

# Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/159361/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Zeng, Zhixuan, Liu, Yang, Yao, Shuo, Liu, Jiqiang, Xiao, Bing, Liu, Chenxue and Gong, Xun 2023. Neural networks based on attention architecture are robust to data missingness for early predicting hospital mortality in intensive care unit patients. Digital Health 10.1177/20552076231171482 file

Publishers page: <https://doi.org/10.1177/20552076231171482>  
<<https://doi.org/10.1177/20552076231171482>>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies.

See


<http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



# Neural networks based on attention architecture are robust to data missingness for early predicting hospital mortality in intensive care unit patients

DIGITAL HEALTH  
Volume 9: 1–15  
© The Author(s) 2023  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20552076231171482  
journals.sagepub.com/home/dhj



Zhixuan Zeng<sup>1</sup>, Yang Liu<sup>2</sup>, Shuo Yao<sup>1</sup>, Jiqiang Liu<sup>1</sup>, Bing Xiao<sup>1</sup>, Chenxue Liu<sup>3</sup>  
and Xun Gong<sup>1</sup> 

## Abstract

**Background:** Although the machine learning model developed on electronic health records has become a promising method for early predicting hospital mortality, few studies focus on the approaches for handling missing data in electronic health records and evaluate model robustness to data missingness. This study proposes an attention architecture that shows excellent predictive performance and is robust to data missingness.

**Methods:** Two public intensive care unit databases were used for model training and external validation, respectively. Three neural networks (masked attention model, attention model with imputation, attention model with missing indicator) based on the attention architecture were developed, using masked attention mechanism, multiple imputation, and missing indicator to handle missing data, respectively. Model interpretability was analyzed by attention allocations. Extreme gradient boosting, logistic regression with multiple imputation and missing indicator (logistic regression with imputation, logistic regression with missing indicator) were used as baseline models. Model discrimination and calibration were evaluated by area under the receiver operating characteristic curve, area under precision-recall curve, and calibration curve. In addition, model robustness to data missingness in both model training and validation was evaluated by three analyses.

**Results:** In total, 65,623 and 150,753 intensive care unit stays were respectively included in the training set and the test set, with mortality of 10.1% and 8.5%, and overall missing rate of 10.3% and 19.7%. Attention model with missing indicator had the highest area under the receiver operating characteristic curve (0.869; 95% CI: 0.865 to 0.873) in external validation; attention model with imputation had the highest area under precision-recall curve (0.497; 95% CI: 0.480–0.513). Masked attention model and attention model with imputation showed better calibration than other models. The three neural networks showed different patterns of attention allocation. In terms of robustness to data missingness, masked attention model and attention model with missing indicator are more robust to missing data in model training; while attention model with imputation is more robust to missing data in model validation.

**Conclusions:** The attention architecture has the potential to become an excellent model architecture for clinical prediction task with data missingness.

## Keywords

Attention architecture, machine learning, mortality prediction, data missingness, intensive care unit

Submission date: 4 January 2023; Acceptance date: 6 April 2023

<sup>1</sup>Department of Emergency Medicine, The Second Xiangya Hospital of Central South University, Changsha, China

<sup>2</sup>Department of Rehabilitation, The Second Xiangya Hospital of Central South University, Changsha, China

<sup>3</sup>Wales Optometry Postgraduate Education Centre, School of Optometry and Vision Sciences Cardiff University, Cardiff, UK

### Corresponding author:

Xun Gong, Department of Emergency Medicine, The Second Xiangya Hospital of Central South University, Changsha, China.

Email: gongxun1246@csu.edu.cn



## Introduction

Accurate and early prediction of hospital mortality for patients in intensive care unit (ICU) is essential for clinicians to recognize high-risk patients and take timely interventions. The Severity-of-Illness score is the most commonly used tool, such as the Acute Physiology and Chronic Health Evaluation, the Simplified Acute Physiology Score and the Mortality Probability Models.<sup>1</sup> These severity scores generally use clinical variables measured within the first 24 h of the ICU stay to predict hospital mortality based on multivariate logistic regression (LR) algorithm.<sup>2-4</sup> Over the past several years, the fast-emerging machine learning (ML) technology and the popularization of electronic health records (EHRs) promote researches that use EHRs to develop ML models for clinical prediction tasks. The most frequently applied ML algorithms for early prediction of hospital mortality include classification and regression tree (CART).<sup>5-7</sup> Naive Bayes model,<sup>5,8,9</sup> support vector machine,<sup>5,8</sup> random forest,<sup>5-11</sup> extreme gradient boosting (XGB)<sup>5-8,10-13</sup> and artificial neural network.<sup>7,11,14</sup> Compared to conventional severity scores, ML models have more sophisticated algorithm for mining data pattern and show improved predictive performance. However, data missingness in EHRs is poorly handled for model development, validation, and implementation in most of the previous related researches,<sup>15</sup> and this is a crucial issue that undermines the credibility of these ML models for clinical application.

Missing data is unavoidable in all types of clinical researches,<sup>16</sup> especially in retrospective research on EHRs, since EHRs are originally designed to monitor patients and improve clinical efficiency rather than to collect complete data for specific research objectives. When missing data is encountered, most ML models are not adaptive and need for preprocessing approaches which delete, impute or indicate missing data. However, these preprocessing approaches which modify missing data may lead to biased estimation of the real association between variables and outcome.<sup>17-20</sup> Another sort of approach is the built-in algorithm mechanism which makes model capable of handling missing data by itself. Tree-based models are representative examples, such as CART and XGB. Specifically, when a missing variable is encountered, CART employs so-called surrogate splits where a surrogate variable similar to the missing variable is used to decide the split direction,<sup>21</sup> while XGB employs sparsity aware splitting where a unified default split direction is used.<sup>22</sup> Nevertheless, such built-in algorithms also involve missing data in their computing processes.

Besides the above approaches, we can also design a model which neglects missing data and makes predictions only based on non-missing data, so as to avoid possible adverse effect caused by involving missing data into the model computation. Unfortunately, most ML algorithms lack flexible

algorithm mechanisms to realize this design. In recent years, neural networks based on attention architecture have become popular in natural language processing<sup>23,24</sup> and computer vision.<sup>25</sup> The core mechanism of attention architecture can be briefly described as: Given a set of inputs, the model lets one input to pay “attention” to the other inputs and to achieve an integrated analysis of these inputs, where the “attention” is obtained by mathematical operations. This architecture is characterized by the capability of capturing the association between any two inputs without regard to their spatial or temporal order and distance, and the flexibility of allocating “attention” to concerned inputs rather than all inputs. These inspire us to design an attention architecture that is competent for mortality prediction and adaptive to missing data in EHRs.

In this study, we propose a simple and effective attention architecture. Based on this architecture, we achieve the design of filtering out missing data from model computation by introducing a mask function into the regular attention mechanism. This masked attention model (MAM) takes a set of clinical variables within the first 24 h during the ICU stay as inputs and outputs the predicted hospital mortality. In addition, we also develop other two neural networks based on this architecture which employ imputation and missing indicator to handle missing data respectively. These attention-based models show a state-of-the-art predictive performance, and furthermore they are robust to data missingness in model training and validation.

## Methods

### Source of data

We implemented a retrospective cohort study on two large public ICU databases: The Medical Information Mart for Intensive Care IV (MIMIC-IV)<sup>26</sup> and the eICU Collaborative Research Database (eICU-CRD).<sup>27</sup> MIMIC-IV database contained clinical records of patients admitted to ICUs of the Beth Israel Deaconess Medical Center between 2008 and 2019, while eICU-CRD contained records of patients admitted to 335 ICUs in 208 hospitals in the US between 2014 and 2015. These two databases were mutually independent, without overlapped data. Local ethical review board (ERB) approvals were achieved for both the two databases and all personal information was deidentified in accordance with the Health Insurance Portability and Accountability Act standards, thus an ERB approval from our institution was exempted.

### Participants and data extraction

In this study, we used clinical data within the first 24 h of an ICU stay to predict hospital mortality. In order to develop a general prediction model, we included all patients from the

two databases rather than restricting our target population in a specific disease group. For patients with multiple hospitalizations, every hospitalization was included; for hospitalizations with multiple ICU stays, only the first ICU stay was considered as it provided the earliest clinical data for mortality prediction. The exclusion criteria were as follows: 1. age not between 18 and 89 years old at ICU admission; 2. not the first ICU stay of a hospitalization. We extracted all available records of demographic characteristics, comorbidities, vital signs, Glasgow Coma Score, laboratory tests, ventilator parameters, vasoactive drugs, etc. Each included ICU stay was treated as a sample in this study. Categorical variables were represented as 0 for absence and 1 for presence. One-hot encoding was employed for gender and admission type. Continuous variables which were probably observed for multiple times during the first 24 h were represented as the maximum, minimum, mean, and standard deviation as appropriate. The finally employed variables and their ID numbers were summarized in Supplemental Table 1. The label of each sample was the survival state of the patient at discharge (0 for survival and 1 for death).

