


# BMJ Open Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan

Chien-An Hu,<sup>1</sup> Chia-Ming Chen,<sup>2</sup> Yen-Chun Fang,<sup>3</sup> Shinn-Jye Liang,<sup>4</sup> Hao-Chien Wang,<sup>5</sup> Wen-Feng Fang,<sup>6,7</sup> Chau-Chyun Sheu,<sup>8,9</sup> Wann-Cherng Perng,<sup>10</sup> Kuang-Yao Yang,<sup>11,12</sup> Kuo-Chin Kao,<sup>13,14</sup> Chieh-Liang Wu,<sup>15,16</sup> Chwei-Shyong Tsai,<sup>3</sup> Ming-Yen Lin,<sup>1</sup> Wen-Cheng Chao <sup>16,17</sup> On behalf of TSIRC (Taiwan Severe Influenza Research Consortium)

**To cite:** Hu C-A, Chen C-M, Fang Y-C, *et al*. Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. *BMJ Open* 2020;**10**:e033898. doi:10.1136/bmjopen-2019-033898

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-033898>).

C-AH and C-MC contributed equally.

Received 27 August 2019  
Revised 17 December 2019  
Accepted 22 January 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

## Correspondence to

Dr Wen-Cheng Chao;  
cwc081@hotmail.com and  
Professor Ming-Yen Lin;  
linmy@mail.fcu.edu.tw

## ABSTRACT

**Objectives** Current mortality prediction models used in the intensive care unit (ICU) have a limited role for specific diseases such as influenza, and we aimed to establish an explainable machine learning (ML) model for predicting mortality in critically ill influenza patients using a real-world severe influenza data set.

**Study design** A cross-sectional retrospective multicentre study in Taiwan

**Setting** Eight medical centres in Taiwan.

**Participants** A total of 336 patients requiring ICU-admission for virology-proven influenza at eight hospitals during an influenza epidemic between October 2015 and March 2016.

**Primary and secondary outcome measures** We employed extreme gradient boosting (XGBoost) to establish the prediction model, compared the performance with logistic regression (LR) and random forest (RF), demonstrated the feature importance categorised by clinical domains, and used SHapley Additive exPlanations (SHAP) for visualised interpretation.

**Results** The data set contained 76 features of the 336 patients with severe influenza. The severity was apparently high, as shown by the high Acute Physiology and Chronic Health Evaluation II score (22, 17 to 29) and pneumonia severity index score (118, 88 to 151). XGBoost model (area under the curve (AUC): 0.842; 95% CI 0.749 to 0.928) outperformed RF (AUC: 0.809; 95% CI 0.629 to 0.891) and LR (AUC: 0.701; 95% CI 0.573 to 0.825) for predicting 30-day mortality. To give clinicians an intuitive understanding of feature exploitation, we stratified features by the clinical domain. The cumulative feature importance in the fluid balance domain, ventilation domain, laboratory data domain, demographic and symptom domain, management domain and severity score domain was 0.253, 0.113, 0.177, 0.140, 0.152 and 0.165, respectively. We further used SHAP plots to illustrate associations between features and 30-day mortality in critically ill influenza patients.

## Strengths and limitations of this study

- This study used machine learning to predict the mortality risk of critically ill influenza patients.
- We used a data set containing medical data of real-world practice at eight Taiwanese referral centres during an influenza epidemic.
- We employed extreme gradient boosting (XGBoost) to establish a prediction model with high accuracy and used domain-based feature importance and SHapley Additive exPlanations plots for visualised realisation to mitigate the concern of black-box issue.
- The number of subjects was relatively small, and large-scale studies are needed to validate our findings.

**Conclusions** We used a real-world data set and applied an ML approach, mainly XGBoost, to establish a practical and explainable mortality prediction model in critically ill influenza patients.

## BACKGROUND

Sepsis is a leading cause of death in the intensive care unit (ICU) worldwide and contributes to approximately 50% of hospital deaths in the USA.<sup>1 2</sup> A number of scoring systems have been developed to predict mortality in patients admitted to ICUs; however, the clinical application remains limited given that sepsis consists of diverse aetiologies and no single scoring system appears to be applicable in diverse patient populations.<sup>3-5</sup> For example, the Acute Physiology and Chronic Health Evaluation (APACHE) II score, a widely used severity scoring system in the ICU,<sup>6 7</sup> has been found to have a limited role

in a number of diseases and settings, including severe influenza, pancreatitis, post-cardiac surgery and burn injury.<sup>8–12</sup> Similarly, the pneumonia severity index (PSI) is currently a widely used scoring system to assess patients with pneumonia,<sup>13</sup> but PSI might underestimate severity in patients with influenza.<sup>8,9</sup> Currently, influenza infection remains a global health threat that is estimated to affect nearly five million people worldwide, resulting in 250 000 to 500 000 deaths in 2015.<sup>14</sup> Our previous studies, investigating 336 patients with severe influenza requiring ICU management, showed the impact of an influenza epidemic on the healthcare system due to an abrupt increase in patients with severe influenza and found that lung-protective ventilation as well as fluid balance were early predictors for 30-day mortality in patients with severe influenza.<sup>15,16</sup> These findings highlight the crucial need to develop an influenza-specific mortality prediction model for both the management of patients and the allocation of ICU resources during an influenza epidemic. Taiwan experienced an influenza epidemic in the spring of 2016,<sup>17</sup> and the Taiwan Severe Influenza Research Consortium was established to survey the management strategies of distinct clinical domains, including the ventilation domain and fluid balance domain, laboratory data domain, demographic and symptom domain, management domain and conventional ICU severity score domain including APACHE II and PSI. Current machine learning (ML) models have enabled us to use all of the collected variables among the aforementioned clinical domains to develop an influenza-specific mortality prediction system, which has the potential to be incorporated into a healthcare information system as an automated decision support system (DSS).<sup>18,19</sup> However, the lack of intuitional understanding of ML models is one of the main obstacles in the implementation of ML in the medical field.<sup>20</sup> Herein, we employed an ML approach to illustrate the domain-specific feature importance, applied a visualised interpretation of the importance of each feature and compared the accuracy of different ML models using a nationwide severe influenza data set.

## METHODS

### Subject enrolment

This multicentre retrospective cohort study enrolled patients admitted to the ICUs at eight tertiary referral centres in Taiwan for virology-confirmed influenza who were enrolled during an influenza epidemic. The reported influenza epidemic occurred between October 2015 and March 2016, and the diagnosis of influenza was confirmed by the Taiwan Centers for Disease Control based on the rapid influenza diagnostic test, reverse transcription-polymerase chain reaction or viral culture. The study was approved by the Institutional Review Boards of the eight participating hospitals. Written informed consent was waived owing to the minimal risk, and all patients' information was de-identified before analysis.

