



AAUP Journal of STEM and Health Sciences (JSH-AAUP)

Biostatistical Models for Predicting Mortality in Intensive Care Units

Running Title: Physician Perspectives on ICU Mortality Prediction Models

Saleh Afaneh^{1,*}

¹Department of Mathematics and Statistics, Faculty of Sciences, Arab American University, Palestine

ORCID: <https://orcid.org/0000-0000-0000-0000>

*Corresponding Author Email: saleh.afaneh@aaup.edu

Date Received: 16/08/2026 Date Revised: 05/09/2026 Date Accepted: 20/09/2026

Date Published: 30/09/2026 DOI: <https://doi.org/10.35517/JSH-AAUP.v1.02.08>

Abstract

Intensive care Units clinicians often rely on risk scores. However, the use of a prediction model depends on how it is built, updated, explained and introduced into clinical work. The researchers surveyed 223 physicians working in intensive care settings in the West Bank and asked about eight features of biostatistical models used for mortality prediction: input variables, model type/methodology, calibration and discrimination, temporal resolution, handling of missing data, interpretability, external validation and clinical integration. Accordingly, the responses were analyzed in SPSS using descriptive statistics, Cronbach's alpha, collinearity diagnostics and multiple linear regression. All the eight dimensions showed positive and statistically significant associations with the physician-reported mortality-prediction outcome. The model type/methodology had the largest standardized coefficient ($\beta = 0.288$; $B = 0.256$; $p < 0.001$) and followed by the temporal resolution ($\beta = 0.229$; $B = 0.202$; $p < 0.001$). The full model yielded $R = 0.905$ and $R^2 = 0.819$. The findings showed how the physicians viewed the usefulness of the model characteristics; they did not represent the validation of an ICU mortality model using patient-level outcomes.

Keywords: intensive care unit, mortality prediction, biostatistical models, machine learning, interpretability, external validation, clinical integration.

1. Introduction

Estimating the risk of death is a routine part of intensive care. APACHE, SAPS and SOFA remain familiar tools because they condense several clinical measurements into a severity score that can be used at the bedside. Their simplicity is useful, but an ICU patient can change quickly, sometimes within hours. This has created interest in statistical and machine-learning methods that can use a broader set of measurements and, when data are available, update risk as the clinical picture changes.

Recent ICU studies have explored several such approaches. Khajehali et al. [1] compared machine-learning methods for mortality prediction and reported their best performance with AdaBoost after preprocessing the data. Using MIMIC-IV records from patients with cardiac arrest, Sun et al. [2] evaluated LASSO, XGBoost, logistic regression and NEWS2; LASSO, XGBoost and logistic regression discriminated outcomes better than NEWS2. Jeon et al. [3] examined patients with severe pneumonia and also found better discrimination with machine-learning models than with SAPS II. Taken together, these reports suggest that the modelling method can matter, although its value still depends on the population, the available data and the way the model is used. For clinical use, the data entering a model are just as important as the algorithm itself. Age, comorbidities, vital signs, laboratory results and severity scores are commonly available in ICU records, but their completeness and timing vary. The choice of method then determines how these inputs are combined. Logistic regression has the advantage of being familiar and relatively easy to explain. More flexible methods, including boosting algorithms and neural networks, can model nonlinear patterns, but added complexity is not automatically an advantage at the bedside.

Model quality cannot also be reduced to a single performance number. Discrimination asks whether patients with different outcomes can be separated, whereas calibration asks whether the estimated risks are close to what is actually observed. A model may perform well in the hospital where it was developed and lose accuracy when moved elsewhere. External validation is therefore important, especially in critical care, where case mix, staffing, treatment pathways and resource availability differ between institutions [2,4,5].

Timing creates another difficulty. That is, ICU measurements are collected repeatedly rather than once, so a risk estimate based only on admission data may miss later deterioration or improvement. Studies using real-time vital signs and multivariate time-series data have shown the value of updating predictions as new information becomes available [6,7]. Missing data are equally common in routine records and need to be handled explicitly. Clinicians also need to understand what is driving a prediction and to see the result in a form that fits the existing workflow rather than in a separate research interface [8,9].