### Study design

We selected the eligible samples in MIMIC-IV as the training set and the eligible samples in eICU-CRD as the test set. Then a 5-fold cross-validation was implemented on the training set, where the training set was randomly and equally split into five mutually exclusive subsets and in each fold four of them were used for model training and the rest one was used for internal validation. Thus, for each type of model, a total of five model instances were developed. Then all instances were evaluated by the external validation on the test set, and the performance of the five instances was aggregated for final evaluation of a model type.

### Neural networks based on attention architecture

In this section, we introduced the three neural networks based on our attention architecture: MAM, attention model with imputation (AM\_imp) and attention model with missing indicator (AM\_ind). The proposed attention architecture contained three major components: embedding layer, multi-head attention layer, and fully connected linear layer. Firstly, the embedding layer was applied to transform clinical variables into numerical vectors, followed by layer normalization.<sup>28</sup> Then layer-normalized vectors were sequentially fed into a multi-head attention layer with the residual connection.<sup>29</sup> Finally, a linear layer followed by Sigmoid function was applied to project the output of the previous layer to predicted mortality. In addition, we also explored the

interpretability of these models by analyzing the allocation of attentions on clinical variables.

**Model architecture of MAM.** MAM was derived from the attention architecture where a mask function was introduced in the multi-head attention layer (Figure 1(a)). We took MAM as an example to provide a detailed explanation of our attention architecture as the following.

**Embedding layer.** The model input was a set of clinical variables, with each variable containing its textual name and numerical value (we used the phrase of “numerical value” here to distinguish it from the conception of “value” used in the attention mechanism). For example, when the age of a patient was 75 years old, the textual name was “age” and the numerical value was “75.” We transformed clinical variables to numerical vectors by the embedding layer. The specific procedures included: (a) erroneous numerical values out of reasonable range were treated as missing values; (b) a word embedding layer<sup>30</sup> was applied to map each textual name to a 2-dimensional numerical vector; (c) numerical values of continuous variables were normalized by subtracting the mean and dividing by the standard deviation, where the mean and the standard deviation were derived from the training set; (d) all missing numerical values were set to zero (although missing variables would be filtered out in the next layer, this step was needed for running python code without null error); (e) each clinical variable was represented as a 3-dimensional vector by concatenating its name-embedding vector and its normalized numerical value (Figure 1(c)).

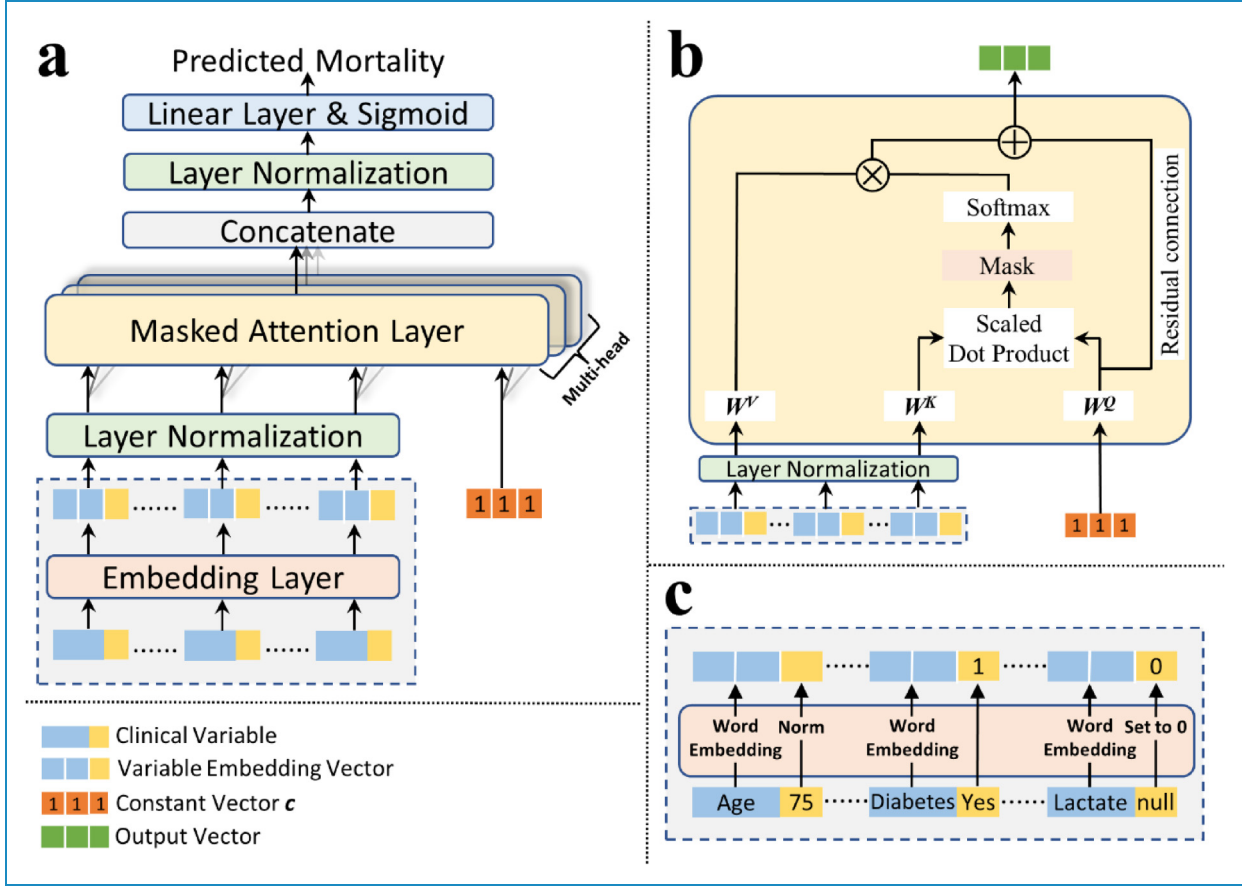
**Masked attention layer.** The attention mechanism could be mathematically described as a function that mapped a query and a set of key-value pairs to an output. Generally, attentions of the query on every key-value pair should be computed. In MAM, we employed a masked attention layer that only allocated attentions to the key-value pairs of non-missing clinical variables. Specifically, we firstly introduced a 3-dimensional constant vector  $\mathbf{c}$  ( $\mathbf{c} = [1, 1, 1]$ ), then the query, keys, and values were computed as:

$$\mathbf{q} = \mathbf{c} \mathbf{W}^Q \quad (1)$$

$$\mathbf{K} = \mathbf{X} \mathbf{W}^K \quad (2)$$

$$\mathbf{V} = \mathbf{X} \mathbf{W}^V \quad (3)$$

where  $\mathbf{W}^Q \in R^{3 \times 3}$ ,  $\mathbf{W}^K \in R^{3 \times 3}$ ,  $\mathbf{W}^V \in R^{3 \times 3}$  were learnable weight matrices for generating query, key, and value, respectively;  $\mathbf{c} \in R^{1 \times 3}$  was the constant vector and  $\mathbf{q} \in R^{1 \times 3}$  was the query vector of  $\mathbf{c}$ ;  $\mathbf{X} \in R^{n \times 3}$  was the matrix containing all clinical variables, where  $n$  was the number of employed variables and each row of  $\mathbf{X}$  was an 3-dimensional vector from the embedding layer;



**Figure 1.** Model architecture of masked attention model (MAM). (a) Overall architecture. (b) Masked attention layer. (c) Embedding layer.

