Scientific Research Publishing

Open Journal of Statistics, 2025, 15(1), 73-80
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

An Immediate Mortality Prediction Score That Is Robust to Missing Data

Tara M. Westover[1*], Marta B. Fernandes[1*], M. Brandon Westover[2,3], Sahar F. Zafar[1#]

[1]Department of Neurology, Massachusetts General Hospital (MGH), Boston, MA, USA
[2]Clinical Data Animation Center (CDAC), Harvard Medical School, Boston, MA, USA
[3]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA
Email: #sfzafar@mgh.harvard.edu

How to cite this paper: Westover, T.M., Fernandes, M.B., Westover, M.B. and Zafar, S.F. (2025) An Immediate Mortality Prediction Score That Is Robust to Missing Data. *Open Journal of Statistics*, **15**, 73-80.
https://doi.org/10.4236/ojs.2025.151005

Received: January 19, 2025
Accepted: February 21, 2025
Published: February 24, 2025

Abstract

Objective: To develop an illness severity score that predicts short-term mortality, based on a small number of readily available measurements, and overcomes limitations of the SOFA score, for use in research involving large-scale electronic health records. **Design:** Retrospective analysis of electronic records for 37,739 adult inpatients. **Setting:** A single tertiary care hospital system from 2016-2022. **Patients:** 37,739 adult ICU patients. **Interventions:** IMPS was developed using logistic regression with the 6 SOFA components, age, sex and missingness indicators as predictors, and 10-day mortality as the outcome. This was compared with SOFA with median imputation. **Measurements and Main Results:** Discrimination was evaluated by AUROC, calibration by comparing predicted and observed mortality. IMPS showed excellent discrimination (AUROC 0.80) and calibration. It outperformed SOFA alone (AUROC 0.70) and with age/sex (0.74). **Conclusions:** By retaining continuous data, adding age, allowing for missingness, and optimizing weights based on empirical mortality association, IMPS achieved substantially better mortality prediction than the original SOFA.

Keywords

Critical Care, Missing Data, Electronic Health Records, Illness Severity, Mortality

1. Introduction

The SOFA score is a widely used measurement of disease severity [1]-[4]. SOFA has 6 integer subscores ranging from 0 to 4 for different organ systems, added up

*Co-first authors.
#Corresponding author.

to obtain a total score [1]. Advantages of SOFA include simplicity, interpretability, wide applicability due to reliance on commonly available measurements, ability to track illness temporally, and correlation with mortality.

However, the SOFA score has drawbacks: it discards information by discretizing analog values; it was designed by consensus rather than optimization; its categories are arbitrarily given equal weight; and it does not include age, which is strongly associated with mortality and arguably should be part of a disease severity score. Finally, it is common for variables required by SOFA to be missing, particularly in retrospective studies such as large-scale medical records data studies [4]-[8]. As SOFA does not accommodate missing data, investigators have proposed a variety of ad hoc imputation methods, including median imputation, zero imputation, multiple imputation, and last observation carried forward [3] [4] [6] [7] [9].

Here, we develop an illness severity score suitable for large-scale studies that involve missing data, such as the IMPS (immediate mortality prediction score). IMPS retains the strengths of SOFA score while addressing the weaknesses described above. IMPS is based on the same variables as SOFA plus age and sex, and allows for missing variables. IMPS also has a stronger association with mortality, enhancing its value as an illness severity score.

2. Methods

This study follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement. The Mass General Brigham IRB reviewed and approved the study (protocol#: 2013P001024), and waived the requirement for informed consent. All study procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (the institutional IRB) and with the Helsinki Declaration of 1975, as most recently amended. Data were from inpatients (age ≥ 18) admitted to intensive care units at a single center between Jan 18, 2016 - Dec 31, 2022. For each patient we collected SOFA component variables from the first 24 hours of hospital admission, plus age, sex, and for those who died, time of death. Where multiple measurements were available, the worst measurements were considered (*i.e.*, minimum Glasgow Coma Score (GCS), mean arterial blood pressure (MAP), platelets, percent arterial oxygen to fraction of inspired oxygen ratio (PaO2/FiO2), and maximum bilirubin, creatinine, dopamine, epinephrine, norepinephrine). Outliers (99^{th} percentile values for PaO2/FiO2 ratio, bilirubin, platelets, creatinine, urine output, dopamine dosage, and pressor doses, and first percentile values for PaO2/FiO2, platelets, creatinine, and urine output) were assumed to be errors because most such values were clinically implausible, and were therefore treated as missing after looking at box plots. Data was randomly split into training and testing sets of equal size (50:50).