Most work in this field has concentrated on the performance of algorithms in patient datasets. Far less is known about which model features physicians themselves consider convincing or workable, particularly in Palestinian hospitals. The researchers, therefore, examined the views of ICU physicians in the West Bank rather than building a new mortality algorithm. The survey covered eight dimensions—input variables, model type/methodology, calibration and discrimination, temporal resolution, missing-data handling, interpretability, external validation and clinical integration—and assessed how each related to the physicians' overall rating of mortality-prediction usefulness.

The conceptual framework used in the study is summarized in Figure 1.

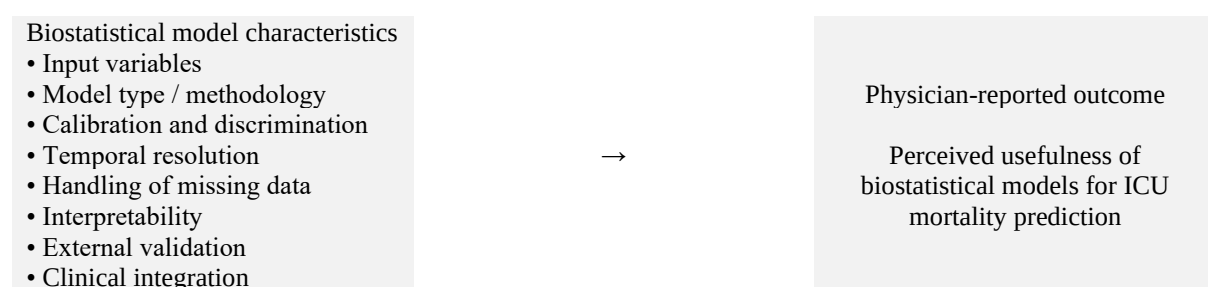


Figure 1. Conceptual framework based on the literature [5,7-12].

Study hypotheses

Main hypothesis: The combined characteristics of biostatistical models are positively associated with the physicians' perceptions of their usefulness for ICU mortality prediction.

H1: Ratings of input variables are positively associated with the physician-reported mortality-prediction outcome.

H2: Ratings of model type/methodology are positively associated with the physician-reported mortality-prediction outcome.

H3: Ratings of calibration and discrimination are positively associated with the physician-reported mortality-prediction outcome.

H4: Ratings of temporal resolution are positively associated with the physician-reported mortality-prediction outcome.

H5: Ratings of missing-data handling are positively associated with the physician-reported mortality-prediction outcome.

H6: Ratings of interpretability are positively associated with the physician-reported mortality-prediction outcome.

H7: Ratings of external validation are positively associated with the physician-reported mortality-prediction outcome.

H8: Ratings of clinical integration are positively associated with the physician-reported mortality-prediction outcome.

2. Materials & Methods

2.1. Study design

The researchers used a cross-sectional survey because the study question was concerned with the physicians' views of mortality-prediction models at one point in time. No patient records were used, and no prediction algorithm was developed or tested in this study.

2.2. Questionnaire development

The questionnaire was prepared after reviewing published work on predictive modelling, interpretability, missing data, temporal modelling, external validation and clinical integration [5,7-12]. It had two sections. The first collected demographic information, including gender, experience and education. The second contained items covering the eight model-related dimensions examined in the study. The physicians recorded their responses on a five-point Likert scale through Google Forms.

2.3. Study setting and participants

The physicians working in intensive care settings in the West Bank, Palestine were population of the study. The researchers invited 290 physicians to participate and retained 223 complete questionnaires after screening, corresponding to a response rate of 76.8%. The sampling frame included government, private, and private or charitable hospitals.