$\mathbf{K} \in R^{n \times 3}$  and  $\mathbf{V} \in R^{n \times 3}$  were matrices for corresponding keys and values of  $\mathbf{X}$ . Then the masked attention was computed as:

$$\mathbf{a}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\text{mask}\left(\frac{\mathbf{q} \mathbf{K}^T}{\sqrt{d_k}}\right)\right) \mathbf{V} \quad (4)$$

where scalar  $d_k$  was the dimension of key ( $d_k = 3$ ) and  $\frac{\mathbf{q} \mathbf{K}^T}{\sqrt{d_k}}$  was a  $n$ -dimensional vector. The  $i$ th element of the vector  $\frac{\mathbf{q} \mathbf{K}^T}{\sqrt{d_k}}$  represented the scaled dot-product attention<sup>23</sup> between the constant vector  $\mathbf{c}$  and the  $i$ th clinical variable. Then a mask function was used to set scaled dot-product attention on missing variable to approximate negative infinity. Given  $s = \frac{\mathbf{q} \mathbf{K}^T}{\sqrt{d_k}}$ , the  $i$ th element of mask function applied to  $s$  was defined as:

$$\text{mask}(s)_i = \begin{cases} s_i, & \text{for non-missing variable} \\ -10^9, & \text{for missing variable} \end{cases} \quad (5)$$

The softmax function in formula (4) ensured that final attentions of  $\mathbf{c}$  on all the clinical variables summed to 1. Given  $\mathbf{m} = \text{mask}(s)$ , the  $i$ th element of softmax

function applied to  $\mathbf{m}$  was defined as:

$$\text{softmax}(\mathbf{m})_i = \frac{e^{m_i}}{\sum_j e^{m_j}} \quad (6)$$

Thus, final attention on missing variables approximately equaled to zero, which meant that missing variables were filtered out from the attention-weighted sum of value vectors and had no impact on the output  $\mathbf{a}(\mathbf{q}, \mathbf{K}, \mathbf{V}) \in R^{1 \times 3}$  in formula (4). At last, we introduced residual connection in the masked attention layer. That was, the final output was computed as:

$$\text{output}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathbf{q} + \mathbf{a}(\mathbf{q}, \mathbf{K}, \mathbf{V}) \quad (7)$$

The above algorithm of masked attention was illustrated in Figure 1(b).

**Masked multi-head attention.** The multi-head attention performed multiple sets of above attention algorithm in parallel, where each set of attention algorithm was referred to as a head. Each head had its own learnable weight matrices  $\mathbf{W}^Q$ ,  $\mathbf{W}^K$ ,  $\mathbf{W}^V$ , thus multiple heads were capable of capturing different data patterns. For masked multi-head attention with  $h$  heads, a total of  $h$  vectors with size of  $1 \times 3$  were

produced, then the  $h$  outputs were concatenated to form an output vector with size of  $1 \times 3h$ . Finally, this output vector was processed by layer normalization and fed into the linear layer (Figure 1(a)). The quantity of heads was an important hyperparameter of the attention architecture. In this study, we respectively used 1, 3, 5, 7, 9 heads in multi-head MAM, and the number of heads with the highest average area under the receiver operating characteristic curve (AUROC) in the 5-fold cross validation was selected for subsequent research. For a fair comparison, AM\_imp and AM\_ind used the same quantity of heads as MAM.

**Attention model with imputation.** AM\_imp had the same architecture as MAM except that it did not employ the mask function. In the embedding layer missing numerical values were not set to zero. We employed multiple imputation (MI) using multivariate imputation by chained equations<sup>31</sup> to preprocess missing data. Specifically, we employed multivariate regression models as imputation models and used the training set to train them. We included all the clinical variables except the outcome variable in the imputation procedure to avoid leaking information of the outcome to prediction model. A total of five imputed datasets were created in MI, then estimated regression coefficients of imputation models in the five imputations were combined using Rubin's rules<sup>32</sup> to form the final imputation model. Notably, as it was irrational to impose imputed ventilator parameters to non-ventilation patients, this part of missing data was to zero as default.

**Attention model with missing indicator.** AM\_ind used a missing indicator instead of the mask function to handle missing data. Its architecture was illustrated in Supplemental Figure 1. In the embedding layer, we set all missing numerical value to zero, added a binary indicator (0 for non-missing variable and 1 for missing variable), so each clinical variable was represented as a 4-dimensional vector. And the mask function was removed from the attention layer.

**Interpretability of the attention-based neural networks.** We explored the interpretability of our attention-based neural networks by analyzing their attention allocations to the employed clinical variables. As mentioned above, the attention architecture integrated multiple clinical variables through the weighted sum of their value vectors, where the weight of each clinical variable was the attention of vector  $c$  allocated to this variable. Thus, a clinical variable acquiring higher attention had greater contribution to model output and was more important for hospital mortality prediction. In order to inspect variable importance captured by our attention-based models, for all heads of all trained instances of MAM, AM\_imp, and AM\_ind, we analyzed the average acquired attention for all employed clinical variables among samples in the external validation. Notably, in MAM, a variable had participated in model

computation only among the samples in which this variable is non-missing (in the other samples the masked mechanism made this variable acquiring zero attention). Thus, the importance of a variable with high missing rate would be underestimated if its average acquired attention was computed among all samples of the test set. For this reason, in each head of MAM, the average acquired attention of the  $i$ th variable was defined as  $\frac{1}{n_i} \sum_{j=1}^{n_i} a_{i,j}$ , where  $n_i$  was the number of samples whose  $i$ th variable was not missing, and  $a_{i,j}$  was the attention value of the  $i$ th variable for the  $j$ th sample in the test set. While for AM\_imp and AM\_ind, missing variables that were imputed or indicated also acquired attention and participated in model computation, so the average acquired attention in these two models was computed over all samples in the test set.

### Baseline models

We employed three baseline models for comparison: XGB, LR with imputation (LR\_imp) and LR with missing indicator (LR\_ind).

XGB was widely applied in previous researches aiming to early predict hospital mortality for ICU patients and showed improved predictive performance over other ML models.<sup>5-8,10-13</sup> As mentioned before, XGB owned a built-in mechanism to handle missing data, which made it competent for our dataset. For optimizing hyperparameters of XGB, we performed a grid search on different combinations of the following hyperparameter settings: n\_estimators (400, 600, 800), learning\_rate (0.01, 0.05, 0.1), colsample\_bytree (0.6, 0.8), subsample (0.4, 0.6, 0.8), max\_depth (4, 6, 8), min\_child\_weight (1.0, 2.0), gamma (0.2, 0.4), and determined the optimal setting to achieve the highest average AUROC in the 5-fold cross-validation on the training set.

LR\_imp was a LR model with L1 weight regularization. And the missing data was preprocessed by the same imputation model used in AM\_imp.

LR\_ind was another LR model which set missing variable to zero and added a binary indicator for each variable (0 for non-missing variable and 1 for missing variable) as model input. Thus, LR\_ind took double-quantity inputs compared to LR\_imp.

### Statistical analysis and evaluation of model performance

For both the training set and the test set, clinical variables were compared between samples in survival group and death group, using either Student  $t$  test, rank-sum test or Chi-square test as appropriate. Continuous variables were described as mean (standard deviation) or median [interquartile range], and categorical features were described as number (percentage). In addition, the number and percentage of missing data for each variable were also counted.

The AUROC and the area under the precision-recall curve (AUPRC) were employed to evaluate the discriminative ability of models. The mean and 95% confidence interval (CI) for each type of model were obtained by aggregating the measurements of five model instances developed in the 5-fold cross-validation. The calibration curve was employed to visualize model calibration.<sup>33</sup> We adopted the average predicted probabilities of five model instances as the final predicted probability for each sample, and plotted means of decile-binned predicted probabilities versus corresponding means of actual probabilities in the samples in each bin. The calibration was assessed by inspecting the proximity between the calibration curve and the identity line of  $y = x$  which represented perfect calibration.

The attention-based models were built using Pytorch version 1.7.1, and the XGB, LR, and imputation model were built using Scikit-learn package version 0.23.1. Statistical analysis was performed using SciPy package version 1.5.2. Two tailed  $P < 0.05$  was considered as statistical significance.

### Model robustness to missing data

We estimated model robustness to data missingness in both model validation and model training, by analyzing the alteration of model performance under increasing missing rate in the test or training set. A total of three analyses were performed. At first, we focused on the impact of the inherent missingness in the test set on model validation. We performed a subgroup analysis in which the

samples in the test set were divided into five subgroups based on their missing rate: 0%–10%, 10%–20%, 20%–30%, 30%–40% and more than 40%. Then, we employed the previously developed prediction models and imputation models without retraining, and evaluated their AUROCs and AUPRCs on the above subgroups respectively. In the second analysis, we focused on the impact of random missingness on model validation. We introduced additional random missingness in the raw test set, by artificially setting every piece of non-missing variable to missing data under a certain probability  $P$ , while the training set, the previously developed prediction models and imputation models were still fixed. Then we validated our models on the modified test sets which were produced under the  $P$  of 0.2, 0.4, 0.6, and 0.8. And for each setting of  $P$ , we repeated this random modification on the test set ten times to obtain the mean and 95% CI of AUROC and AUPRC. In the third analysis, we focused on the impact of random missingness on model training. This time the repeated random modification under different  $P$  values was performed on the raw training set, while the test set was not modified. For each modified training set, we retrained our prediction models and imputation models (for AM\_imp and LR\_imp), where 80% of the modified training set was randomly selected for model training and 20% were for internal validation, and then retrained models were externally validated on the unmodified test set. We did not change any architecture or hyperparameters of our models during model retraining.