IMPS was created by logistic regression with SOFA components, age, sex, and "missingness" indicators for all variables except age, sex, ventilator status, and amounts of medication, which were never missing. The target variable was death

within 10 days. These variables were included to keep IMPS close to the spirit and simplicity of SOFA, while addressing it's drawbacks.

For comparison, we also calculated conventional SOFA scores [1]. Missing values were median imputed, as this is commonly done in the literature. We converted SOFA scores to a [0, 1] scale by fitting a logistic regression model with death within 10 days as the target variable. We also fit a logistic regression model that included SOFA plus age and sex as predictors.

All logistic regression models (SOFA, SOFA + age + sex, IMPS) were trained using LASSO regularization with 5-fold internal cross validation to select the regularization parameter.

Performance Evaluation:

Models were compared using area under the receiver operating characteristic curves (AUROC). We also evaluated calibration of the IMPS model: how closely predicted risk of death matches the observed rate [10]. The calibration curve was computed by dividing IMPS mortality predictions into 100 bins, with bin centers at regular intervals between 0.05 and 0.95, and calculating the percentage of patients who died (within 10 days) whose predictions were in ±0.05 of the bin center. Uncertainty for ROC and calibration curves was quantified by 95% confidence intervals (CIs), calculated via using 10,000 iterations of bootstrapping. Model evaluation was done exclusively using testing data.

3. Results

3.1. Cohort Characteristics

Our cohort comprised 37739 patients (**Table 1**). The average age was 62 years (standard deviation 16.5); most were male (60%), white (79%) and non-Hispanic (82%), and 9.4% (N = 3562) of patients died. The cohort median SOFA score was 5 [IQR: 4 - 9] [range 0 - 22].

Patients who died had, relative to survivors (**Table 1**), higher average age (68 vs 61 years), SOFA score (8.96 vs 6.13), total bilirubin (1.65 vs 1.04 mg/dL), PaO2/FiO2 (198.92 vs 189.32) and creatinine (1.96 vs 1.31 mg/dL), and lower average GCS (8.1 vs 11.6), daily urine output (1250 vs 1395 mL), MAP (66 vs 74), and platelet levels (175 K vs 185 K/mL). Patients who died received more pressors (% on Dobutamine: 4.5% vs 1.2%); average maximum infusion rate of Dopamine (0.89 vs 0.28 μg/kg/min), Epinephrine (10.7 vs 1.2 μg/kg/min), Norepinephrine (43.4 vs 8.4 μg/kg/min). Percentage of values missing were largest for PaO2/FiO2 (63%), GCS (38%), and total bilirubin 20%, with less missingness for urine output (9%), platelets (3%), creatinine (3%), and MAP (2%).

3.2. Model Performance

Discrimination: To assess association between model scores and mortality, we plotted ROC curves (**Figure 1(a)**). SOFA alone shows a moderate association with mortality (AUC 0.7), which increased by 4% when including age (AUC 0.74). IMPS exhibits a substantially stronger association (AUC 0.8).

Table 1. Characteristics of patients in the total dataset.

	All	Died	Survived	SD	% missing
Number (%)	37,739 (100)	3562 (9.4)	**34,177 (90.6)**	--	--
Demographics					
Sex (% male)	60%	57%	**60%**	--	0%
Age (years)	62.22	**68.14**	61.6	16.53	0%
Race					
Black	6%	5%	**6%**	--	--
White	79%	77%	**79%**	--	--
Asian	3%	3%	**4%**	--	--
Unknown	12%	**15%**	12%	--	--
Ethnicity					
Non-hispanic	82%	75%	**83%**	--	--
Hispanic	7%	5%	**7%**	--	--
Unknown	11%	**20%**	10%	--	--
SOFA components					
GCS	11.25	8.09	**11.63**	4.6	38.36%
Platelets (10^3/mL)	183.85	174.61	**184.79**	84.76	2.73%
Urine output (mL/day)	1382	1250	**1395**	1085	8.66%
MAP (mm Hg)	73.48	66.28	**74.22**	15.3	2.32%
SOFA score*	6.39	**8.96**	6.13	3.48	--
PaO2/FiO2	190.53	**198.92**	189.32	117.91	63.2%
MV or CPAP (%)	34.14	**55.42**	31.92	--	0%
Total bili (mg/dL)	1.11	**1.65**	1.04	1.7	20.23%
Creatinine (mg/dL)	1.37	**1.96**	1.31	1.09	2.68%
On dobutamine (%)	1.54	**4.52**	1.23	--	0%
Dopamine (μg/kg/min)	0.34	**0.89**	0.28	4.86	0%
Epinephrine (μg/kg/min)	2.13	**10.68**	1.24	16.19	0%
Norepinephrine (μg/kg/min)	11.69	**43.37**	8.39	35.43	0%

