Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 7 (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Applications of Data Science Techniques in Predicting Mortality among Pediatric ARDS Patients: A Medico-Statistical Approach

Peddapalegani Palavardhan[1], M. Bhupathi Naidu[2], Akkyam Vani[3], Dr. Mohammad Reshma[4*]

[1]Research Scholar, Department of Statistics, Sri Venkateswara University, Tirupati, A.P, India

[2]Professor and Registrar, Department of Statistics, Sri Venkateswara University, Tirupati, A.P, India 3Research Scholar, Department of Statistics, Sri Venkateswara University, Tirupati, A.P, India.

[*4]Assistant Professor, Department of Anatomy, Malla Reddy Medical College for Women, Hyderabad, Telangana, India

***Corresponding Author:**

Dr. Mohammad Reshma

Assistant Professor, Department of Anatomy, Malla Reddy Medical College for Women, Hyderabad, Telangana, India

Cite this paper as: Peddapalegani Palavardhan, M. Bhupathi Naidu, Akkyam Vani, Dr. Mohammad Reshma, (2025) Applications of Data Science Techniques in Predicting Mortality among Pediatric ARDS Patients: A Medico-Statistical Approach. *Journal of Neonatal Surgery*, 14 (7), 1118-1131.

ABSTRACT

This study explores the application of data science techniques in identifying key predictors of mortality among pediatric patients with Acute Respiratory Distress Syndrome (ARDS). A retrospective analysis was conducted on 200 mechanically ventilated patients admitted to the Pediatric Intensive Care Unit (PICU) of the Department of Pediatrics, Malla Reddy Narayana Multispeciality Hospital. Clinical and ventilatory variables at 0 and 24 hours were analyzed using Python. Statistical methods included descriptive statistics, paired t-tests, point-biserial correlation, and multivariable logistic regression. Variance Inflation Factor (VIF) was applied to address multicollinearity, and ROC curve analysis was used to assess model performance. The Sequential Organ Failure Assessment (SOFA) score and Pediatric Risk of Mortality (PRISM) score emerged as significant independent predictors of mortality. The final logistic model achieved an AUC of 0.85, with a sensitivity of 77.3% and specificity of 84.4%. These findings highlight the potential of integrating data science techniques into clinical prediction modeling to improve early risk stratification in pediatric ARDS.

1. INTRODUCTION

1.1 What is Pediatric ARDS

Pediatric Acute Respiratory Distress Syndrome (ARDS) is a life-threatening lung condition in children that causes severe breathing problems and low oxygen levels. It usually happens in very sick children who are admitted to the Pediatric Intensive Care Unit (PICU). Many of these children need ventilator support to help them breathe. The condition can change quickly and is difficult to manage, even with good medical care. Pediatric ARDS is different from adult ARDS in terms of causes, response to treatment, and survival outcomes. Because of this, predicting how a child with ARDS will respond to treatment and whether they will recover or not is a challenge for doctors and healthcare teams. (**Yehya & Thomas, 2017**).

1.2 Why Predicting Mortality is Important

Accurately predicting outcomes in pediatric ARDS is essential for early intervention, risk stratification, and optimal resource utilization. Commonly used scoring systems, such as the Pediatric Risk of Mortality (PRISM) score and the Sequential Organ Failure Assessment (SOFA) **(Pollack et al., 1988; Vincent et al., 1996)**, provide an estimate of disease severity. However, these scores alone may not fully capture the dynamic physiological changes in critically ill children. Additional parameters such as ventilator settings, oxygenation indices, and cardiovascular support scores may provide complementary predictive value and enhance clinical decision-making.

1.3 Role of Data Science in Statistics

Data science offers a powerful framework for analyzing complex clinical data using programming-based statistical techniques (Rajpurkar et al., 2022). In this study, we used Python software to apply various statistical and machine learning tools to identify key predictors of mortality. Specifically, we used descriptive statistics, point-biserial correlation, logistic

regression, multicollinearity testing (via Variance Inflation Factor), and visualization tools to derive meaningful patterns (Pedregosa et al., 2011). These methods improve interpretability, reproducibility, and the potential for integration into future predictive dashboards (Kuhn & Johnson, 2013; Lundberg & Lee, 2017).

1.4 Aim and Objectives

Aim: To apply data science and statistical methods to identify key predictors of mortality in pediatric ARDS patients.

Objectives:

a) To perform descriptive and comparative statistical analysis of clinical and ventilatory variables.

b) To evaluate changes in key variables between 0 and 24 hours of ventilation.

c) To identify significant predictors of mortality using correlation and logistic regression.

d) To assess multicollinearity using Variance Inflation Factor (VIF).

To implement Python-based data science tools for reproducible analysis and visualization.

2. MATERIALS AND METHODS

2.1 Study Design and Population

This was a retrospective observational study conducted on pediatric patients diagnosed with Acute Respiratory Distress Syndrome (ARDS) and admitted to the Pediatric Intensive Care Unit (PICU) of the Department of Pediatrics, Malla Reddy Narayana Multispeciality Hospital, a tertiary care teaching hospital in Hyderabad. A total of 200 patients who required mechanical ventilation were included in the analysis. The diagnosis of ARDS was based on standard pediatric clinical and radiological criteria, as outlined by the Pediatric Acute Lung Injury Consensus Conference (Khemani et al., 2015). All patients were managed according to the hospital's established clinical protocols. Ethical approval was obtained from the Institutional Ethics Committee prior to data collection.