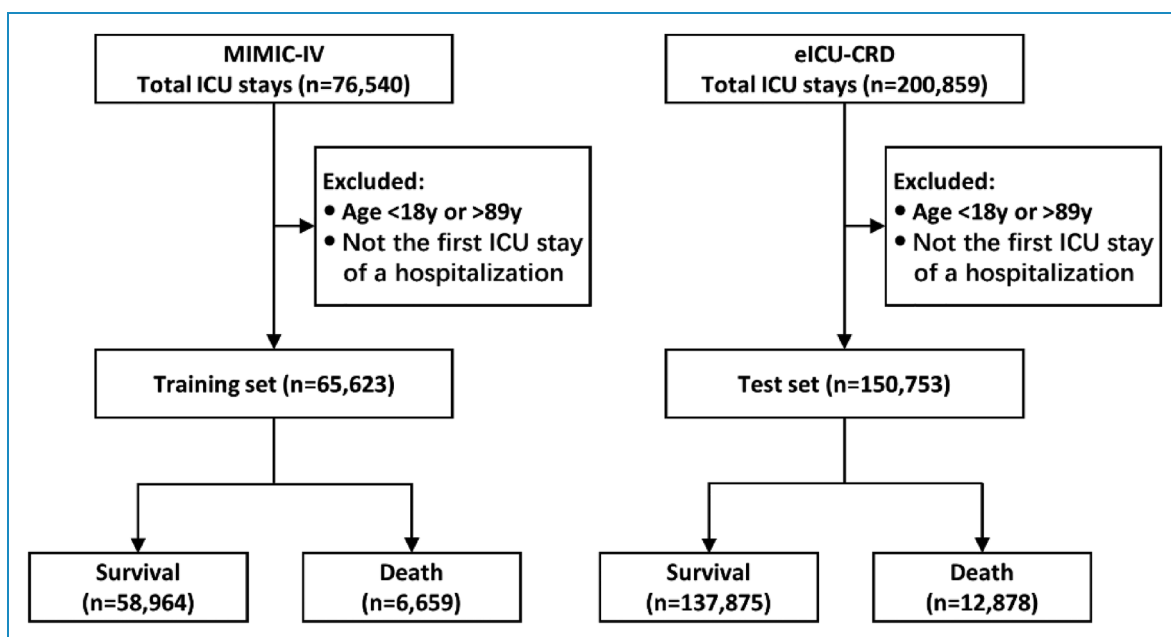


Figure 2. Flow chart of patient selection.

**Table 1.** Comparison of baseline characteristics.

	Training set from MIMIC-IV (n = 65623)			Test set from eICU-CRD (n = 150753)		
	Survival (n = 58964)	Death (n = 6659)	P	Survival (n = 137875)	Death (n = 12878)	P
Gender (male) , n (%)	33430 (56.696)	3716 (55.804)	0.168	75246 (54.6)	6977 (54.2)	0.391
Age (y, mean (SD))	62.53 (16.31)	68.66 (14.35)	<0.001	61.4 (16.7)	68.2 (14.4)	<0.001
Admission type			<0.001			<0.001
Medical, n (%)	41298 (70.0)	5540 (83.2)		109777 (79.6)	11849 (92.0)	
Unscheduled surgical, n (%)	15912 (27.0)	1087 (16.3)		25623 (18.6)	910 (7.1)	
Scheduled surgical, n (%)	1754 (3.0)	32 (0.5)		2475 (1.8)	119 (0.9)	
SOFA (median [IQR])	3.0 [1.0, 5.0]	6.0 [4.0, 9.0]	<0.001	2.0 [1.0, 4.0]	6.0 [3.0, 9.0]	<0.001
SAPS II (median [IQR])	33.0 [25.0, 42.0]	55.0 [43.0, 68.0]	<0.001	29.0 [21.0, 38.0]	50.0 [37.0, 65.0]	<0.001
Length of ICU stay (hours, median [IQR])	44.3 [25.8, 79.9]	67.3 [28.2, 155.1]	<0.001	39.0 [21.0, 70.0]	51.0 [19.0, 119.0]	<0.001

SOFA Sequential Organ Failure Assessment, SAPS Simplified Acute Physiology Score; ICU: intensive care unit; MIMIC-IV: Medical Information Mart for Intensive Care IV; eICU-CRD: eICU Collaborative Research Database.

**Table 2.** AUROCs for MAM with different attention heads in 5-fold cross-validation.

Heads	1	3	5	7	9
AUROC [95%CI]	0.888 [0.880–0.895]	0.892 [0.884–0.900]	0.889 [0.880–0.898]	<b>0.896 [0.885–0.907]</b>	0.894 [0.885–0.903]

AUROC: area under the receiver operating characteristic curve; MAM: masked attention model.

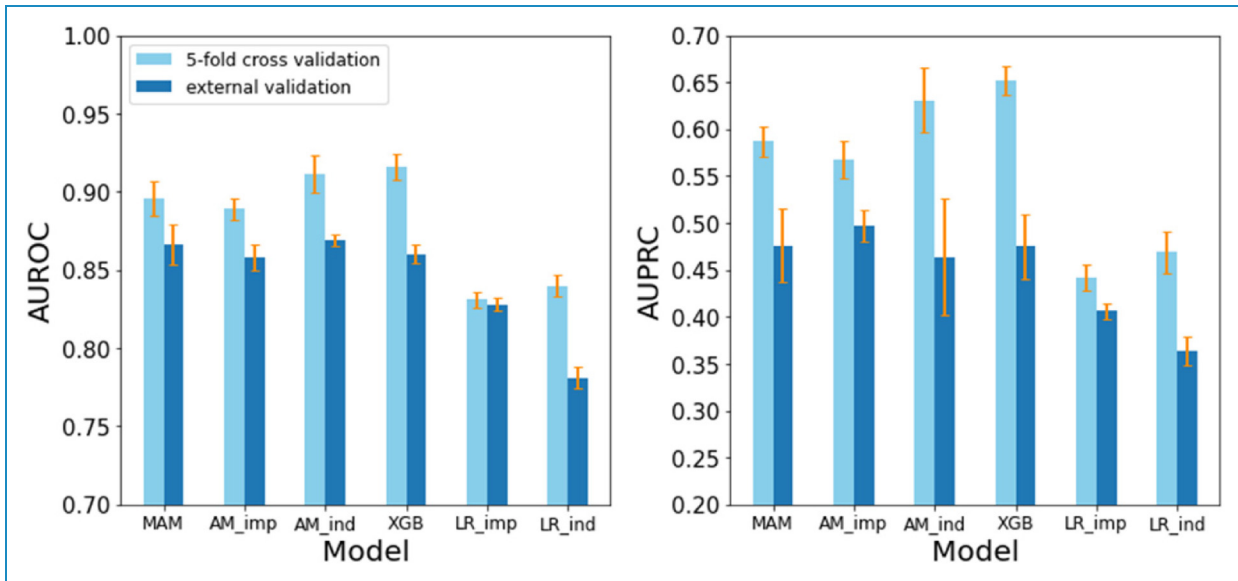
## Results

### Participants and clinical variables

We ultimately included 65,623 ICU stays for 50,354 patients from MIMIC-IV and 150,753 ICU stays for 126,804 patients from eICU-CRD (Figure 2). In-hospital death occurred for 6659 (10.1%) and 12,878 (8.5%) ICU stays in the training set and the test set respectively. Comparison of the baseline characteristics between samples in survival and death group for both the training set and the test set was provided in Table 1. And comparison of the other employed clinical variables and statistics of their missing rate was provided in Supplemental Table 2. Our results demonstrated the statistical difference of variables between survival and death group. Regarding to data missingness, overall missing rate was 10.3% for the training set and 19.7% for the test set. As shown in Supplemental Table 2, the test set had a higher missing

rate for most clinical variables compared to the training set. The four ventilator parameters (Max\_TV\_setting, Max\_Ppeak, Max\_Pplat, Max\_PEEP) showed the highest missing rates in both the training set (>62% in the survival group and >38% in the death group) and the test set (>79% in survival group and >49% in death group). As ventilator parameters for non-ventilation patients were treated as missing variables, this result was related to the corresponding ventilation rate in the training set (37.3% for survival group and 61.3% for death group) and the test set (21.2% for survival group and 53.2% for death group). Other high-missing variables included Mean\_pH, Min\_PaO2, Mean\_PaCO2, Min\_PaO2/FiO2, Max\_Lactate, Max\_TBil, Max\_ALT, Max\_AST, etc. For these high-missing variables, the missing rate was obviously higher in survival group than in the death group, while for the other variables the difference of missing rate between survival and death group was relatively small.