### Measurement of variables

A standardised case report form was used to collect data at the eight participating hospitals. Medical records were reviewed to obtain data, which included demographics, comorbidities, physiological data, laboratory tests and influenza-associated data. Importantly, given that the reported influenza epidemic was characterised by a high proportion of patients with acute respiratory distress syndrome (ARDS), we collected ventilatory parameters and daily fluid status; protective ventilation strategy and dry-lung strategy are key components of the fundamental management of ARDS.<sup>21,22</sup> We also collected severity scores, including APACHE II, which is one of the quality-assessment indicators used across ICUs in Taiwan, and PSI, a widely used scoring system for predicting mortality in patients with pneumonia.<sup>13,23</sup>

### Extreme gradient boosting

We used extreme gradient boosting (XGBoost), an ensemble machine learning method based on decision trees, to establish a prediction model for 30-day mortality using data within the first 7 days after admission to the ICU and to illustrate the feature importance. Gradient boosting is a technique employed in complex prediction models that involves iterative combinations of ensembles of weak prediction models into one strong learner.<sup>24</sup> XGBoost uses second-order Taylor series to approximate the value of the loss function and further reduces the likelihood of overfitting by application of regularisation.<sup>25</sup> In the setup of the hyperparameters, the optimal values were found by performing a grid search on possible value combinations of the parameters. The main fine-tuned parameters in the present study included number of trees ( $n_{\text{estimator}}=100$ ), learning rate ( $\eta=0.007$ ), minimal loss to expand on a leaf node ( $\gamma=0$ ), maximum tree depth ( $\text{max\_depth}=4$ ), subsample proportion ( $\text{subsample}=1$ ), ratio of the number of negative class samples to positive class samples ( $\text{scale\_pos\_weight}=263/73$ ) and minimum sum of instance weight needed in a child node ( $\text{min\_child\_weight}=1$ ). All the other parameters may remain at their default values (see online supplementary table 1 for detailed parameters).<sup>25</sup> Additionally, the ensemble of decision tree methods can be used to obtain a predictive model with high accuracy through sequential (boosting) or parallel (bagging) ensemble methods and to provide estimates of feature importance from a trained predictive model. In the present study, which used F scores in XGBoost, the relative importance of each variable was computed as the sum of Gini improvement among the corresponding splits within a tree averaged over all the trees. Moreover, we implemented SHapley Additive exPlanations (SHAP), which is a recent approach to explain the output of a machine learning model, to illustrate the individual feature-level impacts on the 30-day mortality.<sup>26</sup> In brief, SHAP is an additive feature attribution method that provides an explanation of the tree ensemble's overall impact in the form of particular feature contributions and is relatively consistent with human intuition. In the present study, the training set consisted of a randomly selected 80%

of the patients, and the testing set was composed of the remaining 20% of the patients. The model establishment was based on data from the training set, and the testing set was independent of the training process and was used only for performance evaluation after the establishment of the model. The same training and testing sets were used in all three machine learning models in the present study (online supplementary figure 1).

### Logistic regression and random forest

In addition to XGBoost, we also employed other ML models, including logistic regression (LR) and random forest (RF). LR is a widely used method in medicine and is used as an ML model for classification tasks; however, LR is based on the assumption that a linear relationship exists between the input variables and the outcomes.<sup>27</sup> With regards to XGBoost and RF, both models are tree-based classifiers; however, these two ML models have substantial differences in ensemble method: XGBoost is based on boost, whereas RF is based on bagging.<sup>24 28</sup> In detail, XGBoost is based on the ensemble of weak learners and is characterised by high bias and low variance.<sup>25</sup> In contrast, RF is designed as fully grown decision trees and is hence characterised by low bias and high variance.<sup>29</sup> In RF, `max_depth` was 4 and `n_estimators` was 100, while default values were used for the other parameters in RF and LR (see online supplementary table 1 for detailed parameters).

### Statistical analysis

Data were expressed as frequencies (percentages) for categorical variables and as means±SD or median (IQR) based on the test of normality for continuous variables. Kolmogorov-Smirnov test was applied to test the normality. Differences between the survivor and non-survivor groups were analysed using Student's t-test for continuous variables and Fisher's exact test for categorical variables. Mann-Whitney U test was used for the variable which is not normally distributed. For the interpretability of the ML approach, feature importance was used to quantify the variable importance of each variable, and the SHAP summary plot and partial SHAP dependency plots were used to illustrate the distribution of the variable importance of individual variables. The score of feature importance was the average gain across all splits of a feature's used in the construction of the model. The performance of XGBoost, RF and LR for predicting 30-day mortality was determined by using the area under the receiver operating characteristic (ROC) curve metric. The DeLong's test was used to compare the difference between two area under the curves (AUCs). Python V.3.6 was used in the present study.

## RESULTS

### Patient characteristics

A total of 336 patients with virology-proven influenza were enrolled, and 76 variables with complete data of these 336 patients were analysed. [Table 1](#) summarises patients' demographic characteristics and other relevant data. The

median age of patients was 61 (IQR, 53 to 69) years, and 62.8% were men (see online supplementary dataset for details). Given that the eight participating hospitals were all teaching hospitals, the enrolled patients had a high severity, including a high APACHE II score (22, IQR 17 to 29) and PSI score (118, IQR 88 to 151) as well as a low ratio of oxygen PaO<sub>2</sub> to fractional inspired oxygen ratio (PaO<sub>2</sub>/FiO<sub>2</sub> (fraction of inspired O<sub>2</sub>)) (107, IQR 65 to 159.2). To investigate factors associated with hospital mortality, we divided the 336 subjects into survivor and non-survivor groups according to mortality at 30 days. Compared with those in the survivor group, those in the non-survivor group were more likely to have a lower PaO<sub>2</sub>/FiO<sub>2</sub> (85.9 vs 111.2, *p*<0.01), higher PSI (146.5 vs 108, *p*<0.01), APACHE II (28 vs 21, *p*<0.01), serum C-reactive protein (16.3 vs 12.9 mg/dL, *p*=0.07), blood urea nitrogen (30 vs 20.8 mg/dL, *p*=0.01) and creatinine (1.5 vs 1.0 mg/dL, *p*=0.01) and were more likely to receive extracorporeal membrane oxygenation (27.8% vs 12.5%, *p*<0.01), haemodialysis (23.6% vs 7.2%, *p*<0.01) and usage of steroid (60.8% vs 42.3%, *p*=0.01) ([table 1](#)).