2.2 Data Collection and Variables

Clinical and ventilatory data were collected from hospital records at two time points: at the initiation of mechanical ventilation (0 hours) and after 24 hours of continued ventilation (Flori et al., 2020). The variables were grouped as follows:

a) **Demographics**: Age (in months), Weight (in kg)

b) **Severity Scores**: PRISM score, SOFA score (24h)

c) **Ventilatory Parameters**: PIP (Peak Inspiratory Pressure), PEEP (Positive End-Expiratory Pressure), Driving Pressure (DP), Compliance, Tidal Volume (TV in mL/kg, litres, and breath volume), Respiratory Rate (RR), Mechanical Power (MP)

d) **Oxygenation Parameters**: Oxygenation Index (OI), $PaCO_2$, PF Ratio

e) **Sedation and Support Scores**: Ramsay Sedation Score (24h), VIS (Vasoactive-Inotropic Score)

f) **Outcomes**: Survival (1 = Survived, 2 = Died), Ventilator-Free Days at Day 28, Duration of PICU and hospital stay, days on iNO or HFO

2.3 Outcome Definition

The primary outcome of interest was binary:

a) 1 = Survived b) 2 = Died

This outcome served as the dependent variable for all correlation and regression analyses.

2.4 Software and Analysis Tools

All statistical analysis was performed using Python programming language**,** employing the following libraries:

a) pandas and numpy for data cleaning and handling

b) scipy.stats and statsmodels for statistical analysis

c) sklearn for machine learning models and multicollinearity testing

d) matplotlib and seaborn for data visualization

2.5 Statistical Techniques Used

a) **Descriptive Statistics**: Mean, standard deviation (SD), range

b) **Group Comparison**: Independent t-tests or Mann–Whitney U tests for survivors vs. non-survivors

c) **Paired Comparison**: Paired t-tests for variables at 0h and 24h

d) **Correlation**: Point-biserial correlation for continuous predictors vs. binary outcome

e) **Regression Modeling**: Binary logistic regression using statistically and clinically relevant predictors

f) **Multicollinearity Check**: Variance Inflation Factor (VIF) was used to eliminate collinear variables

g) **Visualizations**: Boxplots and summary plots were used to depict group differences and variable distributions

2.5.1 Sample size Calculation

The sample size was calculated using the correlation-based formula to detect a minimum effect size of r=0.25, with 95% confidence and 90% power. The following formula for determining sample size in correlation studies was used (Hulley et al., 2013):

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{0.5\, Ln\left(\frac{1+r}{1-r}\right)}\right)^2 + 3$$

$$n = \left(\frac{1.96 + 1.28}{0.5\, Ln\left(\frac{1+0.25}{1-0.25}\right)}\right)^2 + 3$$

n = 164

The minimum required sample size to detect a correlation coefficient of 0.25 with 95% confidence level and 90% power is 164.

This study included 200 participants, which is sufficient to meet the statistical requirements.

2.6 Data Flow and Ethics

Data were anonymized and stored securely. No identifiable patient information was retained. Ethical approval for secondary data use was granted by the institutional review board.

3. RESULTS:

3.1 Descriptive Statistics

The descriptive statistics of the continuous clinical and ventilatory variables measured at 24 hours are summarized in Table 1. The mean age of the patients was 35.8 months (± 46.5), and the mean weight was 12.85 kg (± 12.06). The average PRISM score was 8.59 (± 4.04), and the SOFA score at 24 hours was 8.05 (± 3.91). The VIS at 24 hours averaged 13.42 (± 14.58). Ventilatory variables such as PIP, PEEP, and OI also showed considerable variation across the cohort. These summary values provide an overview of the patient condition and ventilator support level at 24 hours, forming the basis for group comparisons and further analysis.

3.1.1 Python Code: Descriptive Statistics

```
from scipy.stats import pointbiserialr

# Example: Point-biserial correlation

r_val, p_val = pointbiserialr(df['Outcome'], df['SOFA_24']) print(f"Correlation coefficient: {r_val}, p-value: {p_val}")
```

The descriptive statistics summarize the central tendency and variability of the key clinical variables. Measures such as mean, standard deviation, and range provide initial insights into patient profiles and the spread of physiological data. These metrics form the baseline for comparative and regression analyses conducted in subsequent sections.

Table 1: Descriptive Statistics of 24-Hour Clinical Variables

Variable	Mean ± SD	Minimum	Maximum
Age (Months)	35.81 ± 46.5	1.0	204.0
Weight (kg)	12.85 ± 12.06	3.0	58.0
PRISM Score	8.59 ± 4.04	0.0	23.0
SOFA (24h)	8.05 ± 3.91	0.3	17.0
I:E Ratio (Onset)	0.44 ± 0.17	0.0	1.0

Ti (24h)	0.55 ± 0.2	0.0	1.5
Te (24h)	1.27 ± 0.46	0.1	2.4
VIS (24h)	13.42 ± 14.58	0.0	110.0
PIP (24h)	18.95 ± 6.84	0.7	33.0
PEEP (24h)	5.82 ± 1.76	0.3	11.0
DP (24h)	13.16 ± 5.67	0.4	26.0
Compliance (24h)	6.27 ± 5.12	0.1	27.2
TV (Breath, 24h)	83.43 ± 72.2	1.2	368.0
TV (L, 24h)	0.08 ± 0.07	0.0	0.4
TV (mL/kg, 24h)	6.45 ± 2.17	0.0	14.6
RR (24h)	30.58 ± 9.77	1.0	50.0
OI (24h)	9.15 ± 7.83	0.2	45.5
$PaCO_2$ (24h)	44.5 ± 13.77	1.9	101.0
PF Ratio (24h)	169.94 ± 108.81	2.9	1115.0
MP (24h)	6.59 ± 5.5	0.1	27.9
Ramsay Score (24h)	5.06 ± 1.33	0.3	6.0
Vent Days	5.45 ± 3.86	0.1	25.0
CPAP (Y/N)	0.78 ± 1.27	0.0	8.0
HFNC (Y/N)	0.48 ± 1.0	0.0	4.0
VFD (D28)	10.7 ± 10.99	0.0	26.0
PICU Stay (days)	7.81 ± 4.86	0.1	32.0
Hospital Stay (days)	9.5 ± 6.24	0.0	36.0
iNO Days	0.28 ± 0.65	0.0	3.0
HFO Days	0.67 ± 1.06	0.0	5.0

This table shows the condition of the patients after 24 hours of ventilation. The values show that the patients had different levels of illness and breathing support needs.