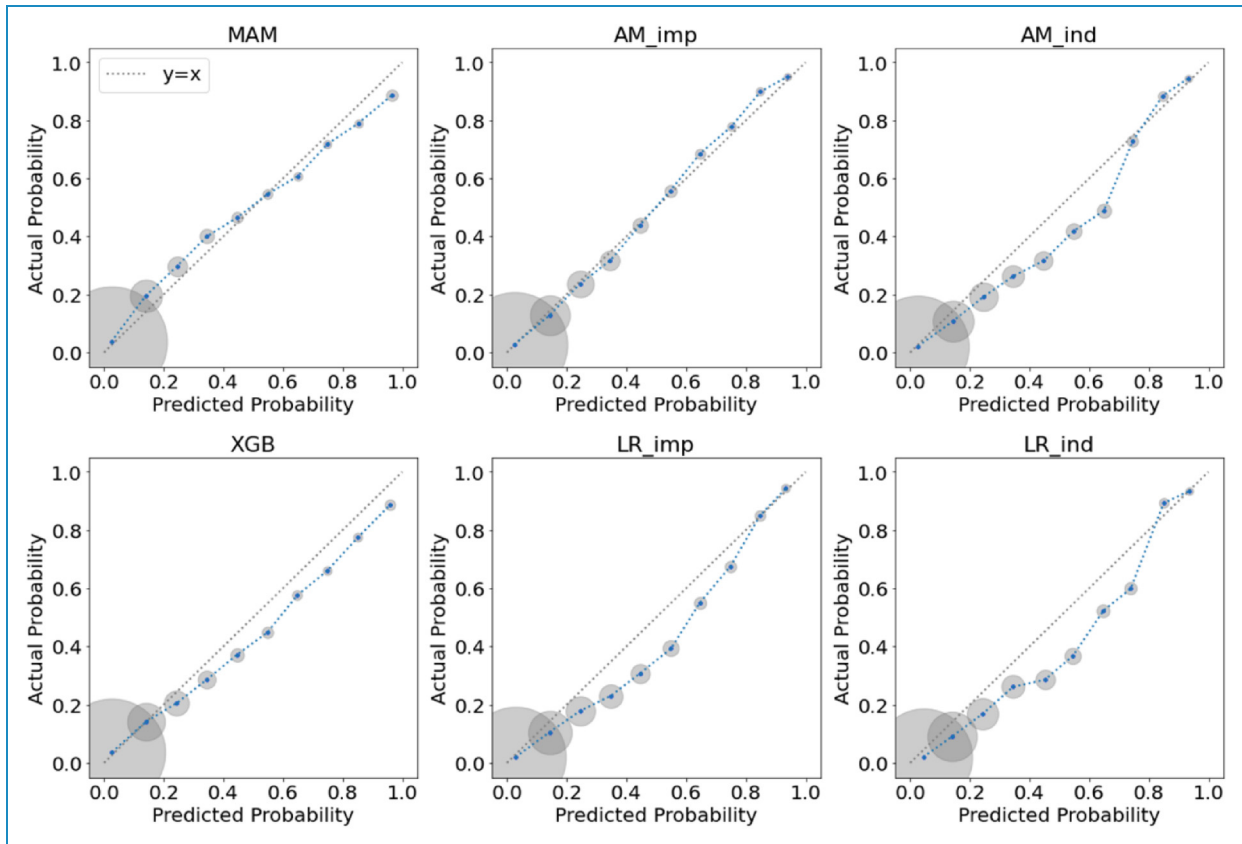
**Figure 3.** AUROCs and AUPRCs for 5-fold cross validation and external validation. MAM: masked attention model, AM\_imp: attention model with imputation, AM\_ind attention model with missing indicator, XGB: extreme gradient boosting, LR\_imp: logistic regression with imputation, LR\_ind: logistic regression with missing indicator; AUROC: area under the receiver operating characteristic curve.

### Model performance

Our result showed that 7-head MAM had the highest average AUROC in the 5-fold cross-validation (Table 2). Thus, we selected 7-head MAM for subsequent research, and the same setting was used in AM\_imp and AM\_ind for comparison. The optimized hyperparameters of XGB were as following: `n_estimators = 600`, `learning_rate = 0.05`, `colsample_bytree = 0.8`, `subsample = 0.6`, `max_depth = 6`, `min_child_weight = 2.0`, `gamma = 0.2`. In terms of model discrimination, we showed the AUROCs and AUPRCs in the 5-fold cross-validation and the external validation for all the models in Figure 3. In the external validation, AM\_ind had the highest AUROC (0.869; 95% CI: 0.865–0.873) and AM\_imp had the highest AUPRC (0.497; 95% CI: 0.480–0.513), while LR\_ind had the lowest AUROC (0.781; 95% CI: 0.774–0.788) and AUPRC (0.364; 95% CI: 0.349–0.378). In terms of model calibration, we provided the calibration curves of the models in Figure 4. MAM and AM\_imp showed a better calibration with their curves closely around the diagonal, while the curves of the other four models deviated from the diagonal more obviously. MAM slightly underestimated the risk in low-risk bins and slightly overestimated the risk in high-risk bins; AM\_ind, LR\_imp, and LR\_ind overestimated the risk in middle-risk bins (from 0.3 to 0.7 bins); XGB overestimated the risk in almost in all risk bins (from 0.3 to 1.0 bins).

### Model interpretation

For all trained instances of MAM, AM\_imp, and AM\_ind, the average acquired attentions of all employed variables in the external validation were shown in Figure 5. We compared the attention allocations among model types, model instances, and attention heads, respectively. Firstly, at the level of model type, the three models showed different patterns of attention allocation. Some variables were treated as important predictors in one model but were neglected in another. For example, variable 10 (Cerebrovascular disease) and 57 (Mean white blood cell) had high average acquired attention in most heads of the five instances of MAM, but they had relatively low attention in AM\_imp and AM\_ind. Such a difference demonstrated the influence of the approach for handling data missingness on attention allocation. Secondly, at the level of model instance, smaller difference of attention allocation was observed among the five instances of a model type. As shown in Figure 5, for most MAM instances, most variables between 40 and 51 and between 57 and 70 acquired high attention, while most variables between 27 and 34 acquired low attention; for most AM\_imp instances, variables between 35 and 45 mostly acquired high attention and variables between 8 and 18 mostly acquired low attention; for most AM\_ind instances, the attention allocation was more focused on several variables, such as variable 3 (Admissiontype\_medical), 5 (Admissiontype\_unscheduled\_surgical), 6 (Age), 51 (Minimum Glasgow Coma Score), 80 (Urine output) and 81 (Invasive mechanical ventilation). Lastly, at the level of



**Figure 4.** Calibration curves for external validation. For each model, the calibration curve plotted means of decile-binned predicted probabilities versus corresponding means of actual probabilities in the patients in each bin. As shown, each blue point of a calibration curve represented a bin and the size of the gray circle around represented the sample size of this bin. The dotted line was the identity line of  $y = x$  representing perfect calibration. MAM: masked attention model, AM\_imp: attention model with imputation, AM\_ind: attention model with missing indicator, XGB: extreme gradient boosting, LR\_imp: logistic regression with imputation, LR\_ind: logistic regression with missing indicator.

attention head, attention allocations of most heads in one instance were relatively consistent except for several heads which showed a different attention allocation, such as the 7th head of the MAM instance 1, the second and the sixth head of the AM\_imp instance 3, and the fifth head of the AM\_ind instance 4. This indicated that the models were capable of capturing different data patterns through multiple heads.