Given that patients with severe influenza infection were characterised by oxygenation failure, as evidenced by the low PaO<sub>2</sub>/FiO<sub>2</sub> and a high proportion (78.3%, 263/336) of ARDS, we specifically collected dynamic ventilator parameters and data regarding fluid status in this study ([table 2](#)). Compared with those in the survivor group, those in the non-survivor group appeared to require a high FiO<sub>2</sub>, high positive end-expiratory pressure (PEEP), high peak pressure ( $P_{peak}$ ), and a low tidal volume/predicted body weight ( $V_T/PBW$ ) and these trends tended to be apparent on day 3. The dynamic fluid data showed that a positive cumulative fluid balance on day 4 (3801.2±4128.8 vs 1347.3±3137.4 mL, *p*<0.01) and on day 7 (4500.7±4997.8 vs 506.8±4385 mL, *p*<0.01) was associated with high 30-day mortality ([table 2](#)). Taken together, these data highlight the critical role of dynamic ventilatory parameters and fluid status in critically ill influenza patients.

### Visualisation of feature importance

To provide clinicians with a straightforward understanding of feature importance, we categorised the top 30 features by clinical domain ([figure 1](#)). The cumulative feature importance of the fluid balance domain, ventilation domain, laboratory data domain, demographic and symptom domain, management domain and conventional severity score domain was 0.253, 0.113, 0.177, 0.140, 0.152 and 0.165, respectively. Moreover, to allow the visualised interpretation of the selected variables, we used SHAP to illustrate how these variables affect 30-day mortality ([figure 2](#)). As shown in [figure 2](#), the higher PSI and cumulative day-4 fluid balance were associated with a higher 30-day mortality, whereas a lower PaO<sub>2</sub>/FiO<sub>2</sub> was associated with a higher 30-day mortality, in critically ill influenza patients. SHAP can also be used to illustrate the impact of an individual feature on 30-day mortality. As shown in [figure 3A](#), PSI mainly contributed to high

**Table 1** Basic characteristics of the 336 patients categorised by 30-day mortality

	All n=336	Survivor n=264	Non-survivor n=72	P value
<b>Demographic data</b>				
Age (years)	61(53-69)	62 (55.8–72.3)	61 (51–69)	0.09
Sex	211 (62.8%)	164 (62.1%)	47 (65.3%)	0.68
Body mass index	24.5 (21.5–27.7)	24.33 (21.4–26.2)	24.71 (21.6–28.2)	0.01
<b>Disease severity</b>				
APACHE II	22 (17–29)	28 (22.0–35.0)	21 (15–27)	<0.01
PSI	118 (88–151)	146.5 (123.0–172.5)	108 (80–140)	<0.01
PaO <sub>2</sub> /FiO <sub>2</sub>	107 (65–159.2)	85.9 (55.4–130.4)	111.2 (68–174)	<0.01
<b>Comorbidities</b>				
Malignancy	50 (14.9%)	37 (14%)	13 (18.1%)	0.45
Congestive heart failure	37 (11%)	27 (10.2%)	10 (13.9%)	0.40
Diabetes mellitus type 2	102 (30.4%)	78 (29.5%)	24 (33.3%)	0.56
Chronic pulmonary diseases	31 (9.2%)	25 (9.5%)	6 (8.3%)	0.99
Chronic renal failure	24 (7.1%)	18 (6.8%)	6 (8.3%)	0.61
<b>Symptoms</b>				
Fever	239 (71.1%)	191 (72.3%)	48 (66.7%)	0.38
Myalgia	83 (24.7%)	67 (25.4%)	16 (22.2%)	0.65
Headache	17 (5.1%)	13 (4.9%)	4 (5.6%)	0.77
Haemoptysis	23 (6.8%)	20 (7.6%)	3 (4.2%)	0.43
<b>Complication of influenza</b>				
Pulmonary complication	327 (97.3%)	255 (96.6%)	72 (100%)	0.21
Neurological complication	7 (2.1%)	5 (1.9%)	2 (2.8%)	0.65
Myocarditis	17 (5.1%)	12 (4.5%)	5 (6.9%)	0.38
<b>Virology data</b>				
Rapid influenza diagnostic test	162 (48.2%)	127 (48.1%)	35 (48.6%)	0.99
Virus culture	102 (30.4%)	75 (28.4%)	27 (37.5%)	0.15
RT-PCR	266 (79.2%)	207 (78.4%)	59 (81.9%)	0.62
Influenza subtype A	255 (75.9%)	204 (77.3%)	51 (70.8%)	0.28
Influenza subtype B	27 (8%)	18 (6.8%)	9 (12.5%)	0.14
<b>Laboratory data</b>				
White blood cell counts (/ml)	8900 (6008–13 500)	8800 (5850–16 155)	8920 (6100–13 210)	0.52
Haemoglobin (mg/dl)	11.9 (9.6–13.7)	11.2 (9.2–13.8)	12 (9.7–13.5)	0.60
Platelet count (10 <sup>3</sup> /ml)	145 (102–202.2)	135 (81.3–196.0)	149 (107–204.8)	0.21
C reactive protein (mg/dl)	13.9 (6.1–22.6)	16.3 (7.495–24.935)	12.9 (6.0–20.8)	0.07
Blood urea nitrogen (mg/dl)	22 (14.4–43.3)	30 (19–54.5)	20.8 (13.9–40)	0.01
Creatinine (mg/dl)	1.1 (0.8–1.9)	1.5 (1.0–2.6)	1.0 (0.8–1.7)	0.01
Sodium (mg/dl)	137 (133–140)	138 (132–141)	137 (133–140)	0.55
Potassium (mg/dl)	3.9 (3.5–4.4)	4 (3.5–4.7)	3.9 (3.5–4.3)	0.20
<b>Management</b>				
First-dose oseltamivir (days)	0.6 (0.1–1.9)	0.5 (0.1–1.9)	0.6 (0.1–1.9)	0.77
ICU wait (days)	0.2 (0–1.0)	0.2 (0–0.9)	0.2 (0–1.0)	0.70
Prone ventilation	65 (19.3%)	48 (18.2%)	17 (23.6%)	0.31
ECMO	53 (15.8%)	33 (12.5%)	20 (27.8%)	<0.01
Haemodialysis	36 (10.7%)	19 (7.2%)	17 (23.6%)	<0.01

Continued

Table 1 Continued

	All n=336	Survivor n=264	Non-survivor n=72	P value
Steroid usage	144 (46.3%)	103 (42.4%)	41 (60.3%)	0.01
Vasopressor usage	110 (32.7%)	80 (30.3%)	30 (41.7%)	0.09
Sedation usage	222 (66.1%)	175 (66.3%)	47 (65.3%)	0.89

Data were presented as frequencies (percentages) or median (IQR).