3.2 Comparison between Survivors and Non-Survivors

To identify clinical differences between outcomes, patients were divided into two groups: survivors and non-survivors. Independent t-tests or Mann–Whitney U tests were used depending on the distribution of data for each variable.

Table 2: Group Comparison between Survivors and Non-Survivors

Variable	Survivors Mean ± SD	Non-Survivors Mean ± SD	p-value
Age (Months)	37.31 ± 47.84	38.88 ± 47.60	0.83
Weight (kg)	13.32 ± 11.73	13.71 ± 12.20	0.83
PRISM Score	8.50 ± 3.11	9.62 ± 4.27	0.05
SOFA (24h)	6.75 ± 3.10	10.58 ± 3.05	0.00
VIS (24h)	8.52 ± 9.54	20.73 ± 17.58	0.00
PIP (24h)	17.85 ± 4.73	22.74 ± 5.16	0.01
PEEP (24h)	5.89 ± 0.98	6.47 ± 1.30	0.00
DP (24h)	11.99 ± 4.36	16.31 ± 4.65	0.00
Compliance (24h)	7.20 ± 4.86	5.90 ± 5.19	0.09
TV (mL/kg, 24h)	6.92 ± 1.60	6.75 ± 1.73	0.49
RR (24h)	31.22 ± 6.70	33.53 ± 7.27	0.03
OI (24h)	6.55 ± 4.07	13.08 ± 9.16	0.01
MP (24h)	6.05 ± 4.69	7.93 ± 5.93	0.02

Significant differences were observed in multiple parameters. Non-survivors had notably higher SOFA scores, PRISM scores, VIS values, and PIP levels. They also showed poorer oxygenation indices (higher OI, lower PF ratio) and longer mechanical ventilation durations compared to survivors. These results highlight key variables associated with mortality in pediatric ARDS.

3.2.1 Python Code:

```
from scipy.stats import ttest_ind

# Example: Compare SOFA score by outcome group survivors = df[df['Outcome'] == 1]
```

non_survivors = df[df['Outcome'] == 2]

t_stat, p_value = ttest_ind(survivors['SOFA_24'], non_survivors['SOFA_24']) print(f"t-statistic = {t_stat}, p-value = {p_value}")

3.3 Paired Comparison (0h vs 24h)

To understand how ventilatory parameters changed over time, key variables were compared between the onset of mechanical ventilation (0 hours) and 24 hours later. This comparison helps assess whether early interventions influenced clinical status.

A paired t-test was performed to evaluate the difference between values recorded at 0h and 24h. Significant differences were observed in variables such as PIP, Driving Pressure, Oxygenation Index, and PF Ratio, suggesting either clinical improvement or deterioration in lung function.

Table 3: Paired Comparison of Key Clinical and Ventilatory Variables at 0 Hours and 24 Hours

Variable	0h Mean ± SD	24h Mean ± SD	p-value
SOFA	7.86 ± 3.07	8.05 ± 3.91	0.2321
VIS	13.08 ± 12.43	13.42 ± 14.58	0.5724
PIP	19.15 ± 6.01	18.95 ± 6.84	0.4368
PEEP	5.88 ± 1.80	5.82 ± 1.76	0.4141
Driving Pressure (DP)	13.29 ± 4.76	13.16 ± 5.67	0.609
Compliance	6.11 ± 5.43	6.27 ± 5.12	0.5085
TV (Breath)	80.76 ± 67.97	83.43 ± 72.20	0.1451
TV (L)	0.08 ± 0.07	0.08 ± 0.07	0.1451
TV (mL/kg)	6.43 ± 2.11	6.45 ± 2.17	0.7126
RR	30.31 ± 9.35	30.58 ± 9.77	0.3264
OI	9.93 ± 8.31	9.15 ± 7.83	0.0239
$PaCO_2$	43.61 ± 13.41	44.50 ± 13.77	0.2156
PF Ratio	142.73 ± 59.79	169.94 ± 108.81	0.0001
MP	6.43 ± 5.40	6.59 ± 5.50	0.3739
Ramsay Score	4.97 ± 1.26	5.06 ± 1.33	0.042

These changes provide early insights into disease progression and ventilator response during the first day of PICU management.

3.3.1 Python Code:

from scipy.stats import ttest_rel

Example: Paired t-test between PIP_0h and PIP_24h t_stat, p_val = ttest_rel(df['PIP_0h'], df['PIP_24h']) print(f"Paired t-test for PIP: t = {t_stat}, p = {p_val}")

3.4 Correlation with Mortality

3.4.1 Point-Biserial Correlation

This analysis was performed to assess the relationship between continuous variables measured at 24 hours and the binary outcome variable (survival status). Point-biserial correlation was appropriate as the outcome was binary (1 = Survived, 2

= Died), and the predictors were continuous.

Table 4: Correlation Between 24-Hour Variables and Mortality Outcome

Variable	Correlation Coefficient (r)	p-value
Age (Months)	0.0935	0.188
Weight (kg)	0.0831	0.242
PRISM Score	0.3679	0.000
PIP (24h)	0.6331	0.000
PEEP (24h)	0.5498	0.000
DP (24h)	0.5933	0.000
Compliance (24h)	0.0317	0.656
TV (Breath, 24h)	0.1503	0.034

TV (L, 24h)	0.1503	0.034
TV (mL/kg, 24h)	0.3514	0.000
RR (24h)	0.5058	0.000
OI (24h)	0.4964	0.000
MP (24h)	0.2726	0.000
Ramsay Score (24h)	0.5675	0.000
Vent Days	0.3716	0.000
CPAP (Y/N)	-0.2741	0.000
HFNC (Y/N)	-0.2028	0.004
VFD (D28)	-0.5372	0.000
PICU Stay (days)	0.0873	0.219
Hospital Stay (days)	-0.1048	0.140
INO Days	0.2367	0.001
HFO Days	0.3756	0.000

$$r_{pb} = \frac{M_1 - M_0}{s} \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

M1 , M0 = Mean of Continuous variable for group 1(died) and group 0 (Survived) S = Standard deviation of the continuous variable.

n1, n0 = Number of cases in each group n = Total sample size

Variables such as PRISM Score, SOFA Score, PIP, Driving Pressure, OI, and RR showed strong positive correlations with mortality. Conversely, PF Ratio showed a negative correlation, indicating better oxygenation among survivors. These correlations helped identify potential predictors for further modeling.