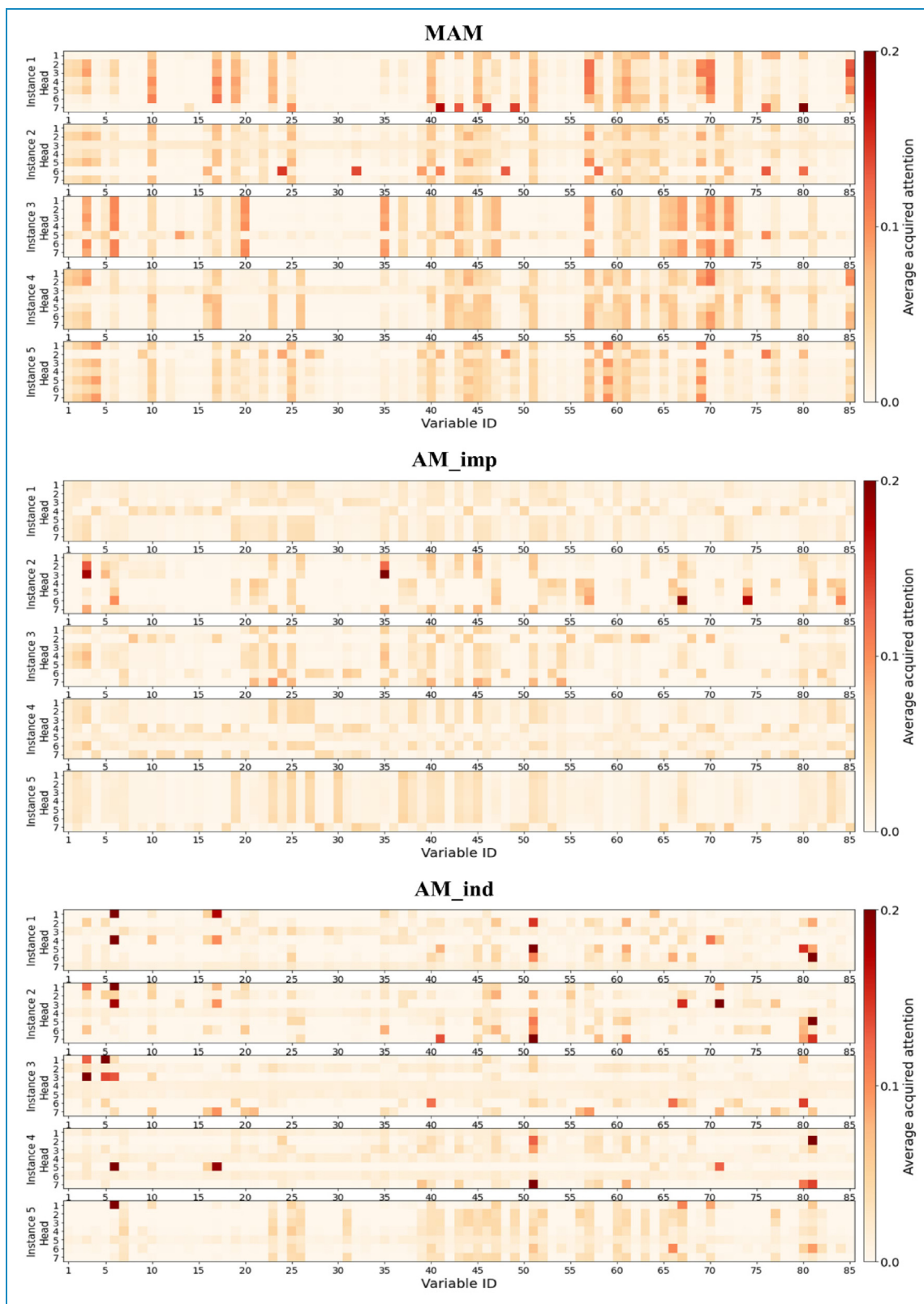
### Model robustness to data missingness

The results of our three analyses about model robustness to data missingness were demonstrated in Figure 6. Each subgraph in Figure 6 showed the means and 95% CIs of AUROC or AUPRC for all types of models in external validations under corresponding settings.

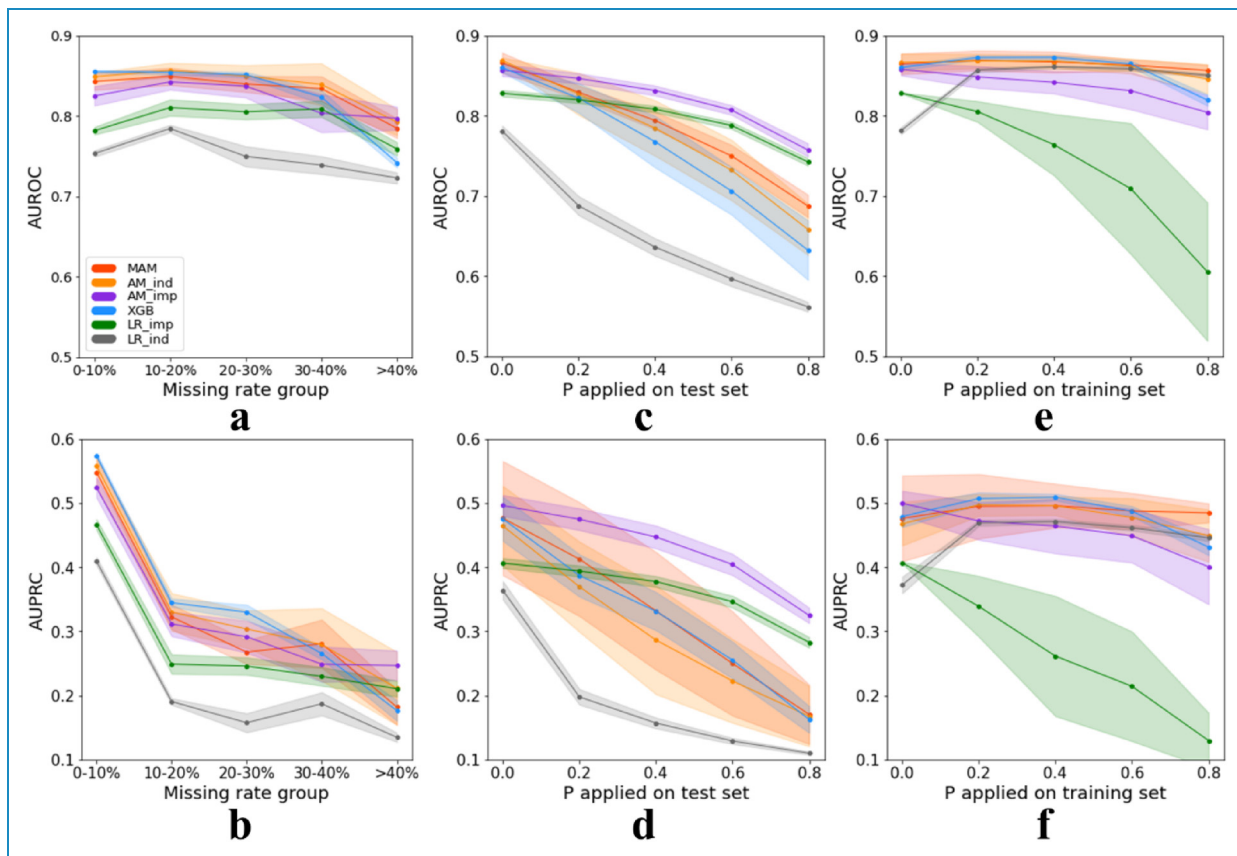
The first was the subgroup analysis and its result was provided in Figure 6(a) and Figure 6(b). The sample size and hospital mortality of the five subgroups with missing rates of 0%–10%, 10%–20%, 20%–30%, 30%–40% and >40% were 27,233 (mortality: 18.8%), 80,900 (5.6%), 15,858 (5.9%), 13,462 (7.9%), and 13,300 (8.9%),

respectively. Overall, most models showed lower AUROCs in subgroups with higher missing rate, especially in 30%–40% and >40% subgroups. The AUROCs of MAM, AM\_ind and LR\_imp kept stable in the first four subgroups and started to decline in the last >40% subgroup. The AUROCs of AM\_imp and XGB started to decline in the 30%–40% subgroup, but the AUROC of AM\_imp kept more stable in the >40% subgroup. The AUROCs of LR\_ind started to decline in the 20%–30% subgroup. In the last >40% subgroup, the three attention-based models showed higher AUROC than the other baseline models. Compared to AUROC, the AUPRCs of all models declined more obviously, especially in 10%–20% subgroup. And AM\_imp showed the most stable and highest AUPRC in the last >40% subgroup.

The second analysis was supplementary to the first analysis for evaluating the impact of random missingness on model validation, and the results were shown in Figure 6(c) and Figure 6(d). Overall, the AUROCs and AUPRCs of all the models declined as the missing probability  $P$  increased, which indicated that increasing random



**Figure 5.** Model interpretation by attention allocations. For the three model types: MAM, AM\_imp and AM\_ind, five heat-map subgraphs were used to show attention allocations for their five trained instances. Each small colored square in a heatmap showed the average acquired attention of a variable in a head of this instance. The color bar on the right indicated the value of the average acquired attention, from low (light red) to high (dark red). MAM: masked attention model, AM\_imp: attention model with imputation, AM\_ind: attention model with missing indicator.