Table 1. Study population source

Source	#
government hospitals	18
private hospitals	18
private or charitable hospitals	22
Total	58

2.4. Statistical analysis

SPSS was used for data screening and statistical analysis. The researchers first described the participants with frequencies and percentages and summarized questionnaire items and dimension scores with means (μ) and standard deviations (σ). Cronbach's alpha was calculated for internal consistency. Tolerance and variance inflation factor (VIF) values were examined before regression analysis. The researchers then fitted a multiple linear regression model with the eight model-related dimensions as predictors of the physician-reported ICU mortality-prediction outcome. A p value below 0.05 was considered statistically significant.

Table 2. Internal consistency of study variables (Cronbach's α)

Variable	α
Input Variables	0.973
Model Type / Methodology	0.938
Calibration and Discrimination	0.948
Temporal Resolution	0.939
Handling of Missing Data	0.928
Interpretability	0.954
External Validation	0.952
Clinical Integration	0.943
Predicting Mortality in (ICU)	0.926

3. Results

3.1. Participant characteristics

Of the 223 respondents, 116 (52.0%) were men and 107 (48.0%) were women. Out of the respondents, 107 physicians (48.0%) were reported to have more than 18 years of professional experience. According to the education categories used in the questionnaire, 151 participants (67.7%) were in the bachelor-level group and 72 (32.3%) were in the high studies category. Table 3 shows the complete distribution.

Table 3. Demographic characteristics of participants

	f	%
Gender		
Male	116	52.0
Female	107	48.0
Experience		
Less than 5 years	33	14.8
6-11	60	26.9
12-17	23	10.3
More than 18 years	107	48.0
Education		
BA	151	67.7
HIGH STUDIES	72	32.3

	f	%
Total	223	100.0

3.2. Questionnaire responses

All of the eight-dimension means were above 3.00, which is the midpoint of the five-point scale. The mean for the overall physician-reported mortality-prediction outcome was 3.840 (SD = 0.897). Input variables received the highest dimension score (mean = 3.819), followed by temporal resolution (3.719), model type/methodology (3.706), external validation (3.611), interpretability (3.527), calibration and discrimination (3.392), clinical integration (3.367), and handling of missing data (3.332). The item-level values are shown in Table 4.

Table 4. Questionnaire item and dimension scores

Statement	μ	σ
The model shows age, sex and comorbidities of patients all the time	3.888	1.018
Vital signs are always flashing including (heart rate, blood pressure, respiratory rate)	3.821	1.080
Indicators of labs values can be seen on screen all the time	3.839	1.044
The Glasgow Coma Scale (GCS) flashes when there are changes	3.848	1.067
Score are always available including APACHE, SAPS, or SOFA scores	3.700	1.228
Input Variables	3.819	1.035
The models have the ability to determine interactions	3.767	1.135
They are able to capture nonlinear relationship	3.381	1.104
Models are all improved in their accuracy	3.803	1.101
They can also improve interpretability of results	3.888	1.163
They are based on logistic regression and XGBoost	3.700	1.129
Model Type / Methodology	3.706	1.009
Models adopted are reliable and able to predict risks	3.350	1.175
They are crucial for clinical utility	3.305	1.222
They depend on calibration plots	3.466	1.273
They can show the Area Under the Receiver Operating Characteristic Curve (AUROC)	3.377	1.224
They are based on Brier Score	3.413	1.193
Calibration and Discrimination	3.392	1.103
Model incorporate time varying data	3.673	1.195
All predictions are updated over time	3.601	1.158
Model are based on real time updating	3.744	1.128
They utilize time-series analysis	3.744	1.108
They operate on recurrent neutral networks (RNNs)	3.834	1.105
Temporal Resolution	3.719	1.021
All models enjoy a good level of robustness and reliability	3.305	1.089
We make sure that models are able to avoid biased mortality prediction	3.256	1.163
Model can ignore and exclude incomplete cases and data	3.121	1.090
Model are always aware of lost information (imputation)	3.345	1.148
ICU environment must be able to handle missing data	3.632	.991
Handling of Missing Data	3.332	.967
There is a continuous risk assessment	3.623	.973
Models aren't based on one-time scoring of data	3.547	1.007
The updating is always in real-time	3.493	.999
This can increase trust and adoption	3.565	1.002
Modifiable factors are always up-to-date	3.390	1.157
Interpretability	3.527	.951
Models can generate data outside dataset	3.668	1.251
Models can prevent overfitting	3.318	1.194
They present a broader clinical application	3.691	1.230
They can generate data from different patients at the same time	3.789	1.289