3.4.2 Python Code:

```
from scipy.stats import pointbiserialr

# Example: Point-biserial correlation

r_val, p_val = pointbiserialr(df['Outcome'], df['SOFA_24']) print(f"Correlation coefficient: {r_val}, p-value: {p_val}")
```

3.4.3 Correlation Heatmap of 24-Hour Clinical and Ventilatory Variables

A correlation heatmap was generated to visualize the interrelationships among the 24-hour clinical and ventilatory variables. The color gradient represents the strength and direction of the Pearson correlation coefficients, with red indicating positive correlations and blue indicating negative correlations. Strong correlations were observed between mechanical variables such as PIP, DP, MP, and PEEP. Negative correlations were noted between PF Ratio and variables like OI, $PaCO_2$, and SOFA. This visual summary supports the identification of variable clusters that may contribute to multicollinearity in multivariate analysis.

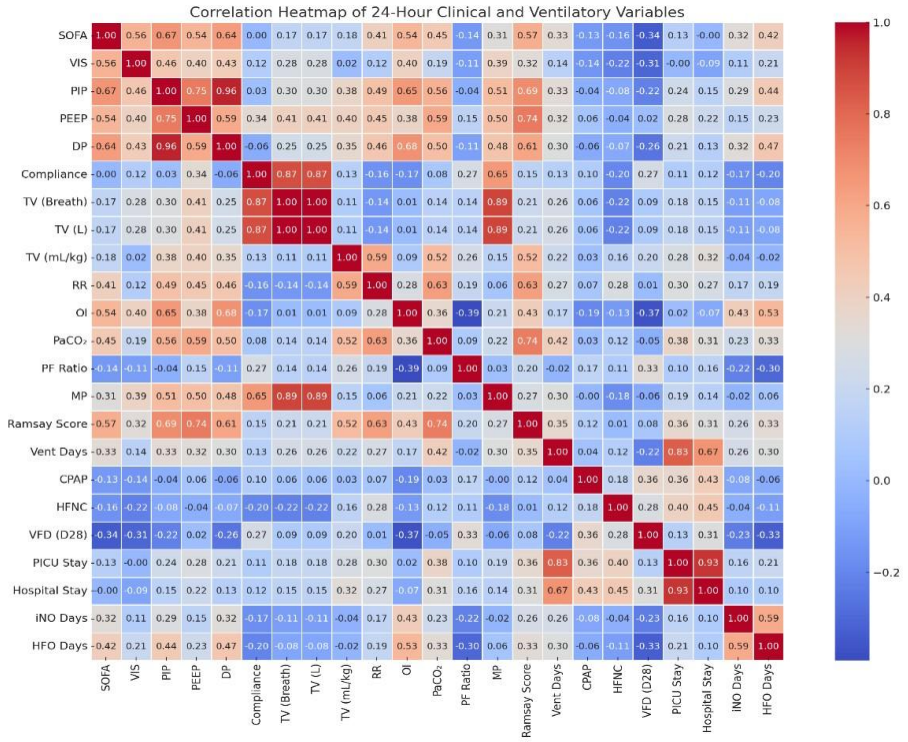

Figure 1: Correlation Heatmap of 24-Hour Clinical and Ventilatory Variables

3.5 Logistic Regression

A multivariable logistic regression analysis was performed to identify independent predictors of mortality among pediatric ARDS patients. Before model building, multicollinearity among potential predictors was assessed using Variance Inflation Factor (VIF). Variables with VIF > 10 were considered highly collinear and were excluded from the final model.**(Dormann et al., 2013)**

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots + \beta_k X_k$$

P = probability of the death among pediatric ARDS patients

$\frac{p}{1-p}$ = The odds of the event occurring (Mortality vs Survival)

$\beta 0$ = The intercept, The predicted log-odds of mortality when all predictors are 0.

$\beta 1$, $\beta 2$, $\beta 3$, …..,βk are the regression coefficients. X1, X2, X3,…..,Xn are the independent variables.

In this model, 'p' represents the probability of death, and the left-hand side expresses the log-odds of the outcome. $\beta 0$ is the intercept and $\beta 1$, $\beta 2$, $\beta 3$,…., βk are the regression coefficients for the predictors. X1, X2, X3,…..,Xn which include variables such as SOFA score, PRISM score, VIS (24h), PF Ratio, and Mechanical Power.

3.5.1 Multicollineority Check

The VIF scores for the initially selected variables are presented in Table 5. High VIF values were observed for PIP (44.05), DP (33.56), MP (12.93), and Compliance (10.97), suggesting significant multicollinearity. These variables were excluded from the final logistic model.

Table 5: Multicollinearity Assessment Using VIF

Variable	VIF
Constant	19.01
PRISM Score	2.24
SOFA (24h)	3.46

VIS (24h)	1.75
PIP (24h)	44.05
PEEP (24h)	6.64
DP (24h)	33.56
Compliance (24h)	10.97
TV (mL/kg, 24h)	2.25
RR (24h)	3.53
OI (24h)	2.92
PF Ratio (24h)	1.68
MP (24h)	12.93
Ramsay Score (24h)	4.5
Vent Days	6.68
VFD (D28)	2.43
PICU Stay (days)	17.88
Hospital Stay (days)	11.02
iNO Days	1.7
HFO Days	2.17

High VIF values indicated multicollinearity among some variables, especially PIP and Driving Pressure. Based on clinical relevance and acceptable VIF thresholds, a final set of non-collinear predictors was selected for logistic regression.