Numbers in the all, died, and survived columns are mean values or percentages. SD = standard deviation. GCS = Glasgow coma score, PaO2/FiO2 = ratio of % oxygen in arterial blood to % inspired oxygen, Bili = bilirubin, mL = milliliters, dL = deciliters, MAP = mean arterial blood pressure, μg = micrograms, kg = kilograms. The all group includes all patients; died and survived included those who died and survived within 10 days of the measured SOFA variables, respectively. For comparing the "died" and "survived" groups the larger value is bolded. * Where values are missing, SOFA scores are computed with median imputation.

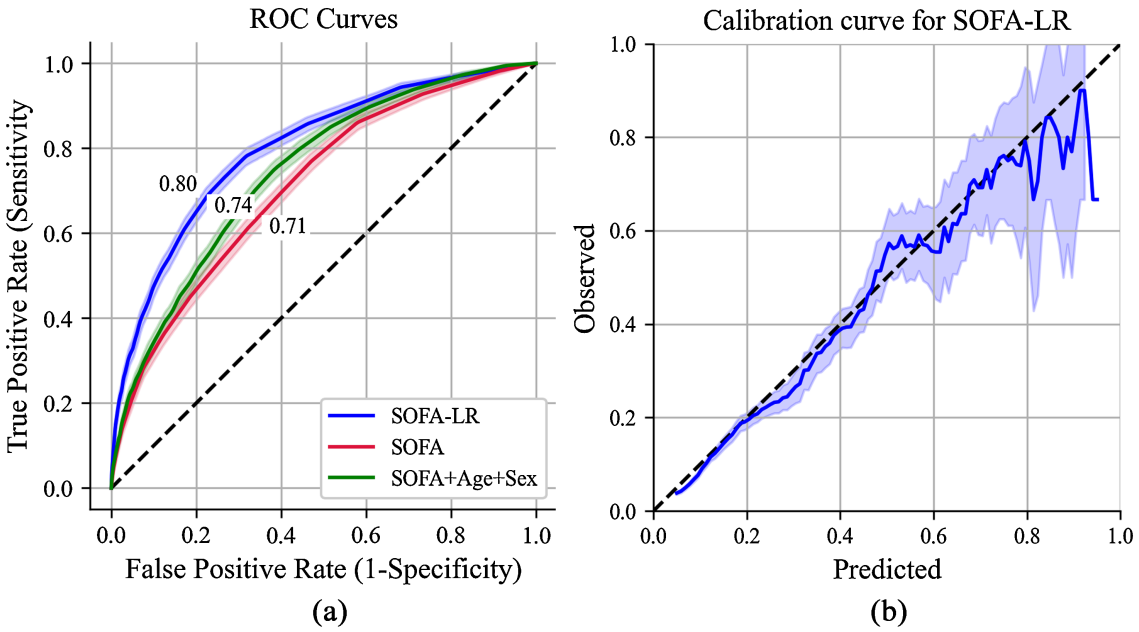

Figure 1. Comparison of model performance. (a) Model discrimination. Receiver operating characteristic (ROC) curves are shown for logistic regression models containing the SOFA score (red), SOFA score plus age and sex (green), and the new IMPS model. Area under the ROC curve (AUC) is shown for each model. IMPS is better than the SOFA model (10% higher AUC) and the SOFA + age + Sex model (6% higher AUC). (b) Model calibration. Calibration curve for IMPS, with 95% confidence band. Predicted (x-axis) is the probability of death within 10-days from IMPS, and observed (y-axis) is the percentage of patients who died within 10-days of the given score.