APACHE II, Acute Physiology and Chronic Health Evaluation II; ECMO, extracorporeal membrane oxygenation; ICU, intensive care unit; PSI, pneumonia severity index; RT-PCR, reverse transcriptase-polymerase chain reaction.

variable importance when PSI was higher than approximately 130, and this finding is consistent with previous studies on PSI. In line with the results of our previous study, a positive day-4 cumulative fluid balance appeared to be associated with a higher 30-day mortality, as shown in [figure 3B](#).<sup>16</sup> Taken together, these data showed the feature importance in accordance with clinical domains and illustrated the importance of individual features by using SHAP plots so that physicians might have an intuitive understanding of feature importance.

### Comparisons among XGBoost, RF and LR

We then attempted to compare the performance of the three ML models to predict 30-day mortality using data within the first 7 days after admission to the ICU. Using ROC analysis, we found that the AUC value for predicting 30-day mortality in the XGBoost was 0.842 (95% CI 0.749 to 0.928), which was slightly higher than those in RF (AUC: 0.809, 95% CI 0.629 to 0.891) and much better than those in LR (AUC: 0.701; 95% CI 0.573 to 0.825) ([figure 4](#)). The detailed metrics of the performance of these three models were provided (online supplementary table 2). Furthermore, we used DeLong's test to determine the difference between two AUCs and confirmed that XGBoost outperformed RF and LR (XGBoost against RF,  $p=0.002$ ; XGBoost against LR,  $p=0.003$ ). Additionally, we examined the use of standard severity scores in the ICU, including APACHE II and PSI, to predict 30-day mortality in critically ill influenza patients, and the performance of APACHE II and PSI was 0.720 (95% CI 0.653 to 0.784) and 0.720 (95% CI 0.654 to 0.7897), respectively ([figure 5](#)). Collectively, these data demonstrated the value of XGBoost and SHAP plots for giving physicians an intuitive understanding of key features and for establishing a model that predicts the 30-day mortality of critically ill influenza patients with high accuracy.

### DISCUSSION

Using a multicentre severe influenza data set along with XGBoost and SHAP plots, we demonstrated that the ML approach can illustrate key features and establish a mortality prediction model with high accuracy in critically ill influenza patients. The illustration of cumulative domain-specific feature importance and visualised interpretation of feature importance may give physicians

an intuitive understanding of the key features within XGBoost. Furthermore, the prediction model, using clinical data in the routine practice of ICU care instead of advanced molecular biomarkers, can potentially be integrated into a computerised clinical decision support system in the future.

Consistent with other studies on severe influenza,<sup>8,9</sup> we found that both APACHE II and PSI had a discriminative power of approximately 0.72 in the ROC analysis, which is generally deemed acceptable accuracy for a single parameter ([figure 5](#)). The expected finding that the ML approach outperformed PSI and APACHE II in critically ill influenza patients should reflect the capability of ML to establish an influenza-specific weighting model using all of the obtained data. Unlike other critical diseases, which are mainly characterised by shock-associated manifestations, severe influenza is characterised by severely compromised oxygenation, so-called ARDS. Our previous studies have also found that early lung-protective ventilation and early fluid balance were associated with mortality in patients with severe influenza,<sup>15,16</sup> a result that is consistent with current concepts of the early lung-protective ventilation strategy and dry-lung strategy in the management of patients with ARDS.<sup>21,22</sup> Therefore, it was reasonable to find that the ventilation domain (cumulative feature importance: 0.113) and fluid balance domain (cumulative feature importance: 0.253) were of particular importance in the present study, reflecting the additional weights on these two domains in XGBoost compared with those in conventional severity scores, including APACHE II and PSI. Notably, the parameters in the ventilation domain and fluid balance domain were all recorded on a routine daily basis in the ICU, and hence, the established prediction model should be generalisable to other settings. Taken together, the results of this study demonstrate that the use of an ML approach, mainly XGBoost, in a real-world data set was capable of establishing a practical DSS in critically ill patients with a specific aetiology, namely, severe influenza.

Critical care medicine requires prompt decision-making based on clinical data that can be interpreted with the assistance of automated DSS. Increasing evidence has shown the potential of ML-based automated DSS in critical care medicine. Taylor *et al*, using electronic health records and ML algorithms, including the RF model,

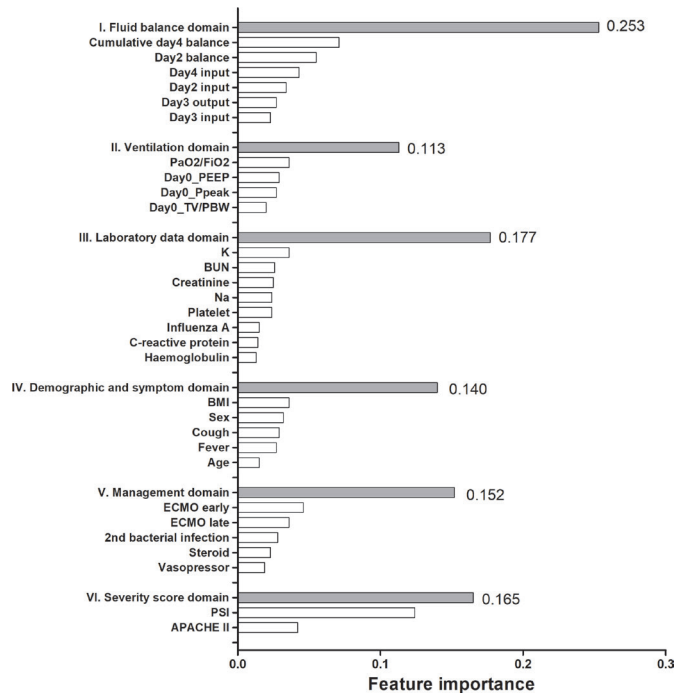
**Table 2** Ventilatory and dynamic fluid parameters of the 336 patients categorised by 30-day mortality