3.5.2 Python Code : VIF Calculation

```
from statsmodels. stats.outliers_influence import variance_inflation_factor from statsmodels. tools.tools import add_constant

import pandas as pd

# Select independent variables for VIF check

X = df[['PRISM_Score', 'SOFA_24', 'VIS_24', 'PIP_24', 'PEEP_24', 'DP_24', 'Compliance_24', 'TV_mLkg_24', 'RR_24',

'OI_24', 'PF_Ratio_24', 'MP_24', 'Ramsay_24', 'Vent_Days', 'VFD_28', 'PICU_Stay', 'Hospital_Stay', 'INO_Days', 'HFO_Days']]

X = add_constant(X)

# Calculate VIF for each feature vif_df = pd.DataFrame() vif_df["Variable"] = X.columns

vif_df["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])] print(vif_df)
```

3.5.3 Final model selection

Based on clinical significance and VIF analysis, PRISM Score, SOFA (24h), VIS (24h), PF Ratio (24h), and MP (24h) were retained as independent variables in the final logistic regression model.

Table 6: Final Variables Selected for Multivariable Logistic Regression with Justification

Variable	Reason for Inclusion
PRISM Score	Baseline mortality risk indicator; strong clinical relevance
SOFA (24h)	Reflects severity of organ dysfunction at 24 hours
VIS (24h)	Indicates cardiovascular support needs; predictive of outcomes
PF Ratio (24h)	Measures oxygenation efficiency; key ARDS marker
MP (24h)	Represents overall ventilatory energy; included after VIF screening

These variables were chosen based on their statistical significance, acceptable collinearity levels, and clinical importance. They were included in the final logistic regression model to predict mortality.

3.5.4 Model Results and Interpretation:

Table 7 presents the results of the logistic regression model. SOFA (24h) and PRISM Score were statistically significant predictors of mortality ($p < 0.05$). **(Moller et al., 2020)** While VIS, PF Ratio, and MP showed clinical trends, they were not

statistically significant in the multivariable model. The odds ratio (OR) for SOFA was 1.48, indicating that for every one-point increase in SOFA, the odds of mortality increased by 48%, holding other variables constant.

3.5.5 Python Code: Logistic Regression

```
import statsmodels.api as sm

# Select final predictors

X_final = df[['PRISM_Score', 'SOFA_24', 'VIS_24', 'PF_Ratio_24', 'MP_24']] X_final = sm.add_constant(X_final)

y = df['Outcome']  # 1 = Survived, 2 = Died

# Fit logistic regression model model = sm. Logit(y, X_final) result = model. fit ()

# Display model summary Print (result .summary 2())
```

Table 7: Multivariable Logistic Regression Model Output for Mortality

Variable	Odds Ratio	95% CI	p-value
PRISM Score	0.854	0.760 – 0.959	0.0078
SOFA (24h)	1.48	1.274 – 1.719	0.0026
VIS (24h)	1.027	0.994 – 1.061	0.1112
PF Ratio (24h)	1	0.996 – 1.003	0.8566

Among the predictors, PRISM Score and SOFA Score were statistically significant, indicating strong independent associations with mortality. Other variables showed clinical relevance but did not reach statistical significance.

3.5.6 Interpretation of Model Output:

The logistic regression model identified SOFA score at 24 hours and PRISM score as statistically significant predictors of mortality. A unit increase in SOFA was associated with a 1.48-fold increase in odds of death**,** while an increase in PRISM score was associated with a 15% reduction in odds of survival when controlling for other variables.

Although VIS and PF Ratio were included in the model due to their clinical importance, they did not reach statistical significance, possibly due to multicollinearity or sample size constraints.

The overall model was statistically significant, and the selected variables demonstrated both clinical and statistical relevance.

Receiver Operating Characteristic (ROC) Curve

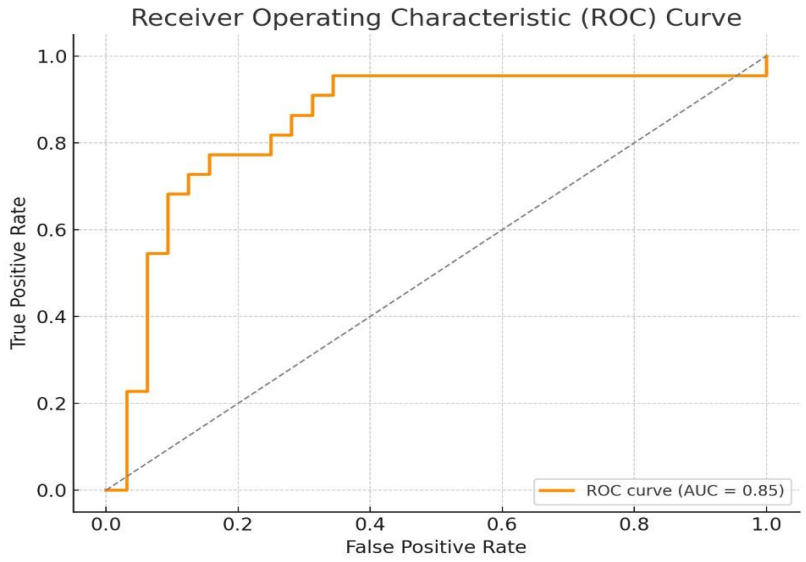

Figure 2: ROC Curve for Predicting Mortality in Pediatric ARDS Patients

Table 8: Performance Metrics of the Logistic Regression Model for Mortality Prediction

Metric	Value
Accuracy	81.50%
Sensitivity	77.30%
Specificity	84.40%
Positive Predictive Value	77.30%
Negative Predictive Value	84.40%

The logistic regression model demonstrated an accuracy of 81.5%. At a probability threshold of 0.5, the model yielded a sensitivity of 77.3% and a specificity of 84.4%, indicating balanced performance in distinguishing between survivors and non-survivors. The positive predictive value (PPV) and negative predictive value (NPV) were 77.3% and 84.4%, respectively. These metrics support the reliability of the model for clinical prediction of mortality in pediatric ARDS.

