California/07/2009 |
| 2012–2013  | 3.27  | 3.14             | A/California/07/2009 | A/South Australia/55/2014 |
| 2013–2014  | 3.59  | 3.37             | A/California/07/2009 | A/Wyoming/03/2010 |
| 2014–2015  | 3.48  | 3.48             | A/California/07/2009 | A/California/07/2009 |
| 2015–2016  | 3.62  | 3.62             | A/California/07/2009 | A/California/07/2009 |
| 2016–2017  | 3.30  | 3.06             | A/California/07/2009 | A/Brisbane/10/2012 |
| 2017–2018  | 3.17  | 3.53             | A/Michigan/45/2015 | A/Brisbane/10/2007 |
| 2018–2019  | 3.26  | 3.26             | A/Michigan/45/2015 | A/Michigan/45/2015 |

Bold values identify a lower weighted distance whenever the columns WHO and Robust do not match.

TABLE A5  Retrospective 10-year comparison of vaccine-strains (for H1N1-like virus) proposed by EMA versus robust prescriptive model for Europe

| Year       | Weighted distance | Proposed vaccine | Robust          |
|------------|-------------------|------------------|-----------------|
|            | EMA   | Our model        | EMA             | Our model       |
| 2010–2011  | 3.64  | 3.64             | A/California/07/2009 | A/California/07/2009 |
| 2011–2012  | 3.38  | 3.38             | A/California/07/2009 | A/California/07/2009 |
| 2012–2013  | 3.19  | 3.26             | A/California/07/2009 | A/Victoria/361/2011 |
| 2013–2014  | 3.49  | 3.29             | A/California/07/2009 | A/Victoria/361/2011 |
| 2014–2015  | 3.03  | 3.03             | A/California/07/2009 | A/Christchurch/16/2010 |
| 2015–2016  | 3.18  | 3.18             | A/California/07/2009 | A/California/07/2009 |
| 2016–2017  | 3.67  | 3.42             | A/California/07/2009 | A/Michigan/45/2015 |
| 2017–2018  | 3.71  | 3.44             | A/Michigan/45/2015 | A/Brisbane/10/2012 |
| 2018–2019  | 3.37  | 3.51             | A/Michigan/45/2015 | A/Wisconsin/67/2016 |

Bold values identify a lower weighted distance whenever the columns WHO and Robust do not match.
TABLE A6  Retrospective 10-year comparison of vaccine-strains (for H3N2-like virus) proposed by FDA versus robust prescriptive model for United States

| Year          | Weighted distance | Proposed vaccine          |  |  |
|---------------|------------------|---------------------------|  |  |
|               | FDA   | Robust | FDA   | Robust |
| 2010–2011     | 3.64  | 3.36   | A/Perth/16/2009 | A/California/7/2009 |
| 2011–2012     | 3.34  | 3.07   | A/Perth/16/2009 | A/Uruguay/716/2007 |
| 2012–2013     | 3.09  | 3.32   | A/Victoria/361/2011 | A/Wisconsin/15/2009 |
| 2013–2014     | 3.75  | 3.57   | A/Victoria/361/2011 | A/Texas/50/2012 |
| 2014–2015     | 3.46  | 3.46   | A/California/7/2009 | A/California/7/2009 |
| 2015–2016     | 3.23  | 3.12   | A/Switzerland/9715293/2013 | A/Norway/466/2014 |
| 2016–2017     | 3.21  | 3.06   | A/Hong Kong/4801/2014 | A/Stockholm/6/2014 |
| 2017–2018     | 2.92  | 2.92   | A/Singapore/INFIMH-16-0019/2016 | A/Singapore/INFIMH-16-0019/2016 |
| 2018–2019     | 3.04  | 3.04   | A/Singapore/INFIMH-16-0019/2016 | A/Singapore/INFIMH-16-0019/2016 |

Bold values identify a lower weighted distance whenever the columns WHO and Robust do not match.

TABLE A7  Retrospective 10-year comparison of vaccine-strains (for H3N2-like virus) proposed by EMA (European Centre for Disease Prevention and Control) versus robust prescriptive model for Europe

| Year          | Weighted distance | Proposed vaccine          |  |  |
|---------------|------------------|---------------------------|  |  |
|               | EMA  | Robust | EMA  | Robust |
| 2010–2011     | 3.46  | 3.29   | A/Perth/16/2009 | A/California/7/2009 |
| 2011–2012     | 3.53  | 3.53   | A/Perth/16/2009 | A/Perth/16/2009 |
| 2012–2013     | 3.21  | 3.06   | A/Massachusetts/2/2012 | A/Victoria/210/2009 |
| 2013–2014     | 3.67  | 3.39   | A/Victoria/361/2011 | A/Wisconsin/1/2010 |
| 2014–2015     | 3.35  | 3.03   | A/Texas/50/2012 | A/Wisconsin/15/2009 |
| 2015–2016     | 3.42  | 3.17   | A/Switzerland/9715293/2013 | A/Phuket/3073/2013 |
| 2016–2017     | 3.43  | 3.43   | A/Hong Kong/4801/2014 | A/Hong Kong/4801/2014 |
| 2017–2018     | 3.38  | 3.51   | A/Singapore/INFIMH-16-0019/2016 | A/Michigan/45/2015 |
| 2018–2019     | 3.35  | 3.12   | A/Singapore/INFIMH-16-0019/2016 | A/Switzerland/8060/2017 |

Bold values identify a lower weighted distance whenever the columns WHO and Robust do not match.

TABLE A8  Retrospective 10-year comparison of vaccine-strains (for influenza type B virus) proposed by FDA versus robust prescriptive model for United States

| Year          | Weighted distance | Proposed vaccine          |  |  |
|---------------|------------------|---------------------------|  |  |
|               | FDA   | Robust | FDA   | Robust |
| 2010–2011     | 3.64  | 3.64   | B/Brisbane/60/2008 | B/Brisbane/60/2008 |
| 2011–2012     | 3.38  | 3.21   | B/Brisbane/60/2008 | B/Wisconsin/01/2010 |
| 2012–2013     | 3.17  | 3.17   | B/Wisconsin/01/2010 | B/Wisconsin/01/2010 |
| 2013–2014     | 3.49  | 3.26   | B/Massachusetts/02/2012 | B/Malaysia/2506/2009 |
| 2014–2015     | 3.12  | 2.97   | B/Massachusetts/02/2012 | B/Brisbane/33/2008 |
| 2015–2016     | 3.18  | 3.04   | B/Phuket/3073/2013 | B/Brisbane/9/2014 |
| 2016–2017     | 3.67  | 3.42   | B/Brisbane/60/2008 | B/Utah/09/2014 |
| 2017–2018     | 3.71  | 3.44   | B/Brisbane/60/2008 | B/Phuket/3073/2013 |
| 2018–2019     | 3.37  | 3.51   | B/Phuket/3073/2013 | B/Colorado/06/2017 |

Bold values identify a lower weighted distance whenever the columns WHO and Robust do not match.