	All n=336	Survivor n=264	Non-survivor n=72	P value
<b>Ventilatory data</b>				
Day-1 FiO <sub>2</sub>	61.6±38.1	59.2±37.5	70.6±39.4	0.02
Day-1 PEEP	8.3±5.6	8.3±5.7	8.2±5.3	0.84
Day-1 V <sub>T</sub> /PBW	6.4±4	6.3±3.9	6.6±4.3	0.54
Day-1 P <sub>peak</sub>	22.1±13.2	22±13	22.5±13.8	0.79
Day-2 FiO <sub>2</sub>	61.5±22.3	57.7±20.9	75.7±21.7	0.00
Day-2 PEEP	11.3±3.7	11.1±3.8	11.9±3.5	0.12
Day-2 V <sub>T</sub> /PBW	8.3±2	8.4±2	8.2±2.1	0.49
Day-2 P <sub>peak</sub>	28.1±4.9	27.8±5.1	29.1±4.3	0.10
Day-3 FiO <sub>2</sub>	53.7±20.1	51.1±18.7	64.5±22.1	<0.01
Day-3 PEEP	11.2±3.7	10.9±3.8	12.5±3.2	<0.01
Day-3 V <sub>T</sub> /PBW	8.1±2	8.2±2	7.7±1.9	0.13
Day-3 P <sub>peak</sub>	27.7±5.3	27.2±5.5	29.7±4.1	0.00
Day-7 FiO <sub>2</sub>	48.9±18.5	45.3±15.2	65±23.2	0.00
Day-7 PEEP	9.8±3.8	9.6±3.7	11.2±4.1	0.02
Day-7 V <sub>T</sub> /PBW	8.3±2.1	8.3±2	7.9±2.5	0.27
Day-7 P <sub>peak</sub>	26.3±6.4	25.7±6.4	29±5.6	0.00
<b>Dynamic fluid status</b>				
Day-1 input	2135.9±1897.2	2036.5±1651.6	2481.5±2561.9	0.22
Day-1 output	1403.9±1052.2	1463.9±1021.4	1185.5±1139.7	0.08
Day-2 input	2626.3±1148.4	2576±1104	2809.2±1290	0.22
Day-2 output	1891.2±1195.5	2003.4±1193.1	1478.6±1119.4	<0.01
Day-3 input	2558.6±1170.7	2481.8±990.3	2875±1701.2	0.13
Day-3 output	2211.6±1438.3	2213.6±1215.3	2203.2±2146.2	0.34
Day-4 input	2424.5±944.1	2392.6±922.3	2563.6±1031.2	0.39
Day-4 output	2338.1±1294.7	2435.8±1259.5	1906.7±1370.4	0.01
Day-1–4 fluid balance	1864.7±3509.7	1347.3±3137.4	3801.2±4128.8	<0.01
Day-5 input	2395.1±935	2365.2±896.9	2534.7±1094.3	0.32
Day-5 output	2552±1362.2	2582.3±1322.5	2411.8±1539.1	0.41
Day-6 input	2345.2±855	2294.2±762.3	2604.4±1199.7	0.09
Day-6 output	2626.1±1882.2	2723.5±1956.9	2130.8±1354.6	0.05
Day-7 input	2395.6±898.8	2322.6±826.3	2785.5±1150.4	0.01
Day-7 output	2510.1±1265	2549.7±1213.2	2299.8±1509.7	0.23
Day-1–7 fluid balance	1351.6±4799.7	506.8±4385	4500.7±4997.8	<0.01

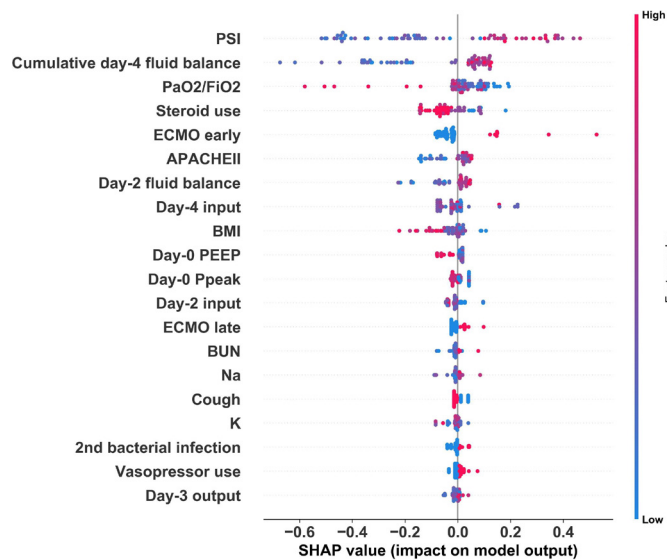
FiO<sub>2</sub>, fraction of inspired O<sub>2</sub>; PBW, predicted body weight; PEEP, positive end-expiratory pressure; P<sub>peak</sub>, peak pressure; V<sub>T</sub>, tidal volume.

recently reported that the ML-based approach was superior to traditional logistic regression for predicting in-hospital mortality in patients admitted to the emergency department for sepsis.<sup>30</sup> In line with our ML approach, Allyn *et al* compared an ML-based approach to traditional scoring systems (EuroSCORE I and EuroSCORE II) for predicting mortality in patients receiving elective cardiac surgery and found that the ML-based approach (AUC: 0.795, 95% CI 0.755 to 0.834) outperformed EuroSCORE I (AUC: 0.737, 95% CI 0.691 to 0.783) and EuroSCORE II (AUC: 0.742, 95% CI 0.698 to 0.785).<sup>31</sup> In addition to

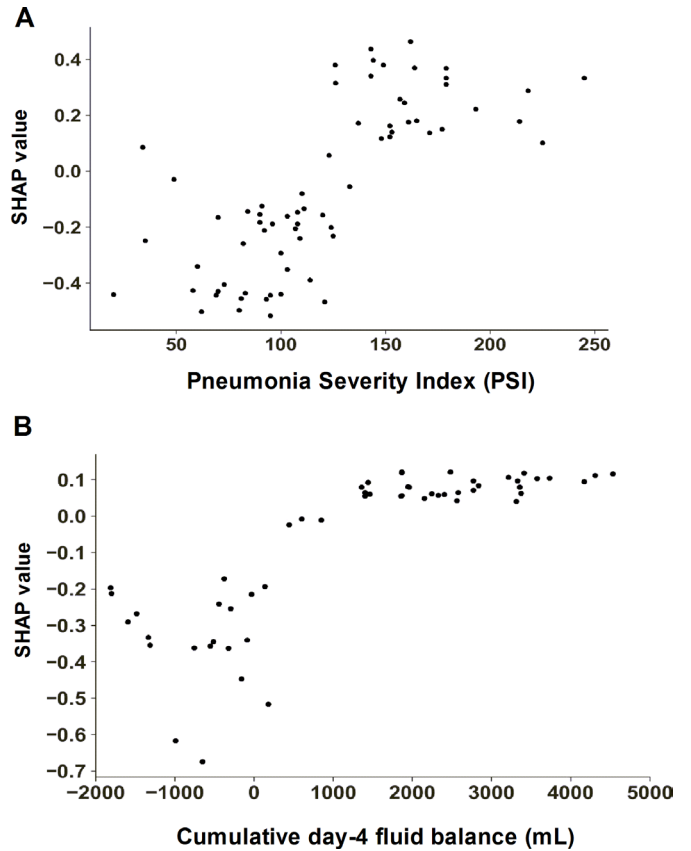
predicting mortality, Horng *et al* recently reported an ML-based model that included vital signs, demographic data and free-text data to identify patients with infection in the emergency department (AUC: 0.85, 95% CI 0.84 to 0.86) and proposed a potential automated trigger for sepsis clinical decision support in the emergency department.<sup>32</sup> Therefore, the accuracy of the prediction model in the field of critical care medicine appears to increase from approximately 0.7 in the conventional scoring system to nearly 0.8 in ML models.