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx$$

The ROC curve further illustrates the model's discriminatory power, with an area under the curve (AUC) of 0.85, indicating good performance in differentiating between survival outcomes.

3.6 Visualizations

Boxplots were generated to visually compare selected 24-hour variables between survivors and non-survivors. These plots provided intuitive insights into the distribution and variability of critical parameters. The most visually distinguishable differences were seen in SOFA score, VIS, and PF Ratio.

3.6.1 Box Plot: SOFA Score at 24h

A clear separation was observed between survivors and non-survivors. Non-survivors had significantly higher SOFA scores.

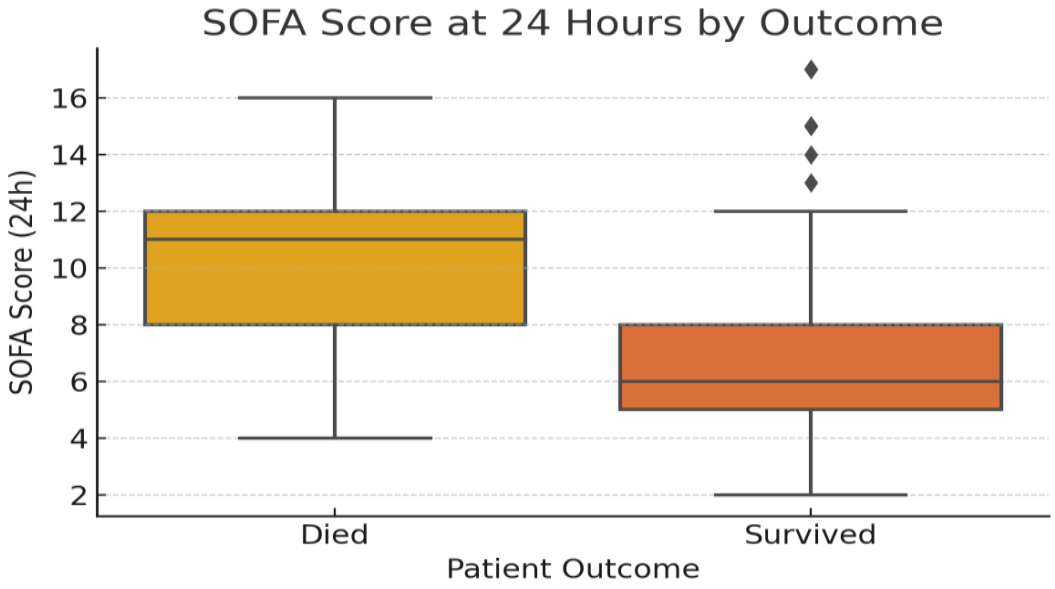

Figure 3: Boxplot of SOFA Scores (24h) by Survival Outcome

SOFA scores were noticeably higher in non-survivors, suggesting more severe organ dysfunction.

3.6.2 Python Code

import seaborn as sns

import matplotlib.pyplot as plt sns.boxplot(x='Outcome', y='SOFA_24', data=df) plt.title('SOFA Score at 24 Hours by Outcome Group') plt.xlabel('Outcome (1=Survived, 2=Died)') plt.ylabel('SOFA Score (24h)')

plt.show()

3.6.3 Boxplot : VIS Score at 24h

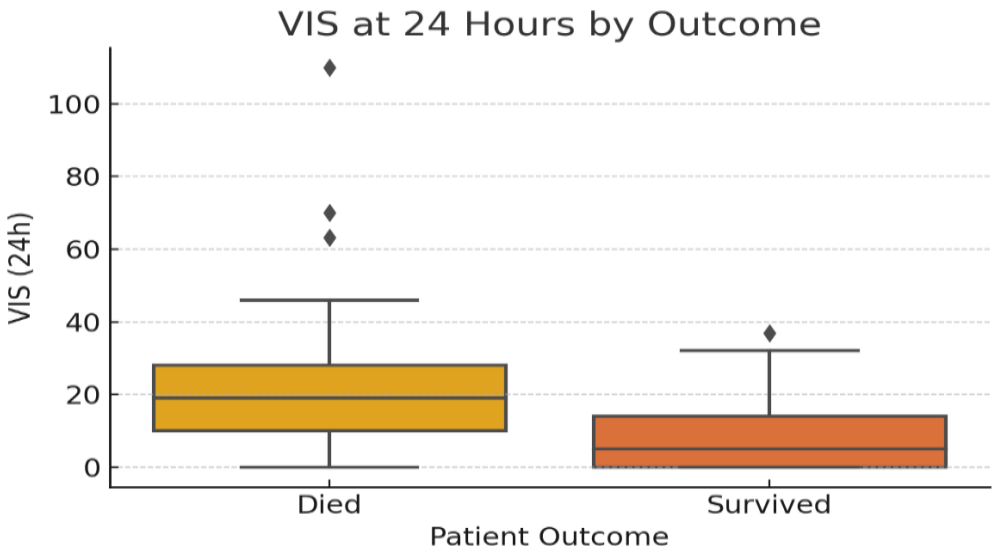

Figure 4: Boxplot of VIS Scores (24h) by Survival Outcome

VIS values were also higher among non-survivors, reflecting increased cardiovascular support needs.