Statement	μ	σ
Generalizability is possible based on certain conditions that can't be ignored	3.619	1.231
External Validation	3.611	1.131
All model fit with ICU's electronic health record (EHR) systems	3.332	1.158
They provide a real-time decision-making	3.274	1.212
They present highly integrated alert tools	3.408	1.273
They are based on clinicians point of care	3.372	1.193
They make sure that all tools are usable for predictions	3.448	1.176
Clinical Integration	3.367	1.086
ICU environment depends highly on biostatistical models	3.839	1.099
All ICU's tasks are complex and must be maintained in a developed approach	3.816	1.064
Predicting patients' risk of death is highly crucial medical aid	3.946	1.043
Biostatistical models can help ICU make decisions and allocate resources	3.964	.990
Biostatistical characteristics are crucial for effective ICU efforts	4.013	1.002
Predicting mortality rate is important when it comes to clinical integration	3.462	1.093
Predicting Mortality in (ICU)	3.840	.897

3.3. Collinearity assessment

VIF values ranged from 1.519 to 5.142, with corresponding tolerance values of 0.194 to 0.659. None reached the conventional VIF threshold of 10 used in the original analysis. The highest value was observed for model type/methodology (VIF = 5.142), indicating some overlap with the other predictors and suggesting that its individual coefficient should be interpreted with care.

Table 5. Collinearity statistics

	Tolerance	VIF
Input Variables	.535	1.870
Model Type / Methodology	.194	5.142
Calibration and Discrimination	.332	3.008
Temporal Resolution	.659	1.519
Handling of Missing Data	.550	1.818
Interpretability	.406	2.461
External Validation	.294	3.403
Clinical Integration	.288	3.472

3.4. Regression analysis

The eight predictors were strongly associated with the physician-reported mortality-prediction outcome in the multiple regression model ($R = 0.905$; $R^2 = 0.819$). In other words, the model accounted for 81.9% of the variation in the questionnaire outcome. This percentage concerns variation in respondents' ratings; it is not the proportion of real ICU deaths explained or predicted by the model.

Each regression coefficient was positive and significant at $p < 0.05$. Model type/methodology had the largest standardized coefficient ($\beta = 0.288$; $B = 0.256$; $p < 0.001$), and temporal resolution was next ($\beta = 0.229$; $B = 0.202$; $p < 0.001$). The remaining coefficients were external validation ($\beta = 0.136$; $p = 0.012$), calibration and discrimination ($\beta = 0.112$; $p = 0.027$), clinical integration ($\beta = 0.109$; $p = 0.045$), interpretability ($\beta = 0.094$; $p = 0.042$), input variables ($\beta = 0.085$; $p = 0.034$), and handling of missing data ($\beta = 0.083$; $p = 0.036$). The direction of H1-H8 was therefore supported within this questionnaire-based analysis.

Table 6. Multiple regression results for the study hypotheses

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.292	.126		2.317	.021
Input Variables	.074	.034	.085	2.134	.034
Model Type / Methodology	.256	.059	.288	4.357	.000
Calibration and Discrimination	.091	.041	.112	2.225	.027
1 Temporal Resolution	.202	.032	.229	6.397	.000
Handling of Missing Data	.077	.036	.083	2.112	.036
Interpretability	.088	.043	.094	2.050	.042
External Validation	.108	.043	.136	2.532	.012
Clinical Integration	.090	.045	.109	2.012	.045

4. Discussion

The survey shows that the physicians did not view ICU mortality prediction as a matter of algorithm choice alone. Ratings across all eight dimensions were positively related to the overall questionnaire outcome, with model type/methodology and temporal resolution showing the strongest associations. Because the outcome was based on the physicians' responses, the result is best read as a picture of what clinicians consider useful in a prediction system, not as evidence that a particular model predicts death accurately.