**Figure 1** Relative feature importance categorised by the six main clinical domains. APACHE, Acute Physiology and Chronic HealthEvaluation; BMI,body mass index; BUN, blood urea nitrogen; ECMO, extracorporeal membraneoxygenation; FIO2, fraction of inspired O2; K,potassium; Na, sodium, PBW, predicted body weight; PEEP,positive end-expiratory pressure; PSI, pneumoniaseverity index; P<sub>peak</sub>, peak pressure; V<sub>T</sub>, tidal volume.

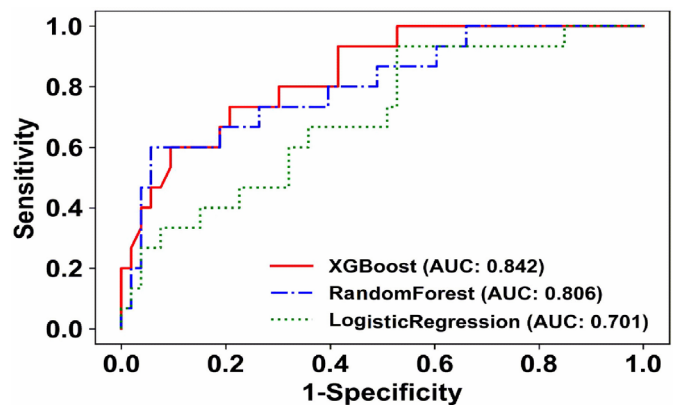


**Figure 2** SHAP summary plots for 30-day mortality predictors in critically ill influenza patients. APACHE, Acute Physiology and Chronic HealthEvaluation; BMI,body mass index; BUN, blood urea nitrogen; ECMO, extracorporeal membraneoxygenation; FIO2, fraction of inspired O2; K,potassium; Na, sodium, PBW, predicted body weight; PEEP,positive end-expiratory pressure; PSI, pneumoniaseverity index; P<sub>peak</sub>, peak pressure; SHAP, SHapley Additive exPlanations; V<sub>T</sub>, tidal volume.

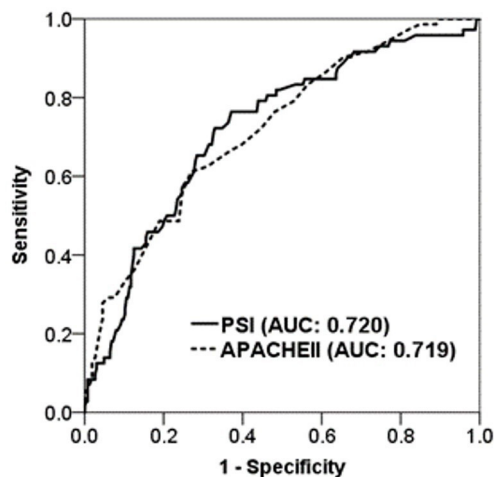


**Figure 3** Partial SHAP dependence plot of two representative features. (A) PSI score. (B) Cumulative day-4 fluid balance. PSI,pneumonia severity index; SHAP, SHapley Additive exPlanations.

In the present study, we found a similar test accuracy between XGBoost (AUC 0.842) and RF (AUC 0.809), whereas the accuracy of LR was poor (AUC 0.701). Indeed, LR is a widely used interpretable algorithm and works well if a single decision boundary exists. However, LR is based on a number of assumptions, including the



**Figure 4** Receiver operating characteristic curves showing the performance of the XGBoost model (AUC 0.842, 95% CI 0.749 to 0.928), RF (AUC 0.809, 95% CI 0.629 to 0.891) and LR (AUC 0.701, 95% CI 0.573 to 0.825) for predicting 30-day mortality in critically ill influenza patients. AUC, area under the curve; LR,logistic regression; RF, random forest; XGBoost,extreme gradient boosting.



**Figure 5** Receiver operating characteristic curves showing the performance of PSI (AUC 0.720, 95% CI 0.654 to 0.7897) and APACHE II (AUC 0.720, 95% CI 0.653 to 0.784) for predicting 30-day mortality in critically ill influenza patients supporting information files. APACHE, Acute Physiology and Chronic HealthEvaluation; AUC, area under the curve; PSI, pneumonia severity index.

independence between input variables and a linear correlation between input and output variables, whereas the real-life data set may not meet all the assumptions of LR. We postulate that the assumption of a linear relationship between the input variables and the outcomes might at least partly account for the relatively low accuracy of LR in this study, as the complex biological events in severe influenza may be correlated with each other in a non-linear model. Instead, tree-based classifiers, including RF and XGBoost, based on homogeneity appeared to fit the characteristics of the data set in the present study. We thought that the application of regularisation, using Taylor expansion to approximate the loss function, and high flexibility to allow for fine-tuning might enable XGBoost to perform slightly better than RF. Taken together, our findings suggest that the XGBoost approach can illustrate the feature importance and establish a mortality prediction model with high accuracy, and this approach has a high potential for practical implementation because it can be integrated with existing computerised healthcare information systems.