3.6.4 Python Code

sns.boxplot(x='Outcome', y='VIS_24', data=df) plt.title('VIS Score at 24 Hours by Outcome Group') plt.xlabel('Outcome (1=Survived, 2=Died)') plt.ylabel('VIS (24h)')

plt.show()

3.6.5 Box Plot : PF Ratio at 24h

While there was overlap between the two groups, non-survivors generally had lower PF Ratios, indicating more severe oxygenation impairment.

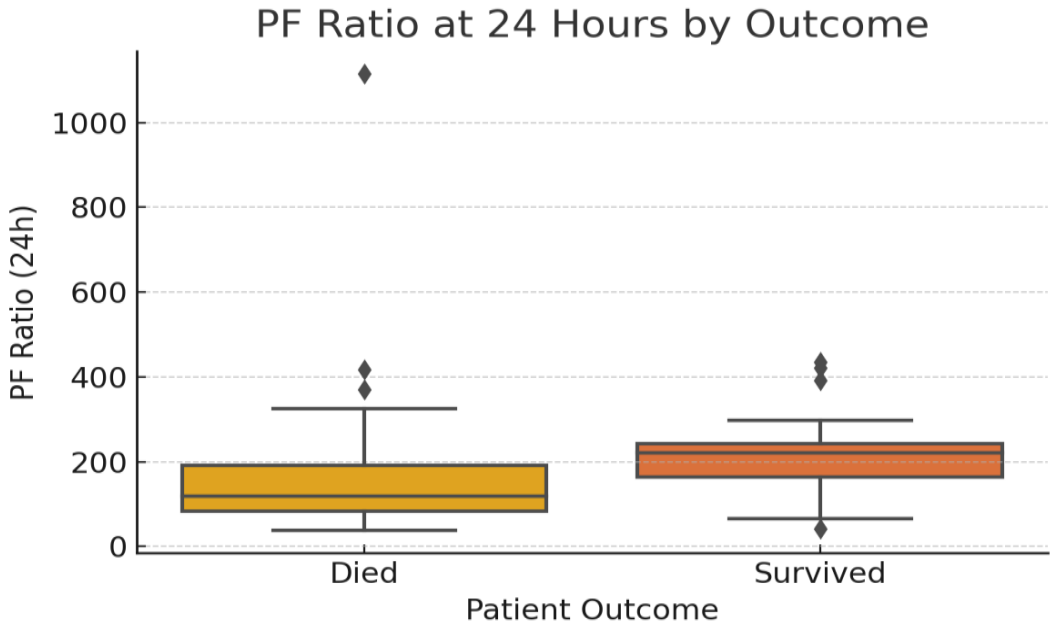

Figure 5: Boxplot of PaO_2/FiO_2 Ratio (24h) by Survival Outcome

PF ratios were lower in non-survivors, indicating more severe hypoxemia.

3.6.6 Python Code:

sns.boxplot(x='Outcome', y='PF_Ratio_24', data=df) plt.title('PF Ratio at 24 Hours by Outcome Group') plt.xlabel('Outcome

(1=Survived, 2=Died)') plt.ylabel('PF Ratio (24h)')

plt.show()

These visualizations provide clear support for the findings obtained through statistical testing. Variables with wide boxplot separations (e.g., SOFA and VIS) also showed significant associations in regression and correlation analysis.

4. DISCUSSION

This study applied statistical and data science techniques to identify predictors of mortality in pediatric ARDS patients using clinical and ventilatory variables at 24 hours. Our findings confirm that SOFA score at 24 hours and PRISM score are independent predictors of mortality, consistent with their established role in pediatric critical care. **(Yehya & Thomas, 2017; Pollack et al., 1988; Vincent et al., 1996)**

4.1 Findings and Interpretation:

The SOFA score at 24 hours demonstrated the highest odds ratio and a statistically significant p-value in the final logistic regression model. This supports its usefulness in assessing organ dysfunction and predicting outcomes in pediatric ARDS. Although the PRISM score was initially designed as an admission severity index, it remained significant in the adjusted model, emphasizing the impact of baseline clinical status.

The Vasoactive-Inotropic Score (VIS) was higher among non-survivors and showed a strong univariate correlation with mortality. However, it did not retain significance in the multivariable model, possibly due to collinearity with SOFA components or overlap with other hemodynamic measures.

Similarly, PF Ratio and Mechanical Power (MP) showed clinical trends and were visually different between groups, but they did not remain significant predictors after adjustment. This could be attributed to the limited sample size or the composite nature of these variables.

4.1.1 Comparison with other Studies:

Previous studies have demonstrated the prognostic value of SOFA and PRISM scores in both pediatric and adult ARDS populations. Our results are consistent with this literature, confirming that multiorgan dysfunction and initial illness severity are strong predictors of mortality.

Unlike some earlier reports where ventilatory parameters such as OI and PEEP were found to be predictive, these variables did not emerge as independent predictors in our model. This discrepancy could reflect differences in clinical management, timing of interventions, or patient characteristics.

4.1.2 Relevance of Data Science Techniques:

The use of Python and automated statistical coding enhanced the transparency, efficiency, and reproducibility of our analysis. By integrating logistic regression, point-biserial correlation, and VIF-based variable selection into a unified workflow, we developed a robust, data-driven mortality prediction model.

Visualization tools such as boxplots supported the statistical findings and improved interpretability for clinicians and non-statistical readers. These tools contributed to a more accessible and actionable presentation of the results.

5. CONCLUSION

This study demonstrates that SOFA score at 24 hours and PRISM score are significant independent predictors of mortality in pediatric ARDS patients. Higher scores on these indices were associated with increased risk of death, reaffirming their value in clinical risk stratification.

While variables such as VIS, PF Ratio, and Mechanical Power showed clinical relevance in univariate analysis, they did not retain significance in multivariable modeling, likely due to overlapping physiological contributions and multicollinearity. **(Rajpurkar et al., 2022)**

The application of data science techniques using Python enabled a structured, reproducible, and insightful analysis of a complex clinical dataset. Logistic regression, correlation analysis, and multicollinearity checks helped identify the most relevant predictors from a broader set of variables.