Model type/methodology ranked first in the regression. That is plausible: the modelling approach affects both the patterns a system can capture and how readily a clinician can understand its output. Logistic regression remains attractive when transparency is important, whereas boosting methods and neural networks can represent more complicated relationships. ICU studies have reported good results with AdaBoost, XGBoost, LightGBM and multilayer perceptron models [1-3]. Our survey did not test one method against another; it indicates that physicians pay attention to the method behind the prediction.

Temporal resolution came next. In intensive care, many of the variables that matter—vital signs, laboratory tests, ventilatory support and treatment—change during the admission. A risk score calculated once may therefore become outdated. Work based on real-time vital signs and multivariate time-series modelling supports the value of repeated updating [6,7]. The relatively high score for temporal resolution in this survey suggests that this issue is also recognized by the clinicians working in the study setting.

External validation, calibration and discrimination were also associated with the outcome. A model can separate survivors from non-survivors reasonably well and still give poorly calibrated risk estimates. Likewise, the results obtained in one hospital may not carry over to another. Sun et al. [2] examined both discrimination and calibration, and Hargovan et al. [4] focused specifically on external validation of sepsis mortality models. The responses in our study suggest that physicians see these checks as part of deciding whether a model is credible enough for wider use.

Interpretability and clinical integration were significant as well. A prediction that sits outside the normal clinical workflow, or one that gives a risk number without an understandable basis, is unlikely to be used consistently. EHR integration, clear displays and access to the main contributors to a prediction may help, but they do not

replace validation. Previous work on interpretable machine learning and transparent ICU decision support raises the same concern [5,7,8]. Tools such as SHAP can assist explanation, but the clinical value of an explanation still depends on the quality of the underlying model.

Input variables and missing-data handling produced smaller standardized coefficients, although both remained significant. The researchers do not interpret the smaller coefficients as evidence that these issues are secondary. ICU records often contain irregular or incomplete measurements, and the way missing values are treated can alter model behavior. The same applies to input selection: even a sophisticated method will perform poorly if the variables are unreliable or not clinically relevant. Bias and missingness in ICU electronic records have been discussed repeatedly in the literature [9,11].

The size of R^2 requires particular caution. Predictor and outcome scores came from the same questionnaire, and several items deal with related aspects of the same general idea. Shared wording, overlap between constructs and common-method effects may therefore have increased the observed association. Model type/methodology also had a VIF just above 5. These points do not remove the associations seen in the survey, but they make it unwise to treat the individual coefficients as if they were independent measures of clinical effectiveness.

The local setting is worth noting. Published work on predictive analytics rarely includes ICU physicians from the West Bank. This survey gives a first indication of the features that clinicians in this setting say they would value in a mortality-prediction system. It does not show that any existing model is accurate for Palestinian ICU patients. That would require patient-level data and formal model validation.

4.1. Limitations and future directions

Several limitations follow from the design. The sample was restricted to ICU physicians in the West Bank, and all measures were self-reported at a single time point. Predictors and outcome were answered by the same respondents, while some constructs are closely related, so common-method variance and item overlap may have inflated the regression. No patient-level clinical data were analyzed; consequently, this study cannot provide AUROC, calibration, sensitivity, specificity or evidence of reduced mortality for any algorithm. A next study should test candidate models on local ICU data, compare simpler transparent approaches with more flexible machine-learning methods, and include internal and external validation. If an EHR-based risk tool is later introduced, its effect on decisions, alert burden and possible bias should also be evaluated prospectively.

5. Conclusion

The physicians surveyed in this study associated the usefulness of ICU mortality-prediction models with a combination of technical and practical features. Model type/methodology and temporal resolution had the strongest associations, but external validation, calibration and discrimination, clinical integration, interpretability, input variables and missing-data handling were also significant. For the West Bank, these findings are better used to guide the design and evaluation of future tools than to justify immediate clinical deployment. Candidate models should first be tested on patient-level ICU data and shown to be calibrated, discriminative and reasonably generalizable. Explainable output and EHR integration can then be considered as implementation features rather than substitutes for validation.