Although ML techniques have made substantial advances in many domains, the clinical application of ML-based algorithms in healthcare has not always been straightforward. One major issue that needs to be tackled is the clarification of the black-box issue, rather than higher accuracy, to reassure clinicians in the application of ML-based algorithms in clinical practice.<sup>20 33 34</sup> Given that the nature of ML is based on accuracy-driven performance metrics, it is likely that the model of ML will continue to become even more opaque. Therefore, the black-box problem will remain an issue in the application of ML-based algorithms for patient care. Tree-based ML algorithms, including RF and XGBoost, are characterised

by the potential for interpretation; however, ensembles of hundreds of trees, which are essential to improve the predictive capability, are essentially not interpretable.<sup>35</sup> In the field of medicine, given that the goal of interpretability is to help the physician make a decision based on numerous clinical variables, the interpretability should mimic the behaviour of physicians in real-world practice, rather than merely providing explanations of the logical concepts behind the black box.<sup>36</sup> In the present study, we not only demonstrated the feature importance, a quantitative score that considered the feature's use in the construction of the tree models, categorised by clinical domain of real-world practice in critical care medicine but also provided a visualised interpretation using SHAP plots. SHAP, developed by Lundberg and Lee, is an additive feature attribution method that provides an explanation of the tree ensemble's overall impact in the form of particular feature contributions and is relatively consistent with human intuition.<sup>26</sup> Additionally, we used local interpretable model-agnostic explanations (LIME) to illustrate the impact of key features at the individual level, and the results of LIME were consistent with the findings from SHAP and will enable physicians to apply the ML model to individual patients (online supplementary figure 2).<sup>34 37</sup> However, it is noteworthy that LIME mainly illustrated key features by applying a local linear model. We believe that the concern of the black-box issue should be mitigated by applying these measures, which are designed to interpret the model.

There are limitations to this study that merit discussion. First, the number of subjects was relatively small. Given that only one per cent of patients with influenza-like illness develop severe influenza, the sample size in studies on severe influenza is generally small.<sup>38</sup> As the main focus of the study was to determine predictors for mortality in ICU patients with proven influenza, the enrollees in this study actually accounted for 44.2% (72/163) of deaths among patients with severe influenza during the reported influenza epidemic in Taiwan.<sup>17</sup> To mitigate the issue of small sample size, we have performed cross-validation ( $k=5$ ) of XGBoost, RF and LR. The evaluation metric for cross-validation was the error rate of classification (ie, the number of wrong predictions divided by the total number of predictions). The accuracy for XGBoost, RF and LR were  $0.792\pm 0.022$ ,  $0.786\pm 0.010$  and  $0.732\pm 0.059$ , respectively. These findings are consistent with the data in the manuscript. Second, the study employed a retrospective design, though most research on influenza is retrospective, and evidence obtained from retrospective studies, including the present investigation, might be valuable for future epidemic or pandemic influenza preparations.<sup>39 40</sup> Third, there was a small improvement in the accuracy of mortality prediction. In the fields of critical care medicine and influenza epidemics, small improvements in the accuracy of predicting mortality may have a major impact on various aspects of healthcare, including patient care and resource allocation. Fourth, we used variables within 7 days as independent variables instead of time-varying



variables to predict 30-day mortality. Finally, we did not include hospital-factor in the present study despite a slight increase of accuracy of XGBoost given that we aimed to establish a mortality prediction model with high generalisability in critically ill influenza patients (online supplementary figure 3).

## CONCLUSIONS

In conclusion, using a multicentre severe influenza data set, we found that the ML approach, particularly XGBoost, outperformed traditional severity scoring systems, including APACHE II and PSI, for predicting mortality among critically ill influenza patients. We used domain-based feature importance and SHAP plots for visualised realisation and these approaches should at least partly mitigate the concern of black-box issue. Future prospective research is warranted to validate the proposed model and to translate the advantages of ML models into improved patient outcomes through automated and real-time DSS.

### Author affiliations

<sup>1</sup>Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan

<sup>2</sup>Department of Applied Mathematics, National Chung Hsing University, Taichung, Taiwan

<sup>3</sup>Department of Management Information Systems, National Chung Hsing University, Taichung, Taiwan

<sup>4</sup>Division of Pulmonary and Critical Care, Department of Internal Medicine, China Medical University Hospital, Taichung, Taiwan

<sup>5</sup>Division of Chest Medicine, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan

<sup>6</sup>Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan

<sup>7</sup>Department of Respiratory Care, Chang Gung University of Science and Technology, Chiayi, Taiwan

<sup>8</sup>Division of Pulmonary and Critical Care Medicine, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan

<sup>9</sup>School of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

<sup>10</sup>Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

<sup>11</sup>Department of Chest Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

<sup>12</sup>Institute of Emergency and Critical Care Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan

<sup>13</sup>Department of Thoracic Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan

<sup>14</sup>Department of Respiratory Therapy, Chang Gung University College of Medicine, Taoyuan, Taiwan

<sup>15</sup>Center for Quality Management, Taichung Veterans General Hospital, Taichung, Taiwan

<sup>16</sup>Division of Chest, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan

<sup>17</sup>Department of Critical Care Medicine, Taichung Veterans General Hospital, Taichung, Taiwan

**Acknowledgements** The Taiwan Severe Influenza Research Consortium (TSIRC) study included the following investigators (all in Taiwan). Taipei: National Taiwan University Hospital–HCW, SCK, JYC and YCC. Taipei Veterans General Hospital–K-YY, W-CC and JYF. Tri-Service General Hospital–WCP and CKP. Taoyuan: Chang Gung Memorial Hospital–KCK, LCC, HCH and KWC. Taichung: Taichung Veterans General Hospital–CLW, MCC, WCC, CHT, YHH and ZRZ. China Medical University Hospital–SJJ and WCC. Kaohsiung: Kaohsiung Medical University Hospital–CCS, JRT, MJT

and WAC. Kaohsiung Chang Gung Memorial Hospital–WFF, YMC, CYL and HCK. Hualien: Hualien Tzu Chi Hospital–YTC.

**Contributors** Study concept and design: CA-H, CM-C, Y-CF, S-JL, H-CW, W-FF, C-CS, W-CP, K-YY, K-CK, C-LW, C-ST, M-YL and W-CC. Acquisition of data: Y-CF, CM-C and W-CC. Analysis and interpretation of data: W-CC, Y-CF, CM-C, CA-H and M-YL. Drafting the manuscript: W-CC.

**Funding** This study was supported in part by grants from Veterans General Hospitals and the University System of Taiwan Joint Research Program (VGHUST108-G2-4-2). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

**Competing interests** None declared.

**Patient consent for publication** Not required.

**Ethics approval** Taichung Veterans General Hospital CE16093A, National Taiwan University Hospital 201605036RIND, Taipei Veterans General Hospital 2016-05-020CC, Tri-Service General Hospital 1-105-05-086, Chang Gung Memorial Hospital 201600988B0, China Medical University Hospital 105-REC2-053(FR), Kaohsiung Medical University Hospital KUMHIRB-E(I)-20170097, Kaohsiung Chang Gung Memorial Hospital 201600988B0.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** All data relevant to the study are included in the article or uploaded as supplementary information. All of the data and materials are provided in the manuscript and the supplemental data. The data set has been put in public Github, and is available via <https://github.com/GitEricLin/BMJOpen>.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