**Figure 6.** Model robustness to data missingness. Each colored point in a subgraph represented the AUROC or AUPRC in external validation under a certain setting, and different colors indicated corresponding model types. Points of a model were connected for reflecting the change tendency and the shadow around indicated the 95% confidence interval. (a, b) AUROCs and AUPRCs for subgroup analysis. (c, d) AUROCs and AUPRCs when random missingness was introduced in the test set under probability of  $P$ . (e, f) AUROCs and AUPRCs when random missingness was introduced in the training set under probability of  $P$ . AUROC: area under the receiver operating characteristic curve; AUPRC: area under precision-recall curve.

missingness in the raw test set impaired the predictive performance of our developed models. The two models using imputation: AM\_imp and LR\_imp, showed relatively less decline of AUROC and AUPRC compared to the other models as  $P$  increased.

Finally, the third analysis was for evaluating the impact of random missingness on model training. As shown in Figure 6(e) and Figure 6(f), the AUROCs and AUPRCs of MAM kept stable when it was retrained using modified training sets with increasing missing data. The performance of AM\_ind was also relatively stable, but its AUROC and AUPRC declined at  $P$  of 0.8. The AUROCs and AUPRCs of AM\_imp kept declining as  $P$  increased, but the decreased extent was obviously less than LR\_imp which also used MI to handle missing data. XGB performed slightly better when it was retrained at  $P$  of 0.2 and 0.4 compared to using the raw training set, but its AUROC and AUPRC declined below MAM, AM\_ind, and even LR\_ind at  $P$  of 0.8. LR\_imp showed the most sharply declined AUROCs and AUPRCs in this analysis. At last, LR\_ind is robust in this

test. Its AUROCs and AUPRCs rose at  $P$  of 0.2, and then kept relatively stable as  $P$  increased.

## Discussion

In this study, we propose an attention architecture for early prediction of hospital mortality. This neural network architecture can achieve a novel approach of filtering out missing data, and is also adaptive to regular MI and missing indicator methods. Our results indicate that the three models based on this attention architecture have excellent performance for early predicting hospital mortality and are robust to data missingness in model training and validation.

Missing data is inevitable in EHRs and should be carefully handled in researches using EHRs to develop prediction model. There are three types of missing mechanism<sup>34</sup>: (a) missing completely at random (MCAR): missingness happens without relationship to any other patient variables; (b) missing at random (MAR): Missingness is related to other observed variables; (c) missing not at random

(MNAR): Missingness is related to some unobserved variables. In theory, the data-missing mechanism should be taken into account for handling missing data.<sup>35</sup> For instance, complete case analysis (deleting samples containing missing data) is generally valid for MCAR but not for MAR; MI is competent for MCAR and MAR<sup>18,19</sup>; missing indicators may introduce bias in handling MAR.<sup>20,36</sup> And all these methods may be inappropriate for MNAR.<sup>34</sup> However, the complication is that sometimes it is difficult and even impossible to distinguish the missing mechanism, especially to recognize MNAR since the unobserved variable is hard to be confirmed. With respect to practical application, we prefer models with excellent predictive performance, and furthermore its performance can keep as stable as possible when increasing missing data is encountered, which is referred to as robustness. Therefore, we design the attention architecture and test it in the most concerned clinical task of mortality prediction. To the best of our knowledge, this is the first study that uses masked attention mechanism to handle missing data and makes a comprehensive analysis of model robustness in both model training and validation.

This study has several advantages. Firstly, we collect sufficient data resources for model development and validation. Two large ICU databases are employed as data source and the extracted clinical variables covers almost all the routine physiological measures for ICU patients. We used MIMIC-IV for model training and eICU-CRD for model external validation, which ensures that the training set and the test set are mutually independent. Our statistical analysis demonstrates heterogeneities between included samples from MIMIC-IV and eICU-CRD, such as the difference in the distribution of admission type, utilization of vasoactive drugs, and proportion of invasive mechanical ventilation. Besides the observable values of clinical variables, their missing rates also show the difference. These challenge the generalization ability of a model when it is trained and validated on these two data sets respectively, and increase the persuasiveness of model performance compared to research on single center or database.

Secondly, we propose a simple and effective attention architecture and a novel approach of filtering out missing data based on the masked mechanism. This architecture only contains one embedding layer, one multi-head attention layer and one linear layer to be tuned during model training. And in the most computationally expensive attention layer, we abandon using the self-attention mechanism proposed in Transformer model,<sup>23</sup> as it needs to compute  $n$  ( $n$  is the number of employed variables) sets of attentions where each set of attentions is computed using query of one variable and key-value pairs of all the variables (including the query variable itself). We introduce a constant vector  $c$  for computing the query instead, and then only one set of attentions of  $c$  on all the variables is computed. The advantage of this design is to avoid that missing variable

which is possibly encountered if we use variables to compute query, and in such a situation this missing variable responsible for computing query is unable to be filtered out from model computation. On the other hand, we reduce the computational cost to  $1/n$  of the self-attention. Based on such an attention architecture, then we can conveniently filter out any missing variable by a mask function.

Thirdly, we explore the interpretability of our proposed attention architecture. We take an insight into the data patterns learned by the three attention-based models through their allocations of average acquired attention among the variables in the external validation. Our results show different patterns of attention allocation among the three models. For MAM, considering the masked mechanism restricting attention allocation to non-missing variables, we wonder whether MAM can capture potential valuable information of high-missing variables as these variables are less likely to be encountered during model training. As the heatmaps of MAM in Figure 5 shown, some previously mentioned high-missing variables (69: Max\_ALT, 70: Max\_AST, 85: PEEP) still acquire high average attention in most MAM instances; while some low-missing variables (29: Mean\_DBP, 30: Std\_DBP, 31: Min\_MAP) acquire low attention. This indicates that a high missing rate will not lead to low attention allocation by MAM. For AM\_imp and AM\_ind, missing variables also obtain attention allocation like non-missing variables. The heatmaps of AM\_imp show a more evenly allocated average attention among variables than AM\_ind (i.e., attention is unlikely to be intensively allocated to minority variables). The probable reason is that the MI model is essentially composed of many multivariate regression models<sup>31</sup> which integrate the information of other non-missing variables to impute missing variables. Therefore, imputed values of an unimportant variable may acquire extra attention when it contains valuable information about other non-missing variables; while the situation is the opposite for an important variable. As a result, the disparity of average acquired attention among all variables will be reduced. Unlike AM\_imp, all missing variables in AM\_ind are uniformly represented by missing indicators. The information about missingness may be valuable when it happens not at random and is related to the outcome.<sup>37,38</sup> For instance, less serious patients have no record of ventilator parameters as they are not intubated. Thus, missingness of ventilator parameters may imply lower mortality. Nevertheless, our result shows that in AM\_ind most variables with high average acquired attention are low-missing variables (variable 3, 5, 6, 51, 81). This is probably because most missing indicators fail to provide sufficient valuable information to the attention architecture for mortality prediction, so high attention is still allocated to the most valuable several non-missing variables. Although the attention allocation makes the attention-based models interpretable rather than to be a black-box model like conventional neural networks, the

clinical rationality of such an interpretation is still needed to be further evaluation.

Fourthly, we provide a comprehensive analysis of model robustness to data missingness in both model training and validation. In the first analysis, our results indicate that all the models generally have lower AUROC and AUPRC in subgroups with higher missing rate. Although data missingness inevitably undermines model performance, our trained attention-based models show advantage of robustness over the baseline models. AM\_imp has higher AUROC and AUPRC in almost all subgroups than LR\_imp (except for AUROC in the 30%–40% subgroup), and so does AM\_ind compared to LR\_ind. This indicates that assisted by the same approach of MI or missing indicator, the attention architecture outperforms LR. MAM has comparable performance as AM\_imp and AM\_ind in most subgroups despite that its AUPRC in the >40% subgroup is relatively low, demonstrating the potential of masked mechanism for handling data missingness. XGB performs slightly better than the attention-based models in the first three subgroups but obviously poorer in the 30%–40% and >40% subgroups, which indicates the limited robustness of XGB for high-missing data. In the second analysis, our results show that the MI model can maintain the robustness of AM\_imp and LR\_imp better than the other approaches when more random missingness is introduced in the test set. However, both AM\_imp and LR\_imp are no longer so robust when we introduce random missingness in the training set in the third analysis, especially LR\_imp. Considering that the MI model integrate non-missing variables to impute missing variables and the missing rate of the training set is lower than the test set (10.3% vs. 19.7%), a probable explanation for the above results is that when MI model is developed on a low-missing training set, it is more likely to learn a valuable data pattern from sufficient non-missing data and effectively impute a high-missing test set; but when a high-missing training set is used, limited available non-missing data may cause the MI to learn a misleading data pattern for imputing the test set. Nevertheless, the final model performance should depend on the prediction model itself as well, and in the second and third analyses, AM\_imp also shows better robustness than LR\_imp, especially in the third analysis, proving the advantage of the attention architecture again. On the other hand, MAM, XGB, AM\_ind, and LR\_ind show opposite results in the second and third analyses. These four models are free of interference by imputed data, and this probably makes them more competent in capturing generalizable data pattern from high-missing training set. In addition, we have not retrained our models using the subgroups in the first analysis to evaluate the impact of inherent missingness on model training. The reason is that sample sizes among these subgroups differ largely, and in this situation the performance of models trained on small subgroups may not only affect by the missing rate but

also by an insufficient sample size, which prevents us to make a fair comparison.