6. Declarations

• Clinical trial number

Not applicable

• Ethics approval and consent to participate

This study was approved by the Institutional Review Board of the Arab American University, Palestine (Approval No. AAUP-IRB: J-2025/A 95/N). All participants provided written informed consent prior to participation.

• Consent for publication

Written informed consent for publication was obtained from all relevant participants.

• Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

• Competing interests

The authors declare that they have no competing interests.

• Funding

This research received no external funding.

• Authors' contributions

S.A. conceived and designed the study, collected the data, performed the analysis, and drafted the manuscript. All authors reviewed and approved the final version.

• Acknowledgements

S.A. would like to acknowledge the Arab American University - Scientific Research Council.

• Authors' information

Saleh Afaneh is an Assistant Professor in Statistics at the department of Mathematics and Statistics, the Arab American University, Palestine.

References

- [1] Khajehali N, Khajehali Z, Tarokh MJ. The prediction of mortality influential variables in an intensive care unit: a case study. *Personal and Ubiquitous Computing*. 2023;27(2):203-219.
- [2] Sun Y, He Z, Ren J, Wu Y. Prediction model of in-hospital mortality in intensive care unit patients with cardiac arrest: a retrospective analysis of MIMIC-IV database based on machine learning. *BMC Anesthesiology*. 2023;23(1):178.
- [3] Jeon ET, Lee HJ, Park TY, Jin KN, Ryu B, Lee HW, Kim DH. Machine learning-based prediction of in-ICU mortality in pneumonia patients. *Scientific Reports*. 2023;13(1):11527.
- [4] Hargovan S, Simpson C, Sivalingam S, Carter A, Gunnarsson R. External validation of eight different models to predict sepsis mortality in intensive care units. *Journal of Critical Care*. 2025;90:155174.
- [5] Li S, Dou R, Song X, Lui KY, Xu J, Guo Z, et al. Developing an interpretable machine learning model to predict in-hospital mortality in sepsis patients: a retrospective temporal validation study. *Journal of Clinical Medicine*. 2023;12(3):915.
- [6] Pan P, Wang Y, Liu C, Tu Y, Cheng H, Yang Q, et al. Revisiting the potential value of vital signs in the real-time prediction of mortality risk in intensive care unit patients. *Journal of Big Data*. 2024;11(1):53.
- [7] Xu Z, Guo J, Qin L, Xie Y, Xiao Y, Lin X, et al. Predicting ICU interventions: a transparent decision support model based on multivariate time series graph convolutional neural network. *IEEE Journal of Biomedical and Health Informatics*. 2024;28(6):3709-3720.
- [8] Liu J, Duan X, Duan M, Jiang Y, Mao W, Wang L, Liu G. Development and external validation of an interpretable machine learning model for the prediction of intubation in the intensive care unit. *Scientific Reports*. 2024;14(1):27174.
- [9] Shi J. Advanced Biostatistical Machine Learning for Complex Biomedical Data: Causal Dose-Response Curve Estimation, ICU EHR Bias and Missing Data Evaluation, and Cancer-Drug Biclustering and Response Prediction [doctoral dissertation]. University of California, Berkeley; 2024.
- [10] Choudhury P, Goel P. Interface of Artificial Intelligence with Conventional Biostatistics in Healthcare Research. In: *Artificial Intelligence (AI) in Cell and Genetic Engineering*. Humana; 2025. p. 411-428.
- [11] Pieniak O. Generative Machine Learning for Precision Medicine [master's thesis]. Oslo Metropolitan University; 2024.
- [12] Yoosoofsah A. Predicting Extubation Failure in Intensive Care: The Development of a Novel, End-to-End Actionable and Interpretable Prediction System. *arXiv preprint arXiv:2412.00105*. 2024.