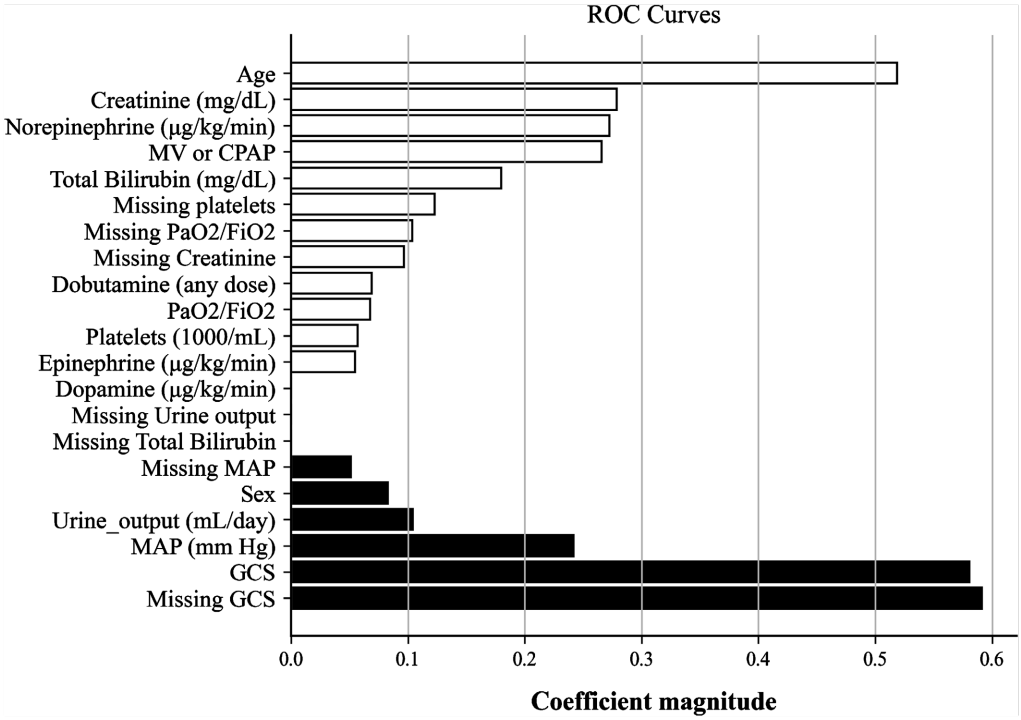

Figure 2. Coefficient values of IMPS Magnitudes of weights in IMPS. Weights with negative values are shown in black, positive weights are shown in white. GCS = Glasgow coma scale, mg = milligrams, dL = deciliter, MV = mechanical ventilation, CPAP = continuous positive airway pressure, PaO2 = pressure of arterial oxygen, FiO2 = fraction of inspired oxygen, mL = milliliters, = micrograms, kg = kilograms, min = minute, mmHg = millimeters of mercury.

Calibration: To assess agreement between risk predicted by IMPS and proportion of people who died, we calculated a calibration curve (**Figure 1(b)**). The calibration curve closely follows the diagonal throughout the range of model predictions, indicating excellent calibration.

3.3. Coefficients

Coefficients of the IMPS model (**Figure 2**) showed clinically reasonable signs, positive and negative. Positive weights, indicating that higher values increase the risk of mortality, were found for age, creatinine, mechanical ventilation or Continuous Positive Airway Pressure (CPAP), total bilirubin, PaO2/FiO2, platelets, and use of pressors. Missingness was assigned positive weights for platelets, PaO2/FiO2, and creatinine. Negative weights, indicating that higher values decrease the predicted risk of mortality, were found for male sex, urine output, MAP, and GCS, and for missingness of MAP and GCS. The model assigned zero weight to presence of dopamine and to missingness of urine output and bilirubin levels.

4. Discussion

IMPS integrates measurements from multiple organ systems that are readily available and easily repeatable, and thus represents a general-purpose hospital illness severity score that can be tracked over time. IMPS is designed via optimization (machine learning), and shows both excellent predictive validity and calibration. IMPS explicitly allows for missing values, making it suitable for retrospective studies that leverage large-scale electronic health records.

4.1. Comparison with SOFA

IMPS has several advantages compared to SOFA. Where SOFA arbitrarily assigns equal weights, IMPS assigns weights based on strength of variable associations with mortality. Where SOFA bins data into discrete categories, IMPS retains raw continuous data, which contributes to prognostic accuracy. SOFA omits age as a component, with the stated intention of enabling comparison of illness levels across patients of different ages. However, the same level of organ dysfunction arguably represents a more severe illness (*i.e.*, higher association with mortality) in older patients, thus IMPS includes age. These three factors (inclusion of raw data and age, and optimized design) likely account for the superior performance of IMPS as a measure of disease severity. Finally, whereas studies that use SOFA often resort to ad hoc imputation resulting in statistical bias [3] [4] [6] [7] [9], IMPS handles missing data through missingness indicators. Overall, IMPS score retains the simplicity and interpretability of an additive score while improving handling of missing data and achieving a stronger association with short-term mortality. It is worth noting that both the conventional SOFA score and IMPS can be computed on a daily schedule.