As mentioned above, the three attention-based models show different patterns of attention allocation and different robustness in model training and validation. Based on their characteristics, we propose a preliminary principle for selecting an appropriate model in practice as following: (a) if the training set is low-missing and contains sufficient information to develop an effective MI model, AM\_imp is preferred; (b) if the training set is high-missing and the missingness is strongly related to the outcome, AM\_ind is preferred; (c) if the training set is high-missing and the missingness is weakly related to the outcome, MAM is preferred.

Our study has several limitations. Firstly, we are unable to strictly simulate the missing mechanism of MCAR and MAR, since there is inherent data missingness in our extracted data sets and this inherent missingness probably belongs to MNAR. It is unrealistic to obtain a complete data set without missing data from EHR database as large as MIMIC-IV and eICU-CRD. This limitation can be partly compensated as we analyze the impact of random missingness where the raw data sets with inherent missingness are treated as baseline. Secondly, this attention architecture is not capable of analyzing clinical time series data and providing dynamic prediction. And the so-called last observation carried forward<sup>39</sup> imputation which uses the last observed value to fill current missingness in a time series is not employed for comparison in this study. We plan to design attention-based dynamic prediction model in our future work. Thirdly, in our attention architecture, the average acquired attention can only interpretate the contribution proportion of a variable for the prediction, but is unable to clarify whether the impact of a variable is positive or negative. For instance, for a variable with high attention, it is not clear whether a higher value will raise the mortality or a lower value. At last, we only evaluate our attention architecture in the task of early predicting hospital mortality, therefore its performance and robustness to data missingness are needed to be further validated in other clinical prediction tasks in the future, and our proposed principle for model selection is also needed to be further concretized and validated (such as the detailed criterion for discriminating low-missing set and high-missing set, and the method for quantifying the relationship between the missingness and the outcome).

## Conclusion

Our proposed attention architecture is a simple and interpretable neural network architecture. It can achieve a novel masked mechanism to filter out missing data, and is also adaptive to conventional imputation and missing indicator for handling missing data. The three attention-based models show the state-of-the-art performance and excellent

robustness to data missing in the task of early predicting hospital mortality in ICU patients. Furthermore, in our prediction task the three models show different patterns of attention allocation and different robustness in model training and validation, so the selection of an appropriate model should depend on the specific situation in practice. Overall, the attention architecture has the potential to become an excellent model architecture for clinical prediction tasks with data missingness, and further research is needed to validate its performance and to clarify its applicable conditions.

**Availability of data and materials:** Data of the MIMIC-IV is available on website at <https://mimic-iv.mit.edu/>. Data on the eICU-CRD is available on website at <https://eicu-crd.mit.edu/>. The extracted dataset used during the current study is available from the corresponding author upon reasonable request.

**Contributorship:** XG conceived the idea and the study design, performed statistical analysis of data, and revised the manuscript. ZXZ performed the literature review and manuscript writing. YL and SY helped to collect data. JQL and BX performed the algorithm program. CXL helped to revise English writing. All authors read and approved the final manuscript.

**Declaration of conflicting interests:** The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethics approval:** This study was an analysis of third-party deidentified publicly available databases with pre-existing ethical review board approval.

**Funding:** The authors received the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Scientific Research Project of Hunan Provincial Health Commission (D202310007453) and Beijing Union Medical Fund—Rui E (Ruiyi) Emergency Medical Research Fund (2022, No.12)

**Guarantor:** XG.

**ORCID iD:** Xun Gong  <https://orcid.org/0000-0003-3492-0190>

**Supplemental material:** Supplemental material for this article is available online.

## References

1. Salluh JI and Soares M. ICU Severity of illness scores: APACHE, SAPS and MPM. *Curr Opin Crit Care* 2014; 20: 557–565.
2. Knaus WA, Draper EA, Wagner DP, et al. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13: 818–829.
3. Le Gall JR, Lemeshow S and Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/north American multicenter study. *JAMA* 1993; 270: 2957–2963.
4. Lemeshow S, Teres D, Klar J, et al. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993; 270: 2478–2486.
5. Hu C, Li L, Huang W, et al. Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study. *Infect Dis Ther* 2022; 11: 1117–1132.
6. Huang T, Le D, Yuan L, et al. Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit. *PLoS One* 2023; 18: e0280606. Published 2023 Jan 26.
7. Pattharanitima P, Thongprayoon C, Kaewput W, et al. Machine learning prediction models for mortality in intensive care unit patients with lactic acidosis. *J Clin Med* 2021; 10: 5021. Published 2021 Oct 28.
8. Liu C, Liu X, Mao Z, et al. Interpretable machine learning model for early prediction of mortality in ICU patients with rhabdomyolysis. *Med Sci Sports Exerc* 2021; 53: 1826–1834.
9. Stenwig E, Salvi G, Rossi PS, et al. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Med Res Methodol* 2022; 22: 53. Published 2022 Feb 27.
10. Kong G, Lin K and Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Med Inform Decis Mak* 2020; 20: 251. Published 2020 Oct 2.
11. Yang J, Lim HG, Park W, et al. Development of a machine learning model for the prediction of the short-term mortality in patients in the intensive care unit. *J Crit Care* 2022; 71: 154106.
12. Luo C, Zhu Y, Zhu Z, et al. A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. *J Transl Med* 2022; 20: 136. Published 2022 Mar 18.
13. Delahanty RJ, Kaufman D and Jones SS. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. *Crit Care Med* 2018; 46: e481–e488.
14. Nimgaonkar A, Karnad DR, Sudarshan S, et al. Prediction of mortality in an Indian intensive care unit. Comparison between APACHE II and artificial neural networks. *Intensive Care Med* 2004; 30: 248–253.
15. Nijman S, Leeuwenberg AM, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J Clin Epidemiol* 2022; 142: 218–229.
16. Wood AM, White IR and Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 2004; 1: 368–376.
17. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006; 59: 1087–1091.
18. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Br Med J* 2009; 338: b2393. Published 2009 Jun 29.

19. Cummings P. Missing data and multiple imputation. *JAMA Pediatr* 2013; 167: 656–661.
  20. Groenwold RH, White IR, Donders AR, et al. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ* 2012; 184: 1265–1269.
  21. Feelders A. Handling missing data in trees: Surrogate splits or statistical imputation? In: Zytkow JM and Rauch J (eds) *Principles of data mining and knowledge discovery [internet]*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999. pp.329–334. Accessed 27 July 2021.
  22. Chen T and Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016 Aug 13. pp.785–794.
  23. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30.
  24. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
  25. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
  26. Goldberger AL, Amaral LA, Glas L, et al. Physiobank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000; 101: E215–E220.
  27. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data* 2018; 5: 180178.
  28. Ba JL, Kiros JR and Hinton GE. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.
  29. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778.
  30. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
  31. Austin PC, White IR, Lee DS, et al. Missing data in clinical research: a tutorial on multiple imputation. *Can J Cardiol* 2021; 37: 1322–1331.
  32. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, 1987.
  33. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users’ guides to the medical literature. *JAMA* 2017; 318: 1377–1384.
  34. Schafer JL and Graham JW. Missing data: our view of the state of the art. *Psychol Methods* 2002; 7: 147–177.
  35. Heymans MW and Twisk JWR. Handling missing data in clinical research [published online ahead of print, 2022 sep 21]. *J Clin Epidemiol* 2022; S0895–4356: 00218–9.
  36. Knol MJ, Janssen KJ, Donders AR, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epidemiol* 2010; 63: 728–736.
  37. Madden JM, Lakoma MD, Rusinak D, et al. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *J Am Med Inform Assoc* 2016; 23: 1143–1149.
  38. van Smeden M, Groenwold RHH and Moons KG. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *J Clin Epidemiol* 2020; 125: 188–190.
  39. Lachin JM. Fallacies of last observation carried forward analyses. *Clin Trials* 2016; 13: 161–168.
-