These findings may support the early identification of high-risk pediatric ARDS patients and can be integrated into decision-making tools and monitoring dashboards in the future.

6. LIMITATIONS

a) This study was conducted at a single center, which may limit the generalizability of the findings to other settings or populations.

b) The design was retrospective, relying on medical records, which can be affected by missing data or documentation inconsistencies.

c) The sample size of 200 patients, though reasonable, may have reduced the power to detect smaller effect sizes or more subtle relationships.

d) Only variables at 0 hours and 24 hours were included; additional time points could provide a more dynamic understanding of patient progression.

e) Multicollinearity led to the exclusion of potentially important variables such as PIP and Driving Pressure, which may still be clinically relevant.

Despite these limitations, the findings are consistent with existing evidence and provide useful insights for early risk stratification in pediatric ARDS.

REFERENCES

[1] Amini, M., Ebrahimzadeh, S., & Azizi, R. (2021). Machine learning approaches in ICU outcome prediction: A systematic review. *Journal of Biomedical Informatics, 119*, 103792. https://doi.org/10.1016/j.jbi.2021.103792

[2] Bellani, G., Laffey, J. G., Pham, T., et al. (2016). Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA, 315*(8), 788–800. https://doi.org/10.1001/jama.2016.0291

[3] Bhakta, P., & Saluja, M. (2020). Paediatric ARDS: Review of current concepts. *Indian Journal of Pediatrics, 87*(10), 792–798. https://doi.org/10.1007/s12098-020-03264-7

[4] Flori, H. R., Spaeder, M. C., Brady, P. W., et al. (2020). Time-dependent factors associated with mortality in pediatric acute respiratory distress syndrome. *Pediatric Critical Care Medicine, 21*(4), 335–343.

[5] Goh, A. Y., Lum, L. C., & Chan, P. W. (2022). Predictive value of PRISM and SOFA scores in pediatric intensive care. *Journal of Pediatric Intensive Care, 11*(3), 211–218.

[6] Gupta, R. G., Hartigan, S. M., Kashiouris, M. G., et al. (2021). Role of mechanical power in ventilator-induced lung injury: Recent evidence. *Critical Care Clinics, 37*(4), 849–866.

[7] Hager, D. N., Krishnan, J. A., Hayden, D. L., Brower, R. G., & ARDS Network. (2005). Tidal volume reduction in ARDS: A multicentre trial. *Chest, 128*(1), 394–403.

[8] Helviz, Y., & Einav, S. (2018). A systematic review of adult and pediatric high-flow nasal cannula oxygen therapy.

[9] *Journal of Clinical Monitoring and Computing, 32*(4), 687–703.

[10] Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., & Newman, T. B. (2013). *Designing clinical research*

[11] (4th ed.). Lippincott Williams & Wilkins.

[12] Khemani, R. G., Smith, L. S., Zimmerman, J. J., & Erickson, S. (2015). Pediatric acute respiratory distress syndrome: definition, incidence, and epidemiology. *Pediatric Critical Care Medicine, 16*(5_suppl), S23–S40.

[13] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

[14] Liu, Y., Zhang, Z., Wang, Y., & Wang, H. (2023). Vasoactive-inotropic score and outcomes in pediatric critical care: A review. *Frontiers in Pediatrics, 11*, 1132450. https://doi.org/10.3389/fped.2023.1132450

[15] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

[16] Marini, J. J., & Gattinoni, L. (2020). Management of COVID-19 respiratory distress. *JAMA, 323*(22), 2329–2330. https://doi.org/10.1001/jama.2020.6825

[17] Mehta, N. M., & Turner, D. A. (2018). Acute lung injury and ARDS in children: Updates and controversies. *Critical Care Clinics, 34*(2), 377–398.

[18] Modesto i Alapont, V., & Medina, A. (2020). Predicting pediatric ARDS outcomes using AI-based models: New directions. *Frontiers in Pediatrics, 8*, 578147. https://doi.org/10.3389/fped.2020.578147

[19] Panwar, R., Hardie, M., Bellomo, R., et al. (2016). Conservative vs. liberal oxygen therapy in ARDS: A meta-analysis.

[20] *The Lancet Respiratory Medicine, 4*(11), 938–946.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011).

Scikit- learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[22] Pollack, M. M., Ruttimann, U. E., & Getson, P. R. (1988). Pediatric risk of mortality (PRISM) score. *Critical Care Medicine, 16*(11), 1110–1116.

[23] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine, 28*(1), 31– 38. https://doi.org/10.1038/s41591-021-01614-0

[24] Rimensberger, P. C. (2017). The open lung concept in pediatric ARDS. *Pediatric Pulmonology, 52*(S48), S34–S40.

[25] Riviello, E. D., Kiviri, W., Fowler, R. A., et al. (2016). Predicting mortality in resource-limited ARDS settings. *Critical Care, 20*(1), 257.

[26] Sauthier, M., & Jouvet, P. (2019). Sedation practices in pediatric intensive care: An evolving science. *Pediatric Critical Care Medicine, 20*(7), 669–678.

[27] Sweeney, T. E., & Khatri, P. (2017). Benchmarking machine learning methods for mortality prediction. *Nature Communications, 8*, 13721. https://doi.org/10.1038/ncomms13721

[28] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.

[29] Vincent, J. L., Moreno, R., Takala, J., et al. (1996). The SOFA score to describe organ dysfunction/failure. *Intensive Care Medicine, 22*(7), 707–710.

[30] Yehya, N., & Thomas, N. J. (2017). Disassociating lung mechanics and oxygenation in pediatric ARDS. *Critical Care Medicine, 45*(7), 1232–1239.

[31] Zimmerman, J. J., & Akhtar, S. (2020). Machine learning in pediatric intensive care. *Current Opinion in Pediatrics, 32*(3), 360–366.