4.2. Limitations and Future Directions

This study has limitations. First, our data derives from a single hospital system,

and validation in external cohorts is needed. Second, we assessed short-term (10-day) mortality; evaluating longer- and shorter-term outcomes is also important, as is investigating how changes in IMPS from day to day relate to mortality. Third, we evaluated overall inpatient mortality, which was in line with our objective of developing a generalized measure of illness severity. Nevertheless, additional information is required to account for mortality specifically due to presenting illnesses. Fourth, we optimized weights based on association with mortality; associations with other clinical outcomes remain to be explored. Fifth, our model is based on only data from the first 24 hours after admission, so its predictive ability using data collected later is unknown. Finally, we did not explore incorporating IMPS into the electronic health record. This is a possible direction for future studies.

5. Conclusion

IMPS is an optimized alternative to the SOFA score. IMPS incorporates strengths of SOFA including simplicity and reliance on commonly available data, while exhibiting a stronger association with short-term mortality and ability to handle missing data. IMPS represents a promising and versatile replacement for the SOFA score.

Availability of Data and Code

All code and data necessary to reproduce the results, including figures, are made available in a GitHub repository https://github.com/bdsp-core/IMPS.

Funding

Dr Zafar was supported by a grant from the NIH (K23NS114201). Dr. Westover was supported by grants from the NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598), and NSF (2014431).

Disclosures

Dr. Zafar is a consultant for Corticare and receives royalties from Springer for the Acute Neurology Survival Guide. Dr. Westover is a co-founder, scientific advisor, consultant, and has personal equity interest in Beacon Biosignals. Other authors report no disclosures.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., de Mendonça, A., Bruining, H., *et al.* (1996) The SOFA (Sepsis-Related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure. *Intensive Care Medicine*, **22**, 707-710.

https://doi.org/10.1007/bf01709751

[2] Lambden, S., Laterre, P.F., Levy, M.M. and Francois, B. (2019) The SOFA Score—Development, Utility and Challenges of Accurate Assessment in Clinical Trials. *Critical Care*, **23**, Article No. 374. https://doi.org/10.1186/s13054-019-2663-7

[3] Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., *et al.* (2016) The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Journal of the American Medical Association*, **315**, 801-810. https://doi.org/10.1001/jama.2016.0287

[4] Vincent, J., de Mendonca, A., Cantraine, F., Moreno, R., Takala, J., Suter, P.M., *et al.* (1998) Use of the SOFA Score to Assess the Incidence of Organ Dysfunction/Failure in Intensive Care Units. *Critical Care Medicine*, **26**, 1793-1800. https://doi.org/10.1097/00003246-199811000-00016

[5] Brinton, D.L., Ford, D.W., Martin, R.H., Simpson, K.N., Goodwin, A.J. and Simpson, A.N. (2022) Missing Data Methods for Intensive Care Unit SOFA Scores in Electronic Health Records Studies: Results from a Monte Carlo Simulation. *Journal of Comparative Effectiveness Research*, **11**, 47-56. https://doi.org/10.2217/cer-2021-0079

[6] Neto, A.S. (2016) Epidemiological Characteristics, Practice of Ventilation, and Clinical Outcome in Patients at Risk of Acute Respiratory Distress Syndrome in Intensive Care Units from 16 Countries (PRoVENT): An International, Multicentre, Prospective Study. *The Lancet Respiratory Medicine*, **4**, 882-893.

[7] Ferreira, F.L. (2001) Serial Evaluation of the SOFA Score to Predict Outcome in Critically Ill Patients. *Journal of the American Medical Association*, **286**, 1754-1758. https://doi.org/10.1001/jama.286.14.1754

[8] Braasch, M.C., Halimeh, B.N. and Guidry, C.A. (2022) Availability of Multiple Organ Failure Score Components in Surgical Patients. *Surgical Infections*, **23**, 178-182. https://doi.org/10.1089/sur.2021.265

[9] Vasilevskis, E.E., Pandharipande, P.P., Graves, A.J., Shintani, A., Tsuruta, R., Ely, E.W., *et al.* (2016) Validity of a Modified Sequential Organ Failure Assessment Score Using the Richmond Agitation-Sedation Scale. *Critical Care Medicine*, **44**, 138-146. https://doi.org/10.1097/ccm.0000000000001375

[10] Alba, A.C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P.J., *et al.* (2017) Discrimination and Calibration of Clinical Prediction Models. *Journal of the American Medical Association*, **318**, 1377-1384. https://doi.org/10.1001/jama.2017.12126