### ORCID iD

Wen-Cheng Chao <http://orcid.org/0000-0001-9631-8934>

## REFERENCES

- 1 Fleischmann C, Scherag A, Adhikari NKJ, *et al*. Assessment of global incidence and mortality of Hospital-treated sepsis. current estimates and limitations. *Am J Respir Crit Care Med* 2016;193:259–72.
- 2 Liu V, Escobar GJ, Greene JD, *et al*. Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA* 2014;312:90–2.
- 3 Fleig V, Brenck F, Wolff M, *et al*. [Scoring systems in intensive care medicine : principles, models, application and limits]. *Anaesthesiologie* 2011;60:963–74.
- 4 Vincent J-L, Moreno R. Clinical review: scoring systems in the critically ill. *Crit Care* 2010;14:207.
- 5 Kuzniewicz MW, Vasilevskis EE, Lane R, *et al*. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. *Chest* 2008;133:1319–27.
- 6 Del Bufalo C, Morelli A, Bassein L, *et al*. Severity scores in respiratory intensive care: APACHE II predicted mortality better than SAPS II. *Respir Care* 1995;40:1042–7.
- 7 Knaus WA, Draper EA, Wagner DP, *et al*. Apache II: a severity of disease classification system. *Crit Care Med* 1985;13:818–29.
- 8 Pereira JM, Moreno RP, Matos R, *et al*. Severity assessment tools in ICU patients with 2009 influenza A (H1N1) pneumonia. *Clin Microbiol Infect* 2012;18:1040–8.
- 9 Bjarnason A, Thorleifsdottir G, Löve A, *et al*. Severity of influenza A 2009 (H1N1) pneumonia is underestimated by routine prediction rules. results from a prospective, population-based study. *PLoS One* 2012;7:e46816.
- 10 Tamayo E, Fierro I, Bustamante-Munguira J, *et al*. Development of the post cardiac surgery (POCAS) prognostic score. *Crit Care* 2013;17:R209.
- 11 Salehi SH, As'adi K, Abbaszadeh-Kasbi A, *et al*. Comparison of six outcome prediction models in an adult burn population in a developing country. *Ann Burns Fire Disasters* 2017;30:13–17.
- 12 Banks PA, Freeman ML. Practice parameters Committee of the American College of G. practice guidelines in acute pancreatitis. *Am J Gastroenterol* 2006;101:2379–400.
- 13 Fine MJ, Auble TE, Yealy DM, *et al*. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med* 1997;336:243–50.

- 14 FluNet. Global influenza surveillance and response system, world Health organization, 2016. Available: [http://www.who.int/influenza/gisrs\\_laboratory/flunet/en/](http://www.who.int/influenza/gisrs_laboratory/flunet/en/)
- 15 Chan M-C, Chao W-C, Liang S-J, *et al.* First tidal volume greater than 8 mL/kg is associated with increased mortality in complicated influenza infection with acute respiratory distress syndrome. *J Formos Med Assoc* 2019;118:378–85.
- 16 Chao W-C, Tseng C-H, Chien Y-C, *et al.* Association of day 4 cumulative fluid balance with mortality in critically ill patients with influenza: a multicenter retrospective cohort study in Taiwan. *PLoS One* 2018;13:e0190952.
- 17 Taiwan national infectious disease statistics system: Taiwan centers for disease control., 2016. Available: <http://nidss.cdc.gov.tw/en/>
- 18 Liu V, Turk BJ, Ragins AI, *et al.* An electronic simplified acute physiology score-based risk adjustment score for critical illness in an integrated healthcare system. *Crit Care Med* 2013;41:41–8.
- 19 Slonim N, Carmeli B, Goldstein A, *et al.* Knowledge-analytics synergy in clinical decision support. *Stud Health Technol Inform* 2012;180:703–7.
- 20 Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017;318:517–8.
- 21 Needham DM, Yang T, Dinglas VD, *et al.* Timing of low tidal volume ventilation and intensive care unit mortality in acute respiratory distress syndrome. A prospective cohort study. *Am J Respir Crit Care Med* 2015;191:177–85.
- 22 Rosenberg AL, Dechert RE, Park PK, *et al.* Review of a large clinical series: association of cumulative fluid balance on outcome in acute lung injury: a retrospective review of the ARDSnet tidal volume study cohort. *J Intensive Care Med* 2009;24:35–46.
- 23 Loke YK, Kwok CS, Niruban A, *et al.* Value of severity scales in predicting mortality from community-acquired pneumonia: systematic review and meta-analysis. *Thorax* 2010;65:884–90.
- 24 Friedman JH, Popescu BE. Importance sampled learning ensembles. *J Mach Learn Res* 2003;4:94305.
- 25 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining, ACM, New York 2016:785–94.
- 26 Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *arXiv:170507874v2* 2018.
- 27 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 2002;35:352–9.
- 28 Bryll R, Gutierrez-Osuna R, Quek F. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognit* 2003;36:1291–302.
- 29 Breiman L, Forests R. *Mach Learn* 2001;45:5–32.
- 30 Taylor RA, Pare JR, Venkatesh AK, *et al.* Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016;23:269–78.
- 31 Allyn J, Allou N, Augustin P, *et al.* A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis. *PLoS One* 2017;12:e0169772.
- 32 Horng S, Sontag DA, Halpern Y, *et al.* Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017;12:e0174708.
- 33 Petkovic D, Kobzik L, Re C. Machine learning and deep analytics for biocomputing: call for better explainability. *Pac Symp Biocomput* 2018;23:623–7.
- 34 Zhang Z, Beck MW, Winkler DA, *et al.* Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann Transl Med* 2018;6:216.
- 35 Tan HF, Hooker G, Wells MT. Tree space prototypes: another look at making tree ensembles interpretable. *arXiv:161107115v* 2016.
- 36 Guidotti R, Monreale A, Turini F, *et al.* A survey of methods for explaining black box models. *arXiv:180201933v3* 2018.
- 37 Pedersen TL MB. Lime: local interpretable Model-Agnostic explanations 2018.
- 38 Oliva J, Delgado-Sanz C, Larrauri A, *et al.* Estimating the burden of seasonal influenza in Spain from surveillance of mild and severe influenza disease, 2010-2016. *Influenza Other Respir Viruses* 2018;12:161–70.
- 39 Meltzer MI, Patel A, Ajao A, *et al.* Estimates of the demand for mechanical ventilation in the United States during an influenza pandemic. *Clin Infect Dis* 2015;60 Suppl 1:S52–7.
- 40 Rasmussen SA, Redd SC. Using results from infectious disease modeling to improve the response to a potential H7N9 influenza pandemic. *Clin Infect Dis* 2015;60 Suppl 1:S9–10.